Abdullah Mobeen

Professor Jeffrey Simonoff

Regression and Multivariate Data Analyses

Kicking the million dollar ball: Data Analysis to estimate the value of the Soccer Players

Soccer industry is one of the most valuable industries with billions of pounds in investments. This could be estimated from the fact that the soccer club leagues (not national teams) in Europe alone have a collective market value of £22 billion. The total market value of a league is the sum of the individual market value of the soccer players in that league, who could be bought by any other club for that amount. Therefore, the player transfer is an important aspect of club-level soccer because every team management wants to optimize the buying of players while remaining in its budget. Naturally, the question of what factors determine a player's value is not only of interest to the team managements and the players, but also for the investors funding the teams.

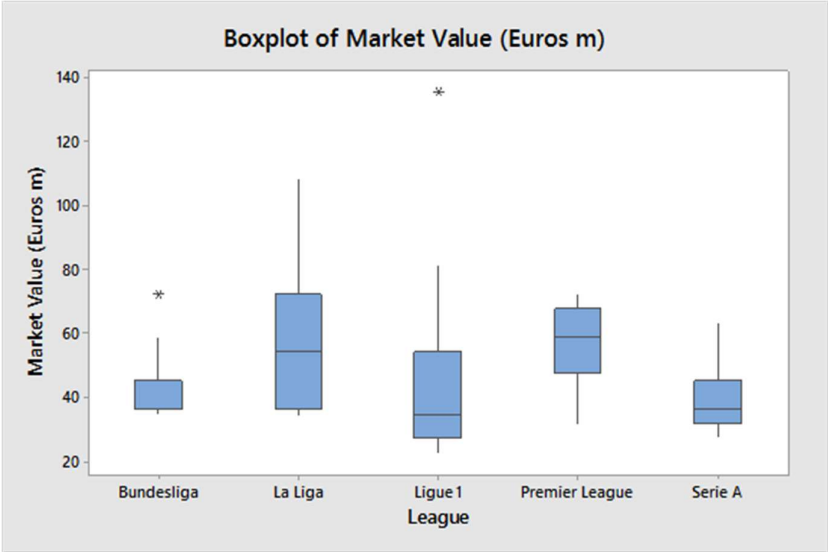
While there exist many factors that influence a player's value, I have focused on two major factors: league the player plays in and the position he plays on. I will try to analyse the data to check how the market value of a player changes across different leagues and different positions. To understand the structure of these 'leagues' and the responsibilities of different positions, let us briefly go over a few dynamics.

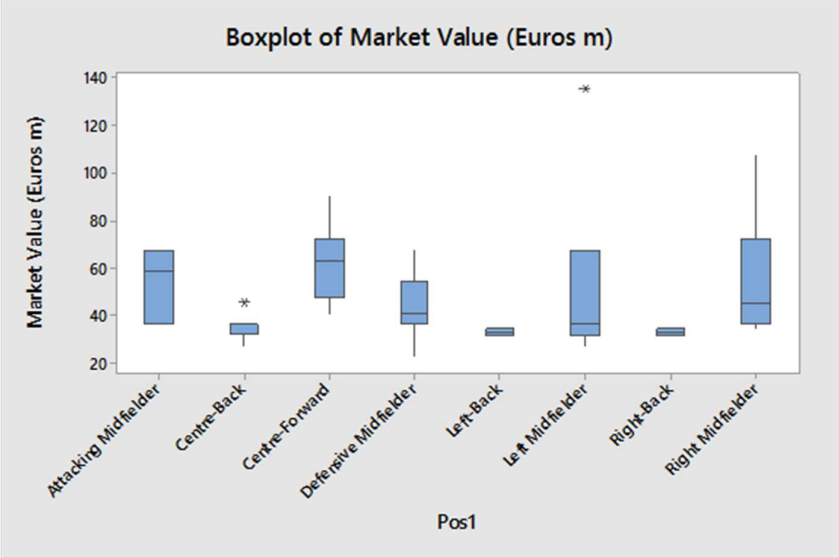
Soccer is a simple game between two teams, each consisting of eleven players. While soccer does have a well-defined set of rules, a high-level description of the sport is that each team has to score goals against each other by putting the soccer ball in the opposition's goal. The central idea is that the players, other than goalkeepers, cannot use their arms to control the ball. The team with most goals wins the match and if the goals are level, the match ends in a draw. Below is an example of a team formation (positions) in a typical soccer match:

I have gathered data for 60 soccer players, where each is valued more than **£**20 million. For each soccer player, the response variable is a continuous numerical variable: the **market value in million pounds**, and the predictor variables are two categorical variables: **league and position**. Anybody unfamiliar with soccer would wonder what league and position are. A soccer league is set of twenty clubs that play each other. Leagues normally correspond to countries e.g. the UK will have its leagues with each league having twenty teams based in the UK, however, the players in the leagues could be from anywhere. Each of the 60 soccer players in my data belongs to one of the five prominent soccer leagues – **Premier League** (UK), **Bundesliga** (Germany), **La Liga** (Spain), **Ligue 1** (France), and **Serie A** (Italy). As for

the positions, there are 9 positions a player can play on – **Centre Forward**, **Attacking Midfielder**, **Left Midfielder**, **Right Midfielder**, **Defensive Midfielder**, **Left Back**, **Right Back**, **Centre Back**, and **Goalkeeper**. These are shown in the picture above. In my data, I am beginning with only 8 positions by excluding the Goalkeeper since none of the 60 players is a Goalkeeper. The data for these 60 players is obtained from an official website for transfer market – www.transfermarkt.co.uk.

Let's look at the side-by-side boxplots to see if they indicate the league and position effects:





From the side-by-side boxplot of Leagues, we could see that La Liga and Premier League have means that are closer to each other while Bundesliga, Serie A, and Ligue 1 have means closer to each other. From the side-by-side boxplot, we could see that Centre-Forward and Attacking Midfield have means closer to each other, while the remaining six positions have their means closer to each other. From both the boxplots, we get some indication of main effects if not a strong one. My first attempt at two-way ANOVA model failed to capture the estimate of the interaction between League and Position and gave the model that only accounted for the main effects. Let's look at the cross-classification table to see if it conveys any reason for the failure of modelling interaction effect:

Rows: League Columns: Pos1

	Attacking Midfielder	Centre-Back	Centre-Forward	Defensive Midfielder	Left-Back	Left Midfielder
Bundesliga	0	2	4	3	0	0
La Liga	1	2	4	5	1	0
Ligue 1	0	1	2	3	0	3
Premier League	2	1	3	1	1	2
Serie A	0	2	3	3	0	2
All	3	8	16	15	2	7

	Right-Back	Right Midfielder	All
Bundesliga	1	1	11
La Liga	0	2	15
Ligue 1	0	2	11
Premier League	0	2	12
Serie A	1	0	11
All	2	7	60

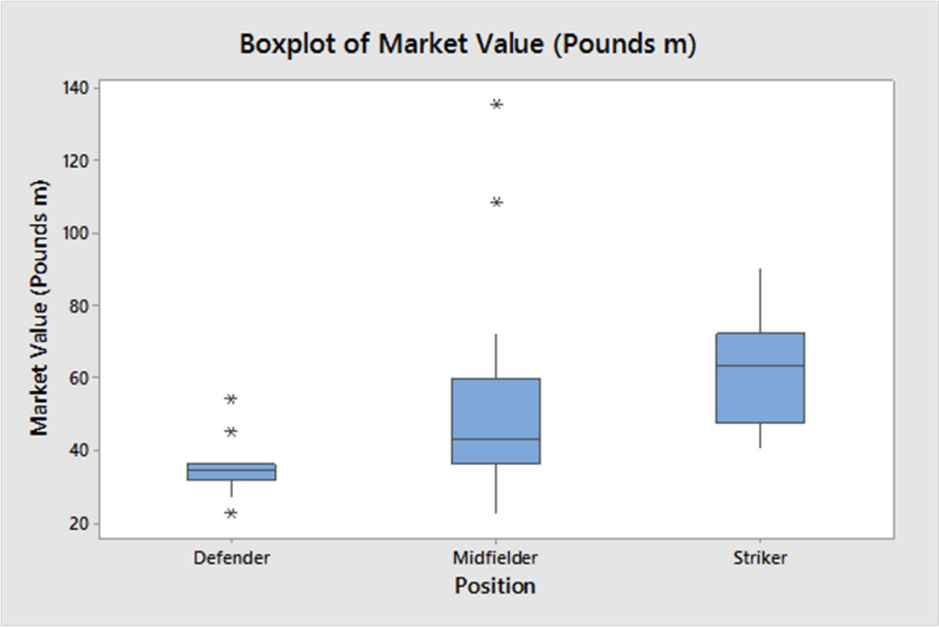
Cell Contents
Count

Now we know the reason for the failure – the holes in our data set. These holes are the zero values in some combinations of Leagues and Positions. With these missing values, it is impossible to fit a model with interaction effect as there are no data points for some combinations of the leagues and players. We could ignore the interaction term and get the model that only represents the main effects, but it will be more interesting to see if the market value of players playing on different positions varies differently per different leagues. So we look for ways to remove these holes and one such way is to change the way we have modelled our predictor variable, Position. If we look at the 8 different positions on the team formation picture above, we see that some positions are really similar to each other, that is, we can categorize our eight positions in only three ways – **Strikers** (Centre Forward), **Midfielders** (Attacking Midfielder, Left Midfielder, Right Midfielder, Defensive Midfielder), and **Defenders** (Centre Back, Right Back, Left Back). This is because, in a typical soccer match, a striker could play at any centre-forward position, a midfielder could play at any midfielder position, and a defender could play at any 'back' position. Therefore, it makes more sense to distinguish players on the bases of more general roles rather than specific locations of the players. The roles of these three classes are such: Strikers are responsible for striking the ball into opponent's net, Midfielders are responsible for assisting the strikers, and Defenders are responsible for defending their own net by tackling the opposition's players. So let's check our cross-classification table again to see if any holes exist with this modelling:

Rows: League Columns: Position

	Defender	Midfielder	Striker	All
Bundesliga	3	4	4	11
La Liga	4	7	4	15
Ligue 1	2	7	2	11
Premier League	2	7	3	12
Serie A	3	5	3	11
All	14	30	16	60

Now we don't see any holes in our data so let's look at the boxplot for these new levels of position:



Let's now proceed with the two-way ANOVA model:

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	2299.9	574.98	1.56	0.202
Position	2	5476.4	2738.21	7.43	0.002
League*Position	8	742.9	92.86	0.25	0.978
Error	45	16592.2	368.72		
Total	59	26088.6			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.2020	36.40%	16.61%	0.15%

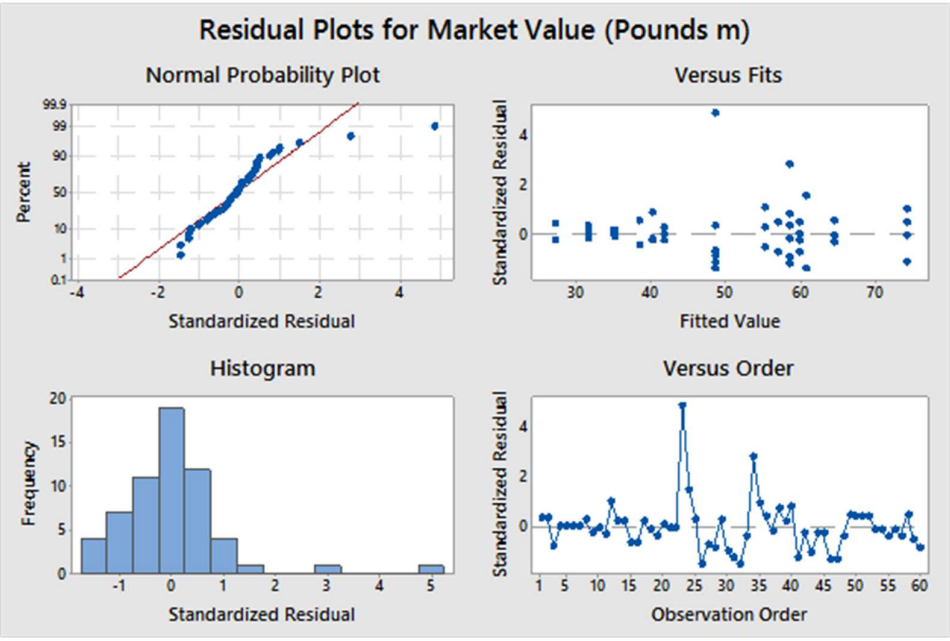
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	48.46	2.70	17.92	0.000	
League					
Bundesliga	-4.61	5.27	-0.87	0.387	1.66
La Liga	9.14	4.81	1.90	0.064	1.61
Ligue 1	-3.01	5.95	-0.51	0.616	2.11
Premier League	5.72	5.60	1.02	0.312	1.95
Position					
Defender	-14.14	4.09	-3.46	0.001	1.36
Midfielder	0.27	3.41	0.08	0.938	1.35
League*Position					
Bundesliga Defender	5.09	7.86	0.65	0.521	2.18
Bundesliga Midfielder	-2.49	7.11	-0.35	0.728	2.18
La Liga Defender	-3.41	7.14	-0.48	0.635	1.94
La Liga Midfielder	0.63	6.16	0.10	0.919	1.95
Ligue 1 Defender	-4.31	9.04	-0.48	0.635	2.22
Ligue 1 Midfielder	2.88	7.09	0.41	0.686	2.30
Premier League Defender	-1.79	8.81	-0.20	0.840	2.31
Premier League Midfielder	5.34	6.79	0.79	0.436	2.25

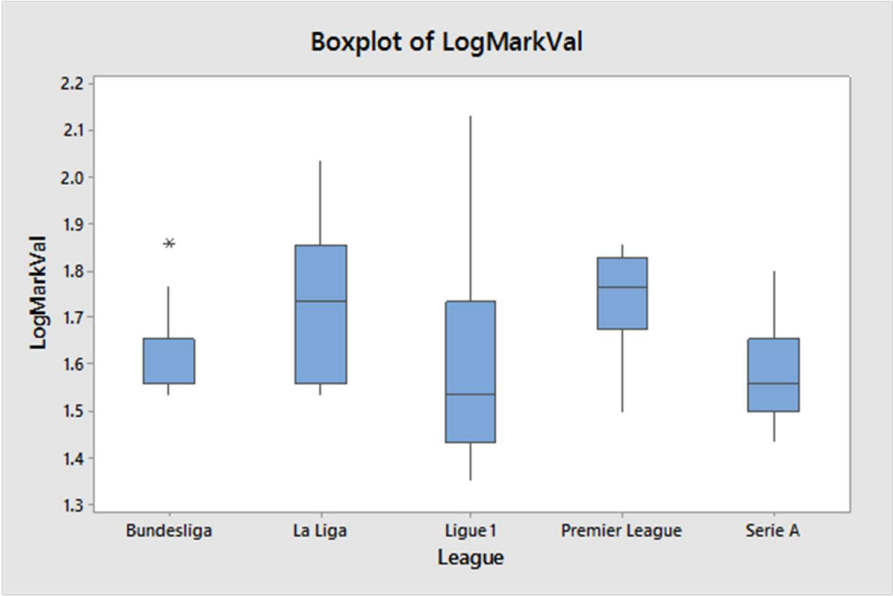
Regression Equation

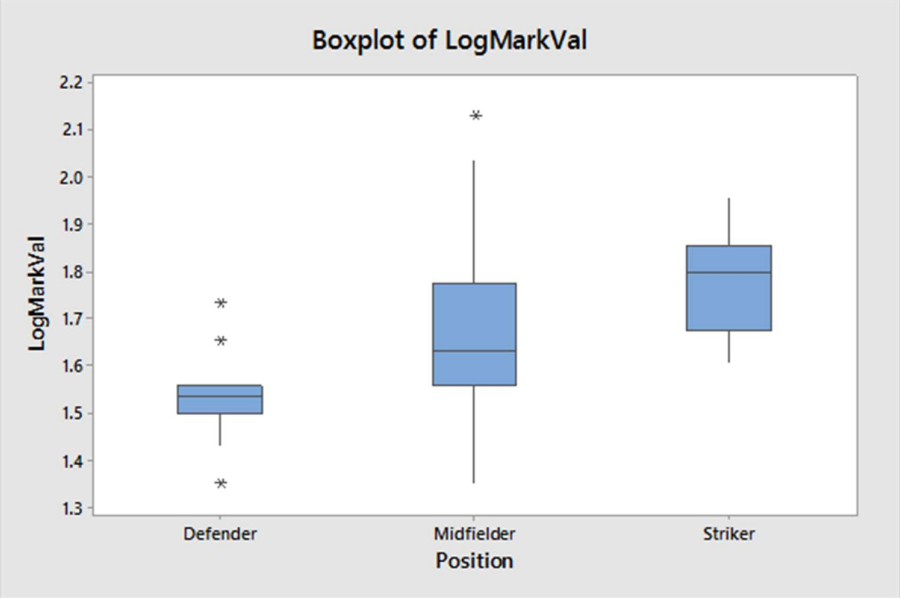
Market Value (Pounds m) = 48.46 - 4.61 League_Bundesliga + 9.14 League_La Liga
- 3.01 League_Ligue 1 + 5.72 League_Premier League
- 7.26 League_Serie A - 14.14 Position_Defender
+ 0.27 Position_Midfielder + 13.87 Position_Striker
+ 5.09 League*Position_Bundesliga Defender
- 2.49 League*Position_Bundesliga Midfielder
- 2.59 League*Position_Bundesliga Striker - 3.41 League*Position_La Liga Defender + 0.63 League*Position_La Liga Midfielder
+ 2.78 League*Position_La Liga Striker - 4.31 League*Position_Ligue 1 Defender + 2.88 League*Position_Ligue 1 Midfielder
+ 1.43 League*Position_Ligue 1 Striker
- 1.79 League*Position_Premier League Defender
+ 5.34 League*Position_Premier League Midfielder
- 3.55 League*Position_Premier League Striker
+ 4.44 League*Position_Serie A Defender
- 6.37 League*Position_Serie A Midfielder
+ 1.93 League*Position_Serie A Striker

We see in the results above that the interaction League*Position is statistically insignificant with a p-value of 0.978. So we cannot reject the null hypothesis that the response mean for the level of one factor does not depend on the value of the other factor level. At this point, we should state that in an unbalanced design situation, an insignificant interaction could make the main effects look statistically significant when in reality they might not be statistically significant. So despite Position being shown as statistically significant, we cannot declare it to be a significant main effect. Let's look at the residual plots:



There are a few very clear outliers in the residual plots that I will talk about in a while. The plots also indicate a very serious case of heteroscedasticity, which could also be observed from the boxplots above. Before dealing with these issues, let's notice that the plots indicate long right-tailed residuals. It does make sense since our response variable is a money variable and we usually do expect to see long right-tailed residuals. So let's log our response variable. Here are the boxplots for the logged response variable:





We still see the same grouping – different means, among the levels of two categorical predictor variables. Let's now take a look at the two-way ANOVA model:

General Linear Model: LogMarkVal versus League, Position

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.18492	0.046230	2.58	0.050
Position	2	0.45815	0.229073	12.81	0.000
League*Position	8	0.05142	0.006428	0.36	0.936
Error	45	0.80500	0.017889		
Total	59	1.58486			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.133749	49.21%	33.40%	15.69%

Coefficients

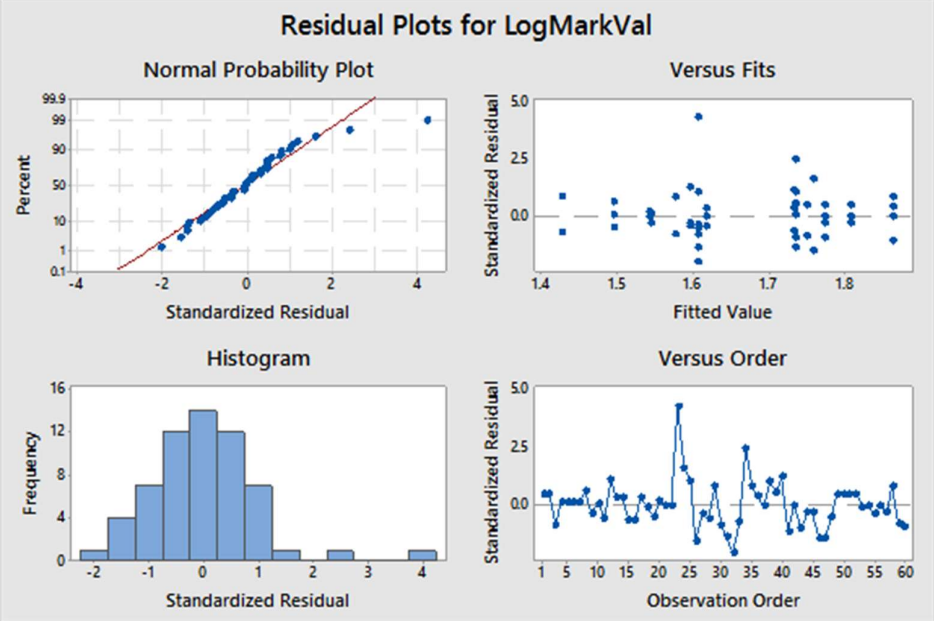
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6546	0.0188	87.85	0.000	
League					
Bundesliga	-0.0240	0.0367	-0.65	0.516	1.66
La Liga	0.0762	0.0335	2.28	0.028	1.61
Ligue 1	-0.0586	0.0414	-1.41	0.165	2.11
Premier League	0.0641	0.0390	1.65	0.107	1.95
Position					
Defender	-0.1281	0.0285	-4.50	0.000	1.36
Midfielder	0.0003	0.0238	0.01	0.991	1.35
League*Position					
Bundesliga Defender	0.0390	0.0548	0.71	0.480	2.18

Bundesliga Midfielder	-0.0133	0.0495	-0.27	0.790	2.18
La Liga Defender	-0.0080	0.0497	-0.16	0.873	1.94
La Liga Midfielder	0.0035	0.0429	0.08	0.934	1.95
Ligue 1 Defender	-0.0427	0.0630	-0.68	0.501	2.22
Ligue 1 Midfielder	0.0087	0.0494	0.18	0.862	2.30
Premier League Defender	-0.0149	0.0614	-0.24	0.810	2.31
Premier League Midfielder	0.0535	0.0473	1.13	0.264	2.25

Regression Equation

LogMarkVal = 1.6546 - 0.0240 League_Bundesliga + 0.0762 League_La Liga
- 0.0586 League_Ligue
1 + 0.0641 League_Premier League - 0.0577 League_Serie A
- 0.1281 Position_Defender + 0.0003 Position_Midfielder
+ 0.1279 Position_Striker + 0.0390 League*Position_Bundesliga Defender
- 0.0133 League*Position_Bundesliga Midfielder
- 0.0257 League*Position_Bundesliga Striker - 0.0080 League*Position_La Liga Defender + 0.0035 League*Position_La Liga Midfielder
+ 0.0044 League*Position_La
Liga Striker - 0.0427 League*Position_Ligue 1 Defender
+ 0.0087 League*Position_Ligue 1 Midfielder + 0.0340 League*Position_Ligue 1 Striker - 0.0149 League*Position_Premier League Defender
+ 0.0535 League*Position_Premier League Midfielder
- 0.0387 League*Position_Premier League Striker + 0.0266 League*Position_Serie A
Defender - 0.0524 League*Position_Serie A Midfielder
+ 0.0259 League*Position_Serie A Striker

The interaction effect League*Position is still statistically insignificant with a p-value of 0.936. So once again, we should not consider the position to be an effective main effect, despite being statistically significant in the table above. The residuals plots now look like this:



The long right-tailed response variable is now somewhat taken care of. There is still the issue of heteroscedasticity. We have to talk about the two obvious outliers as indicated by the residual plots. These two refer to Neymar and Lionel Messi, both once teammates at FC Barcelona (club in La Liga). So what's so unusual about these two? They both have a valuation of more than a £100 million. In fact, Neymar made headlines last summer when he was bought by Paris Saint Germain, a French club in Ligue 1, for a record of £135 million. This became the most expensive trade in the history of soccer, with a lot of people criticising soccer clubs for making the sport too much money oriented. The median market value of players in this French League is £34.2 million and the Upper Quartile is £54 million. So Neymar, with a value of £135 million is an outlier for his League, Ligue 1, and also for his category as a Midfielder, which has a median market value of £42.7 million and Upper

Quartile of £59.7 million. The second outlier, Messi, is often regarded as the best soccer player to exist in the history of soccer with four FIFA Ballon d'Or awards (best player of the year award). So it's not surprising that his market price would be high. With £108 million, Messi is an outlier in his league (La Liga) where the median market value is £54 million and an Upper Quartile is £72 million. He is also a Midfielder, where his market value is also exceptionally high. So let's remove these two outliers and see if our two-way ANOVA model improves.

General Linear Model: LogMarkVal versus League, Position

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.18577	0.046442	5.27	0.002
Position	2	0.47495	0.237474	26.95	0.000
League*Position	8	0.06247	0.007808	0.89	0.536
Error	43	0.37887	0.008811		
Total	57	1.22477			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0938667	69.07%	58.99%	37.18%

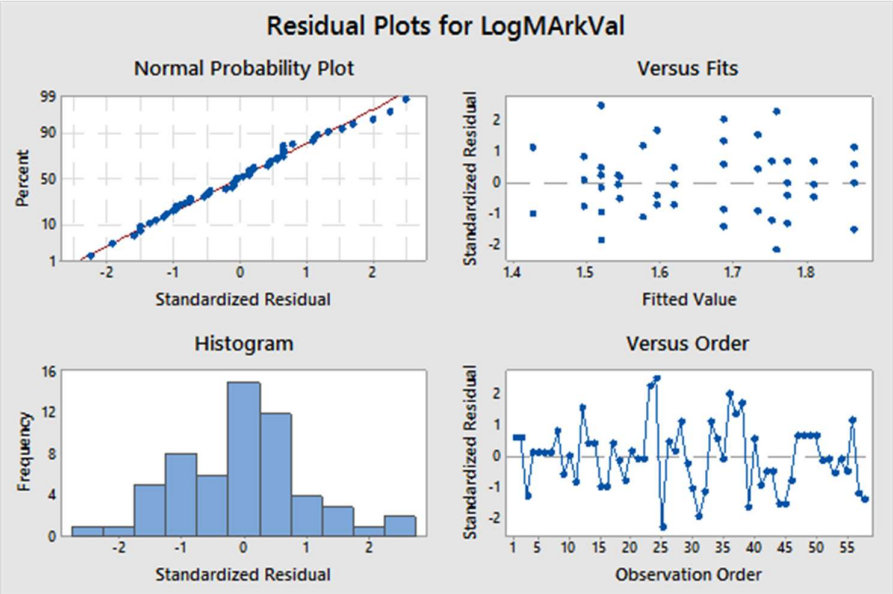
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6455	0.0133	123.82	0.000	
League					
Bundesliga	-0.0149	0.0258	-0.58	0.567	1.66
La Liga	0.0688	0.0238	2.89	0.006	1.60
Ligue 1	-0.0786	0.0294	-2.68	0.010	2.05
Premier League	0.0733	0.0274	2.68	0.011	1.96

Position					
Defender	-0.1190	0.0200	-5.93	0.000	1.36
Midfielder	-0.0181	0.0169	-1.07	0.291	1.34
League*Position					
Bundesliga Defender	0.0298	0.0385	0.78	0.442	2.18
Bundesliga Midfielder	0.0050	0.0349	0.14	0.886	2.20
La Liga Defender	-0.0005	0.0351	-0.01	0.988	1.96
La Liga Midfielder	-0.0113	0.0312	-0.36	0.718	1.98
Ligue 1 Defender	-0.0227	0.0444	-0.51	0.612	2.23
Ligue 1 Midfielder	-0.0314	0.0356	-0.88	0.382	2.29
Premier League Defender	-0.0240	0.0431	-0.56	0.580	2.31
Premier League Midfielder	0.0718	0.0333	2.16	0.037	2.26

Regression Equation

LogMarkVal = 1.6455 - 0.0149 League_Bundesliga + 0.0688 League_La Liga
- 0.0786 League_Ligue
1 + 0.0733 League_Premier League - 0.0486 League_Serie A
- 0.1190 Position_Defender - 0.0181 Position_Midfielder
+ 0.1370 Position_Striker + 0.0298 League*Position_Bundesliga Defender
+ 0.0050 League*Position_Bundesliga Midfielder
- 0.0349 League*Position_Bundesliga Striker - 0.0005 League*Position_La Liga Defender - 0.0113 League*Position_La Liga Midfielder
+ 0.0119 League*Position_La
Liga Striker - 0.0227 League*Position_Ligue 1 Defender
- 0.0314 League*Position_Ligue 1 Midfielder + 0.0541 League*Position_Ligue 1 Striker - 0.0240 League*Position_Premier League Defender
+ 0.0718 League*Position_Premier League Midfielder
- 0.0478 League*Position_Premier League Striker + 0.0174 League*Position_Serie A
Defender - 0.0341 League*Position_Serie A Midfielder
+ 0.0167 League*Position_Serie A Striker



The residual plots look a lot better now. There is a major improvement in the R^2 and R^2 pred. There is still an issue of heteroscedasticity, however not as serious as before. Although we see a general improvement, the interaction effect is still statistically insignificant. Maybe there is not an interaction and we should just look at the main effects. If there is no interaction effect, it indicates that the position effect is not different for different soccer leagues. So let's run the two-way ANOVA model again without the interaction effect. Since the market value of players still shows a long right-tailed histogram, we will stick to the semi-log model. We will put back our two outliers and see if they still appear as outlier once the model is run without the interaction effect. Here's two-way ANOVA model without the interaction coefficient:

General Linear Model: LogMarkVal versus Position, League

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Position	Fixed	3	Defender, Midfielder, Striker
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.25850	0.064625	4.00	0.007
Position	2	0.45823	0.229113	14.18	0.000
Error	53	0.85642	0.016159		
Lack-of-Fit	8	0.05142	0.006428	0.36	0.936
Pure Error	45	0.80500	0.017889		
Total	59	1.58486			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.127118	45.96%	39.84%	32.33%

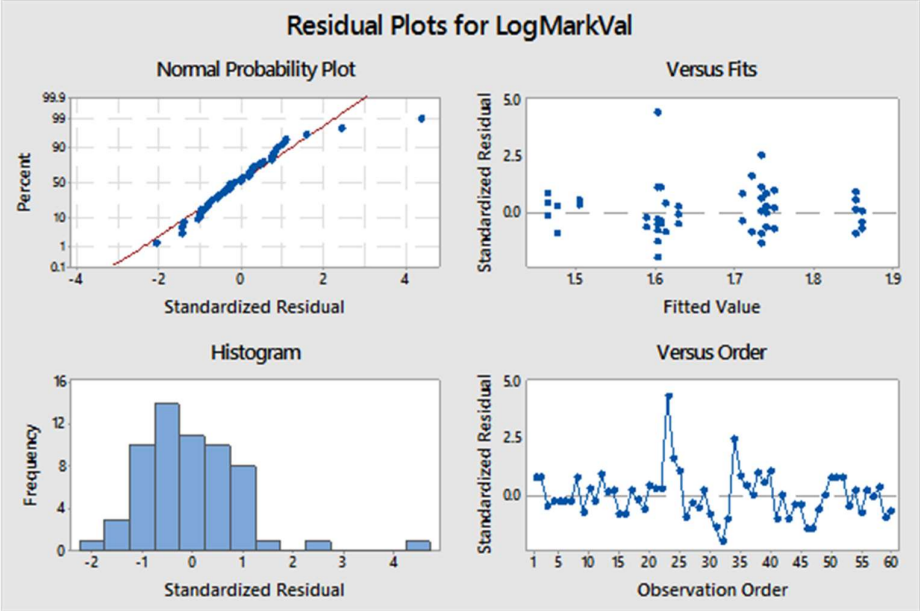
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6565	0.0175	94.73	0.000	
League					
Bundesliga	-0.0292	0.0343	-0.85	0.399	1.60
La Liga	0.0747	0.0304	2.46	0.017	1.47
Ligue 1	-0.0572	0.0343	-1.67	0.101	1.60
Premier League	0.0811	0.0330	2.45	0.017	1.55
Position					
Defender	-0.1250	0.0263	-4.76	0.000	1.28
Midfielder	0.0020	0.0223	0.09	0.930	1.32

Regression Equation

LogMarkVal = 1.6565 - 0.0292 League_Bundesliga + 0.0747 League_La Liga
- 0.0572 League_Ligue
1 + 0.0811 League_Premier League - 0.0694 League_Serie A
- 0.1250 Position_Defender + 0.0020 Position_Midfielder
+ 0.1230 Position_Striker

The main effects are now statistically significant with low p-values. There is a drop in the values corresponding to R^2 and R^2_{pred}. Let's take a look at our residual plots to check the assumptions, especially now that we have added back the outliers identified in the model with the interaction effect.



Here we see the same issue – the same data points showing up as the outliers. So we will remove Neymar and Messi again, with our response variable still logged, and run our two-way ANOVA model this way.

General Linear Model: LogMArkVal versus League, Position

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.29821	0.074552	8.62	0.000
Position	2	0.46647	0.233236	26.95	0.000

Error	51	0.44134	0.008654		
Lack-of-Fit	8	0.06247	0.007808	0.89	0.536
Pure Error	43	0.37887	0.008811		
Total	57	1.22477			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0930251	63.97%	59.73%	53.37%

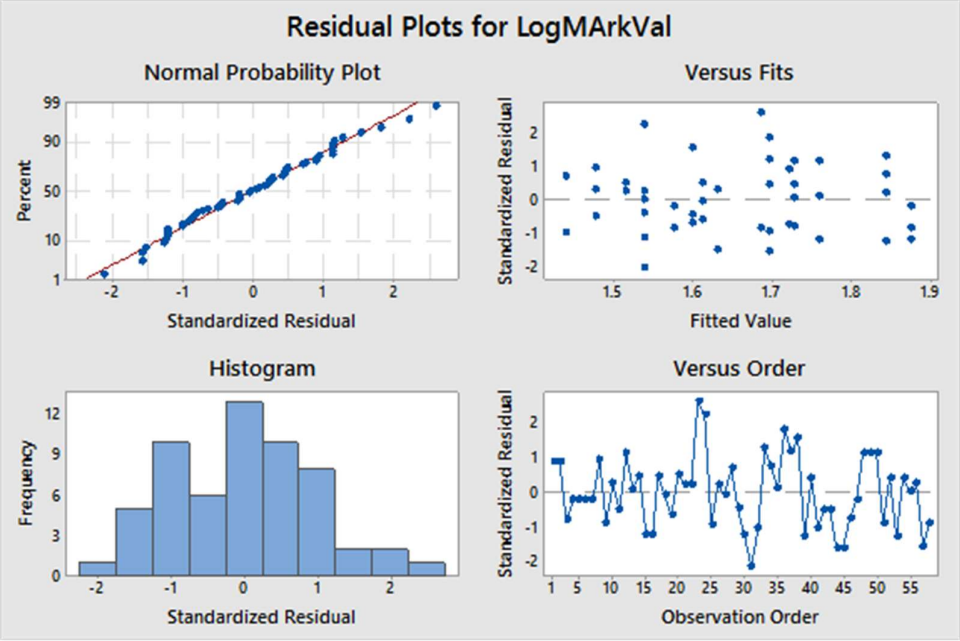
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6458	0.0129	127.70	0.000	
League					
Bundesliga	-0.0176	0.0252	-0.70	0.488	1.61
La Liga	0.0666	0.0229	2.91	0.005	1.51
Ligue 1	-0.0922	0.0260	-3.54	0.001	1.65
Premier League	0.0986	0.0243	4.06	0.000	1.57
Position					
Defender	-0.1150	0.0193	-5.96	0.000	1.29
Midfielder	-0.0163	0.0166	-0.98	0.330	1.32

Regression Equation

LogMArkVal = 1.6458 - 0.0176 League_Bundesliga + 0.0666 League_La Liga
- 0.0922 League_Ligue
1 + 0.0986 League_Premier League - 0.0554 League_Serie A
- 0.1150 Position_Defender - 0.0163 Position_Midfielder
+ 0.1313 Position_Striker

By removing the same outliers, we were able to obtain an improvement in the value of R^2. Now we see our main effects showing up even more statistically significant with a p-value that is zero to at least three decimal places. This implies that the league a player plays in and the position he plays on have a significant effect on his market value. Lack of interaction effect indicates that the effect of playing at a particular position does not vary differently per leagues. Let's take a look at our residual plots for this model:



The plots look a lot better once we have removed the outliers from this model (without the interaction effect). However, now there seem to be a few more outliers. The diagnostics of data with Neymar and Messi removed look like this:

Player	Market Value (Pounds m)	League	Position	HI_1	COOK_1
Paulo Dybala	63	Serie A	Striker	0.137216	0.018356
Gonzalo Higuain	63	Serie A	Striker	0.137216	0.018356
Mauro Icardi	45	Serie A	Striker	0.137216	0.014262
Lorenzo Insigne	36	Serie A	Midfielder	0.112179	0.000741
Marek Hamsik	36	Serie A	Midfielder	0.112179	0.000741
Radja	36	Serie A	Midfielder	0.112179	0.000741
Miralem Pjanic	36	Serie A	Midfielder	0.112179	0.000741

Leonardo Bonucci	36	Serie A	Defender	0.141343	0.020733
Ivan Perisic	31.5	Serie A	Midfielder	0.112179	0.01348
Alex Sandro	31.5	Serie A	Defender	0.141343	0.001667
Kalidou	27	Serie A	Defender	0.141343	0.006126
Lewandowski	72	Bundesliga	Striker	0.12663	0.026213
Emerick	58.5	Bundesliga	Striker	0.12663	0.000159
Naby Keita	45	Bundesliga	Midfielder	0.119875	0.004357
Thomas Muller	45	Bundesliga	Striker	0.12663	0.030979
Timo Werner	45	Bundesliga	Striker	0.12663	0.030979
James Rodriguez	45	Bundesliga	Midfielder	0.119875	0.004357
Christian Pulisic	40.5	Bundesliga	Midfielder	0.119875	5.08E-05
Thiago	36	Bundesliga	Midfielder	0.119875	0.007901
Mats Hummels	36	Bundesliga	Defender	0.14215	0.005919
Boateng	34.2	Bundesliga	Defender	0.14215	0.00138
Alaba	34.2	Bundesliga	Defender	0.14215	0.00138
Mbappe	81	Ligue 1	Striker	0.156613	0.181647
Verratti	54	Ligue 1	Midfielder	0.111438	0.088669
Cavani	40.5	Ligue 1	Striker	0.156613	0.021819
Di Maria	36	Ligue 1	Midfielder	0.111438	0.00084
Lucas	34.2	Ligue 1	Midfielder	0.111438	2.52E-05
Marquinhos	31.5	Ligue 1	Defender	0.16206	0.013575
Julian Braxler	31.5	Ligue 1	Midfielder	0.111438	0.003545
Thomas Lemar	27	Ligue 1	Midfielder	0.111438	0.026157
Adrien Rabiot	22.5	Ligue 1	Midfielder	0.111438	0.079861
Fabinho	22.5	Ligue 1	Defender	0.16206	0.028469
Ronaldo	90	La Liga	Striker	0.116214	0.030024
Suarez	81	La Liga	Striker	0.116214	0.010315
Griezmann	72	La Liga	Striker	0.116214	0.000458
Bale	72	La Liga	Midfielder	0.094772	0.049655
Toni Kroos	63	La Liga	Midfielder	0.094772	0.020364
Sergio Busquets	54	La Liga	Defender	0.120105	0.046689
Benzema	54	La Liga	Striker	0.116214	0.030421
Koke	54	La Liga	Midfielder	0.094772	0.00252
Ivan Rakitic	40.5	La Liga	Midfielder	0.094772	0.014991
Gerard Pique	36	La Liga	Defender	0.120105	0.004319
Sergio Ramos	36	La Liga	Defender	0.120105	0.004319
Isco	36	La Liga	Midfielder	0.094772	0.037295
Modric	36	La Liga	Midfielder	0.094772	0.037295
Marcelo	34.2	La Liga	Defender	0.120105	0.010276
Harry Kane	72	Premier League	Striker	0.133466	0.000996
De Bruyne	67.5	Premier League	Midfielder	0.095821	0.0198
Eden Hazard	67.5	Premier League	Midfielder	0.095821	0.0198

Paul Pogba	67.5	Premier League	Midfielder	0.095821	0.0198
Romelu Lukaku	63	Premier League	Striker	0.133466	0.017136
Coutinho	58.5	Premier League	Midfielder	0.095821	0.002945
Aguero	58.5	Premier League	Striker	0.133466	0.03461
Alexis Sanchez	58.5	Premier League	Midfielder	0.095821	0.002945
Dele Alli	54	Premier League	Midfielder	0.095821	3.5E-05
Kante	45	Premier League	Defender	0.149962	0.001935
Kyle Walker	31.5	Premier League	Defender	0.149962	0.05892
Mesut Ozil	45	Premier League	Midfielder	0.095821	0.010864

The guideline for the leverage points is (2.5)*(7)/58 = 0.30. So we don't see any leverage points in the diagnostics. Here we can see that Mbappe (Ligue 1 and Striker) shows up as an outlier. I conducted the diagnostics over and over again until no outlier appeared. Following Mbappe, Veratti appeared as an outlier. So I removed these two and at this point, I'm left with 56 soccer players. Before I list the two-way ANOVA model for these 56 players and the diagnostics, let's talk about why Mbappe and Veratti appear to be the outliers. Both of these stars play in Paris Saint Germain (Ligue 1). Mbappe is a striker whereas Veratti is a midfielder. Mbappe also made headlines a few months back when PSG signed him for £81 million, making him the youngest player (18 years old) and second most expensive one, next to Neymar who is valued £135 million. This number for Mbappe is unusually high for the Strikers in Ligue 1. As for Veratti, he plays as a midfielder in PSG. With a market value of £54 million, he exceeds the median for Ligue 1 – £34.2 million by a healthy margin. As for his position as a midfielder, after removing Neymar, the mean for this position has dropped down to £34.2 million, so Veratti is unusual for his position as a midfielder in Ligue 1. After removing these two players, the two-way ANOVA and the diagnostics look like this:

General Linear Model: LogMarkVal versus League, Position

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.38549	0.096373	14.38	0.000
Position	2	0.37707	0.188535	28.13	0.000
Error	49	0.32845	0.006703		
Lack-of-Fit	8	0.05034	0.006292	0.93	0.504
Pure Error	41	0.27811	0.006783		
Total	55	1.15128			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0818723	71.47%	67.98%	63.10%

Coefficients

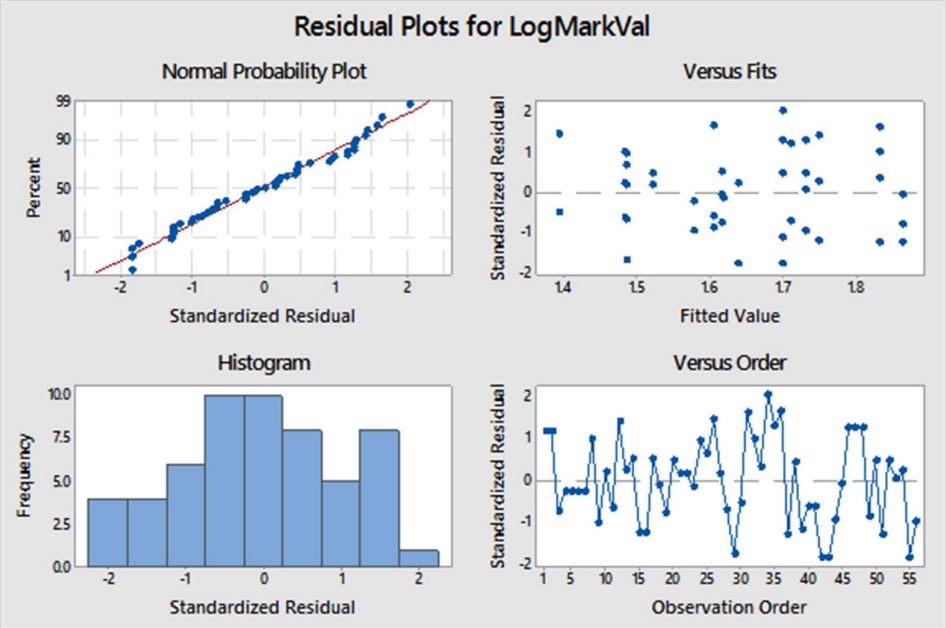
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6347	0.0117	140.09	0.000	
League					
Bundesliga	-0.0057	0.0223	-0.26	0.798	1.64
La Liga	0.0772	0.0203	3.80	0.000	1.53
Ligue 1	-0.1360	0.0253	-5.38	0.000	1.80
Premier League	0.1094	0.0216	5.08	0.000	1.59
Position					
Defender	-0.1070	0.0171	-6.26	0.000	1.26
Midfielder	-0.0132	0.0149	-0.89	0.380	1.30

Regression Equation

LogMarkVal = 1.6347 - 0.0057 League_Bundesliga + 0.0772 League_La Liga - 0.1360 League_Ligue 1 + 0.1094 League_Premier League - 0.0449 League_Serie A - 0.1070 Position_Defender - 0.0132 Position_Midfielder + 0.1202 Position_Striker

Means

Term	Fitted Mean	SE Mean
League		
Bundesliga	1.6289	0.0247
La Liga	1.7119	0.0220
Ligue 1	1.4987	0.0297
Premier League	1.7441	0.0243
Serie A	1.5898	0.0248
Position		
Defender	1.5277	0.0221
Midfielder	1.6215	0.0159
Striker	1.7549	0.0216

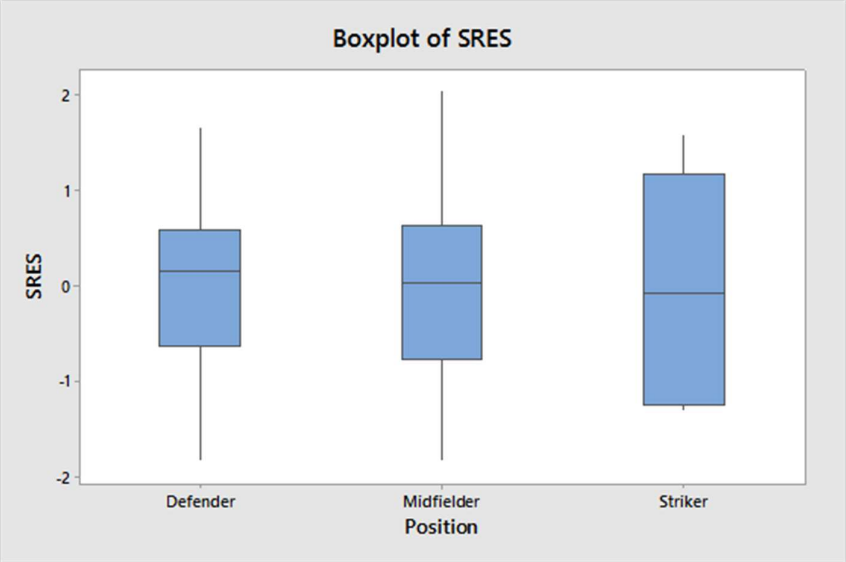


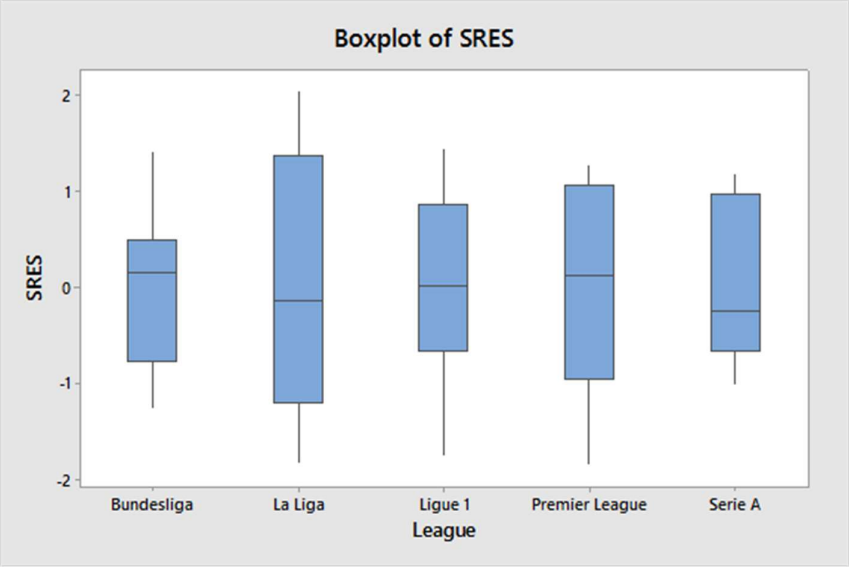
Player	Market Value (Pounds m)	League	Position	HI	COOK
Paulo Dybala	63	Serie A	Striker	0.140321	0.032316
Gonzalo Higuain	63	Serie A	Striker	0.140321	0.032316
Mauro Icardi	45	Serie A	Striker	0.140321	0.013038
Lorenzo Insigne	36	Serie A	Midfielder	0.112965	0.001263
Marek Hamsik	36	Serie A	Midfielder	0.112965	0.001263
Radja	36	Serie A	Midfielder	0.112965	0.001263
Miralem Pjanic	36	Serie A	Midfielder	0.112965	0.001263

Leonardo Bonucci	36	Serie A	Defender	0.141833	0.022186
Ivan Perisic	31.5	Serie A	Midfielder	0.112965	0.018765
Alex Sandro	31.5	Serie A	Defender	0.141833	0.00099
Kalidou	27	Serie A	Defender	0.141833	0.010852
Lewandowski	72	Bundesliga	Striker	0.128965	0.042413
Emerick	58.5	Bundesliga	Striker	0.128965	0.001177
Naby Keita	45	Bundesliga	Midfielder	0.121122	0.004683
Thomas Muller	45	Bundesliga	Striker	0.128965	0.03333
Timo Werner	45	Bundesliga	Striker	0.128965	0.03333
James Rodriguez	45	Bundesliga	Midfielder	0.121122	0.004683
Christian Pulisic	40.5	Bundesliga	Midfielder	0.121122	0.000232
Thiago	36	Bundesliga	Midfielder	0.121122	0.011822
Mats Hummels	36	Bundesliga	Defender	0.142848	0.004893
Boateng	34.2	Bundesliga	Defender	0.142848	0.000605
Alaba	34.2	Bundesliga	Defender	0.142848	0.000605
Cavani	40.5	Ligue 1	Striker	0.196898	0.000854
Di Maria	36	Ligue 1	Midfielder	0.136047	0.019441
Lucas	34.2	Ligue 1	Midfielder	0.136047	0.009125
Marquinhos	31.5	Ligue 1	Defender	0.181551	0.06563
Julian Braxler	31.5	Ligue 1	Midfielder	0.136047	0.000632
Thomas Lemar	27	Ligue 1	Midfielder	0.136047	0.011409
Adrien Rabiot	22.5	Ligue 1	Midfielder	0.136047	0.069102
Fabinho	22.5	Ligue 1	Defender	0.181551	0.009028
Ronaldo	90	La Liga	Striker	0.119184	0.048838
Suarez	81	La Liga	Striker	0.119184	0.019099
Griezmann	72	La Liga	Striker	0.119184	0.002083
Bale	72	La Liga	Midfielder	0.095635	0.062668
Toni Kroos	63	La Liga	Midfielder	0.095635	0.025214
Sergio Busquets	54	La Liga	Defender	0.120609	0.054017
Benzema	54	La Liga	Striker	0.119184	0.032552
Koke	54	La Liga	Midfielder	0.095635	0.00282
Ivan Rakitic	40.5	La Liga	Midfielder	0.095635	0.020773
Gerard Pique	36	La Liga	Defender	0.120609	0.007854
Sergio Ramos	36	La Liga	Defender	0.120609	0.007854
Isco	36	La Liga	Midfielder	0.095635	0.05057
Modric	36	La Liga	Midfielder	0.095635	0.05057
Marcelo	34.2	La Liga	Defender	0.120609	0.016703
Harry Kane	72	Premier League	Striker	0.136955	0.000188
De Bruyne	67.5	Premier League	Midfielder	0.096388	0.024362
Eden Hazard	67.5	Premier League	Midfielder	0.096388	0.024362
Paul Pogba	67.5	Premier League	Midfielder	0.096388	0.024362

Romelu Lukaku	63	Premier League	Striker	0.136955	0.016512
Coutinho	58.5	Premier League	Midfielder	0.096388	0.003307
Aguero	58.5	Premier League	Striker	0.136955	0.036944
Alexis Sanchez	58.5	Premier League	Midfielder	0.096388	0.003307
Dele Alli	54	Premier League	Midfielder	0.096388	5.61E-06
Kante	45	Premier League	Defender	0.150512	0.00116
Kyle Walker	31.5	Premier League	Defender	0.150512	0.085575
Mesut Ozil	45	Premier League	Midfielder	0.096388	0.015184

The model has improved generally. The R^2 has also increased significantly. The main effects have become increasingly statistically significant. In the diagnostics above, the guideline for the leverage points is (2.5)*(7)/56 = 0.31. Clearly, there is no leverage point in the diagnostic above. We cannot identify any clear outlier as well. In the residual plots, we see normality in our data points. We also have to consider the limitation of small data size that might be restricting us from getting perfect normality. On the other hand, the issue of heteroscedasticity persists. There is non-constant variance in our errors. The boxplots of standardised residuals against the types are as follows:





Just as the responses for the players from different leagues and on different positions might have different means, the errors for the players from different leagues and on different positions might have different variances. This is a violation of Ordinary Least Squares and carries many negative impacts on OLS analysis such as the coefficients are inefficient, confidence intervals do not have correct coverage, predictions and prediction intervals are not correct etc. Fear of our model suffering from these negative effects is why we should now consider the Levene's test.

General Linear Model: AbsSRES versus League, Position

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	2.7529	0.6882	2.48	0.056
Position	2	0.2530	0.1265	0.46	0.637
Error	49	13.6010	0.2776		
Lack-of-Fit	8	2.8959	0.3620	1.39	0.231
Pure Error	41	10.7051	0.2611		
Total	55	16.5725			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.526850	17.93%	7.88%	0.00%

The Levene's test indicates that League variable shows up as statistically significant, while the null hypothesis of constant variance could not be rejected for Position variable. So we need to consider the non-constant variance in the errors resulting due to the variable League. Once successfully identified, the non-constant variance could be addressed by the Weighted Least Squares. We would only need different weights for players in different leagues, so the defenders, strikers, and midfielder are all going to have same weights if they are in the same league. To get the weights, let get the standard deviations of the players in different leagues.

95% Bonferroni Confidence Intervals for Standard Deviations

League	N	StDev	CI
Bundesliga	11	0.54872	(0.323985, 1.21350)
La Liga	15	1.08650	(0.646627, 2.20410)
Ligue 1	11	1.80219	(0.633800, 6.69135)
Premier League	12	0.62036	(0.424565, 1.15419)
Serie A	11	0.55203	(0.348357, 1.14226)

Individual confidence level = 99%

Tests

Method	Test Statistic	P-Value
Multiple comparisons	—	0.239
Levene	2.50	0.053

The weight for players in any given league would be $1/(StDev)^2$. So I took each of the standard deviation, squared it, and divided one by that term. I stored the results in a column named 'wt' and carried out the two-way ANOVA model as a WLS instead of OLS. I added again the interaction effect to see how the WLS would assess it now. Also, the model was run on all 60 players, including the OLS outliers, since a player might not be an outlier any more relative to a higher standard deviation.

General Linear Model: LogMarkVal versus League, Position

Method

Factor coding	(-1, 0, +1)
Weights	wt

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.3078	0.07694	5.27	0.001
Position	2	0.3630	0.18148	12.42	0.000

League*Position	8	0.1224	0.01530	1.05	0.416
Error	45	0.6576	0.01461		
Total	59	2.1301			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.120886	69.13%	59.52%	45.16%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6546	0.0186	88.90	0.000	
League					
Bundesliga	-0.0240	0.0243	-0.99	0.328	2.94
La Liga	0.0762	0.0329	2.32	0.025	3.28
Ligue 1	-0.0586	0.0628	-0.93	0.356	8.24
Premier League	0.0641	0.0267	2.40	0.020	3.27
Position					
Defender	-0.1281	0.0282	-4.54	0.000	3.46
Midfielder	0.0003	0.0225	0.01	0.991	3.02
League*Position					
Bundesliga Defender	0.0390	0.0365	1.07	0.291	3.91
Bundesliga Midfielder	-0.0133	0.0312	-0.43	0.672	3.48
La Liga Defender	-0.0080	0.0489	-0.16	0.871	4.33
La Liga Midfielder	0.0035	0.0416	0.09	0.932	4.21
Ligue 1 Defender	-0.0427	0.0954	-0.45	0.657	13.03
Ligue 1 Midfielder	0.0087	0.0738	0.12	0.907	10.74
Premier League Defender	-0.0149	0.0415	-0.36	0.722	3.85
Premier League Midfielder	0.0535	0.0321	1.67	0.103	3.68

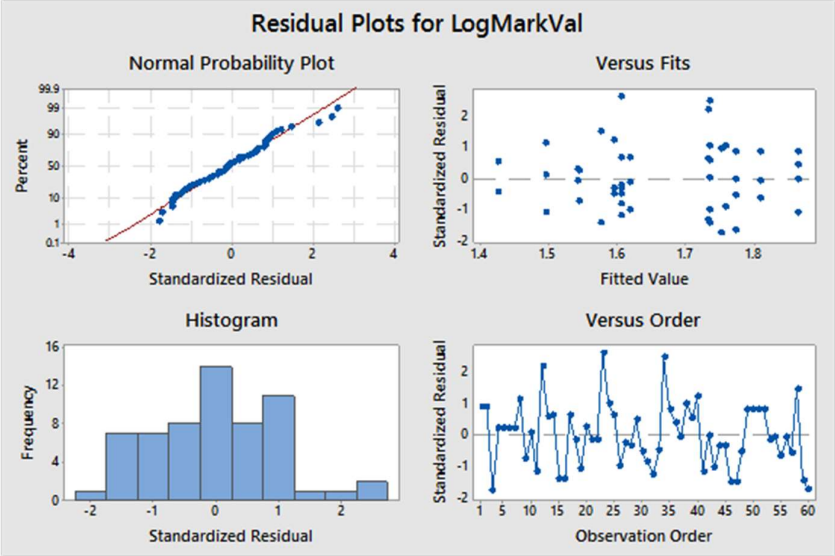
Regression Equation

LogMarkVal = 1.6546 - 0.0240 League_Bundesliga + 0.0762 League_La Liga
- 0.0586 League_Ligue
1 + 0.0641 League_Premier League - 0.0577 League_Serie A
- 0.1281 Position_Defender + 0.0003 Position_Midfielder
+ 0.1279 Position_Striker + 0.0390 League*Position_Bundesliga Defender
- 0.0133 League*Position_Bundesliga Midfielder
- 0.0257 League*Position_Bundesliga Striker - 0.0080 League*Position_La Liga Defender + 0.0035 League*Position_La Liga Midfielder
+ 0.0044 League*Position_La
Liga Striker - 0.0427 League*Position_Ligue 1 Defender
+ 0.0087 League*Position_Ligue 1 Midfielder + 0.0340 League*Position_Ligue 1 Striker - 0.0149 League*Position_Premier League Defender

+ 0.0535 League*Position_Premier League Midfielder
- 0.0387 League*Position_Premier League Striker + 0.0266 League*Position_Serie A
Defender - 0.0524 League*Position_Serie A Midfielder
+ 0.0259 League*Position_Serie A Striker

Means

Term	Fitted Mean	SE Mean
League		
Bundesliga	1.6306	0.0202
La Liga	1.7308	0.0350
Ligue 1	1.5961	0.0774
Premier League	1.7188	0.0247
Serie A	1.5969	0.0207
Position		
Defender	1.5265	0.0367
Midfielder	1.6549	0.0219
Striker	1.7825	0.0360
League*Position		
Bundesliga Defender	1.5415	0.0383
Bundesliga Midfielder	1.6175	0.0332
Bundesliga Striker	1.7327	0.0332
La Liga Defender	1.5948	0.0656
La Liga Midfielder	1.7347	0.0496
La Liga Striker	1.8631	0.0656
Ligue 1 Defender	1.425	0.154
Ligue 1 Midfielder	1.6050	0.0821
Ligue 1 Striker	1.758	0.154
Premier League Defender	1.5758	0.0530
Premier League Midfielder	1.7725	0.0283
Premier League Striker	1.8079	0.0433
Serie A Defender	1.4953	0.0385
Serie A Midfielder	1.5447	0.0299
Serie A Striker	1.7506	0.0385



The interaction effect, again, is statistically insignificant so we will just conclude that there is no interaction effect. In the residual plots above, we see heteroscedasticity somewhat taken care of. However, in this model, we have outliers that could be the OLS outliers or new ones. So let's take a look at the diagnostics to see which players show up as outliers.

Player	Market Value (Pounds m)	League	Position	HI_1	COOK_1
Paulo Dybala	63	Serie A	Striker	0.161888707	0.066702
Gonzalo Higuain	63	Serie A	Striker	0.161888707	0.066702
Mauro Icardi	45	Serie A	Striker	0.161888707	0.018925
Lorenzo Insigne	36	Serie A	Midfielder	0.124763806	0.003832
Marek Hamsik	36	Serie A	Midfielder	0.124763806	0.003832
Radja	36	Serie A	Midfielder	0.124763806	0.003832
Miralem Pjanic	36	Serie A	Midfielder	0.124763806	0.003832
Leonardo Bonucci	36	Serie A	Defender	0.170771781	0.049326
Ivan Perisic	31.5	Serie A	Midfielder	0.124763806	0.037621
Alex Sandro	31.5	Serie A	Defender	0.170771781	0.003484
Kalidou	27	Serie A	Defender	0.170771781	0.016699
Lewandowski	72	Bundesliga	Striker	0.146294313	0.082617
Emerick	58.5	Bundesliga	Striker	0.146294313	0.003373
Naby Keita	45	Bundesliga	Midfielder	0.137413139	0.005145
Thomas Muller	45	Bundesliga	Striker	0.146294313	0.053694
Timo Werner	45	Bundesliga	Striker	0.146294313	0.053694

James Rodriguez	45	Bundesliga	Midfielder	0.137413139	0.005145
Christian Pulisic	40.5	Bundesliga	Midfielder	0.137413139	0.001592
Thiago	36	Bundesliga	Midfielder	0.137413139	0.027123
Mats Hummels	36	Bundesliga	Defender	0.172594343	0.012143
Boateng	34.2	Bundesliga	Defender	0.172594343	0.002181
Alaba	34.2	Bundesliga	Defender	0.172594343	0.002181
Neymar	135	Ligue 1	Midfielder	0.092331172	0.095738
Mbappe	81	Ligue 1	Striker	0.099582741	0.013407
Verratti	54	Ligue 1	Midfielder	0.092331172	0.006135
Cavani	40.5	Ligue 1	Striker	0.099582741	0.004519
Di Maria	36	Ligue 1	Midfielder	0.092331172	0.000573
Lucas	34.2	Ligue 1	Midfielder	0.092331172	0.001359
Marquinhos	31.5	Ligue 1	Defender	0.100758123	1.82E-05
Julian Braxler	31.5	Ligue 1	Midfielder	0.092331172	0.003319
Thomas Lemar	27	Ligue 1	Midfielder	0.092331172	0.00931
Adrien Rabiot	22.5	Ligue 1	Midfielder	0.092331172	0.020297
Fabinho	22.5	Ligue 1	Defender	0.100758123	0.007261
Messi	108	La Liga	Midfielder	0.075054399	0.067619
Ronaldo	90	La Liga	Striker	0.085359352	0.009451
Suarez	81	La Liga	Striker	0.085359352	0.003051
Griezmann	72	La Liga	Striker	0.085359352	6.89E-05
Bale	72	La Liga	Midfielder	0.075054399	0.012149
Toni Kroos	63	La Liga	Midfielder	0.075054399	0.003707
Sergio Busquets	54	La Liga	Defender	0.087722755	0.010673
Benzema	54	La Liga	Striker	0.085359352	0.011307
Koke	54	La Liga	Midfielder	0.075054399	1.54E-05
Ivan Rakitic	40.5	La Liga	Midfielder	0.075054399	0.010479
Gerard Pique	36	La Liga	Defender	0.087722755	0.003709
Sergio Ramos	36	La Liga	Defender	0.087722755	0.003709
Isco	36	La Liga	Midfielder	0.075054399	0.021283
Modric	36	La Liga	Midfielder	0.075054399	0.021283
Marcelo	34.2	La Liga	Defender	0.087722755	0.006671
Harry Kane	72	Premier League	Striker	0.144460803	2.15E-06
De Bruyne	67.5	Premier League	Midfielder	0.099009687	0.026702
Eden Hazard	67.5	Premier League	Midfielder	0.099009687	0.026702
Paul Pogba	67.5	Premier League	Midfielder	0.099009687	0.026702
Romelu Lukaku	63	Premier League	Striker	0.144460803	0.016374
Coutinho	58.5	Premier League	Midfielder	0.099009687	0.002956
Aguero	58.5	Premier League	Striker	0.144460803	0.039916
Alexis Sanchez	58.5	Premier League	Midfielder	0.099009687	0.002956
Dele Alli	54	Premier League	Midfielder	0.099009687	4.38E-05
Kante	45	Premier League	Defender	0.167213648	0.003279
Kyle Walker	31.5	Premier League	Defender	0.167213648	0.10552
Mesut Ozil	45	Premier League	Midfielder	0.099009687	0.021181

Now the outliers are Neymar (Cook's Distance 0.096), Kyle Walker (Cook's Distance 0.11), and Lewandowski (Cook's Distance 0.08). They all belong to different leagues: Neymar in Ligue 1, Lewandowski in Bundesliga, and Kyle Walker in Premier League. After I removed them and ran the two-way ANOVA model on WLS again, other outliers showed up. Therefore, I repeated the process of removing the outliers and conducting the regression until there was no significant outlier. The outliers, in addition to Neymar, Lewandowski, and Kyle Walker, which I removed are Messi, Dybala, and Higuain. Neymar is an outlier due to the extremely high market value. Lewandowski showed up because his market value of £72 million is way above the median value of strikers in Bundesliga. Kyle Walker is a defender, and is one of the most valued defenders in the Premier League, which makes him an outlier. As for Dybala and Higuain, they both are way above the median market value of the strikers in Serie A. Finally, Messi also values way above the median value of the players in La Liga. Our six outliers belong to each league in our model. Here are the regression results without these outliers and without the interaction effect:

General Linear Model: LogMarkVal versus League, Position

Method

Factor coding	(-1, 0, +1)
Weights	wt

Factor Information

Factor	Type	Levels	Values
League	Fixed	5	Bundesliga, La Liga, Ligue 1, Premier League, Serie A
Position	Fixed	3	Defender, Midfielder, Striker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
League	4	0.65615	0.164038	20.26	0.000
Position	2	0.32902	0.164511	20.32	0.000
Error	47	0.38060	0.008098		
Lack-of-Fit	8	0.05795	0.007244	0.88	0.545
Pure Error	39	0.32265	0.008273		

Total	53	1.59622

Model Summary

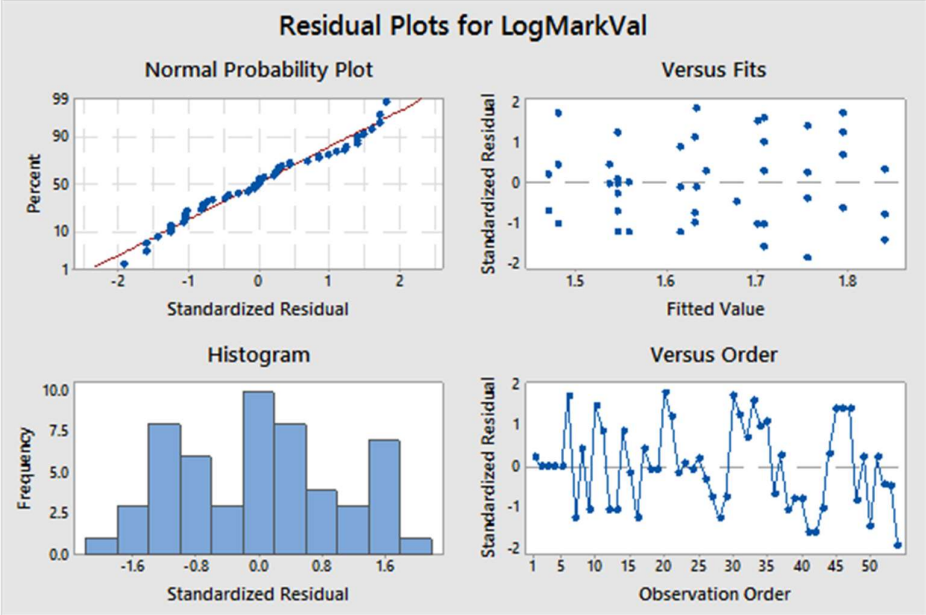
S	R-sq	R-sq(adj)	R-sq(pred)
0.0899887	76.16%	73.11%	69.06%

Coefficients

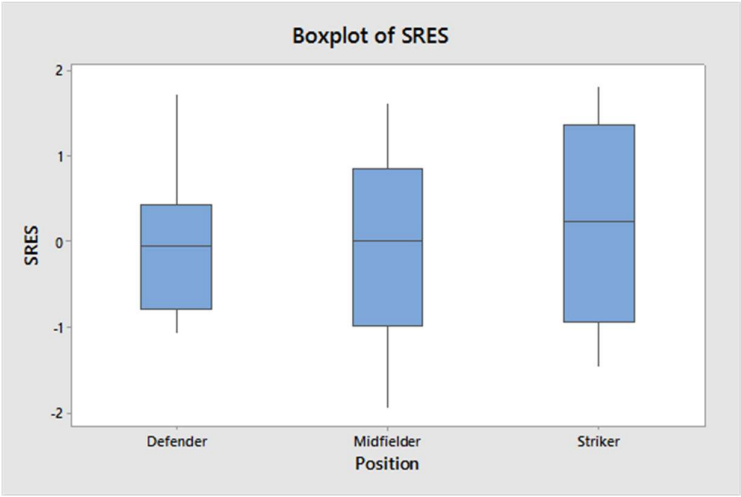
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.6386	0.0132	123.75	0.000	
League					
Bundesliga	-0.0215	0.0177	-1.21	0.232	2.43
La Liga	0.0719	0.0240	3.00	0.004	2.73
Ligue 1	-0.0904	0.0416	-2.17	0.035	5.58
Premier League	0.1192	0.0185	6.43	0.000	2.47
Position					
Defender	-0.0804	0.0141	-5.71	0.000	1.26
Midfielder	-0.0028	0.0119	-0.24	0.814	1.25

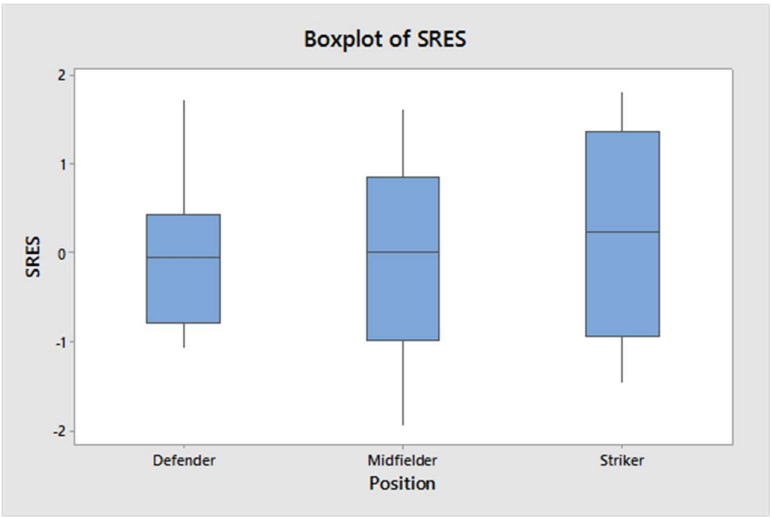
Regression Equation

LogMarkVal = 1.6386 - 0.0215 League_Bundesliga + 0.0719 League_La Liga
- 0.0904 League_Ligue
1 + 0.1192 League_Premier League - 0.0792 League_Serie A
- 0.0804 Position_Defender - 0.0028 Position_Midfielder
+ 0.0833 Position_Striker



The main effects in our model are highly statistically significant with p-values of zero at least to three decimal places. There is no physical interpretation of R^2 in WLS. As for the assumptions, we see normality in our data points, though there are a few points that are off the fitted line, we should be mindful of the small size of the dataset. The issue of heteroscedasticity is now taken care of as indicated by the residuals versus fitted plot. Let's finally take a look at the boxplots of standardized residuals versus the predictors.





The boxplots also indicate that the non-constant variance is now somewhat taken care of. Now that we have conducted our regression with only main effects, we should look at the multiple comparisons. We will use Tukey's pairwise comparisons to assess the difference

Comparisons for LogMarkVal

Tukey Pairwise Comparisons: League

Grouping Information Using the Tukey Method and 95% Confidence

League	N	Mean	Grouping	
Premier League	11	1.74457	A	
La Liga	14	1.71424	A	
Bundesliga	10	1.61673		B
Ligue 1	10	1.56688	A	B
Serie A	9	1.56441		B

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of League Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
La Liga - Bundesliga	0.0975	0.0312	(0.0083, 0.1867)	3.12	0.026
Ligue 1 - Bundesliga	-0.0498	0.0609	(-0.2239, 0.1242)	-0.82	0.923
Premier League - Bundesliga	0.1278	0.0279	(0.0483, 0.2074)	4.59	0.000
Serie A - Bundesliga	-0.0523	0.0261	(-0.1270, 0.0224)	-2.00	0.285
Ligue 1 - La Liga	-0.1474	0.0647	(-0.3321, 0.0373)	-2.28	0.173
Premier League - La Liga	0.0303	0.0353	(-0.0704, 0.1310)	0.86	0.910
Serie A - La Liga	-0.1498	0.0339	(-0.2467, -0.0529)	-4.42	0.001
Premier League - Ligue 1	0.1777	0.0631	(-0.0026, 0.3579)	2.82	0.055
Serie A - Ligue 1	-0.0025	0.0624	(-0.1806, 0.1757)	-0.04	1.000
Serie A - Premier League	-0.1802	0.0308	(-0.2683, -0.0920)	-5.84	0.000

Individual confidence level = 99.32%

Tukey Pairwise Comparisons: Position

Grouping Information Using the Tukey Method and 95% Confidence

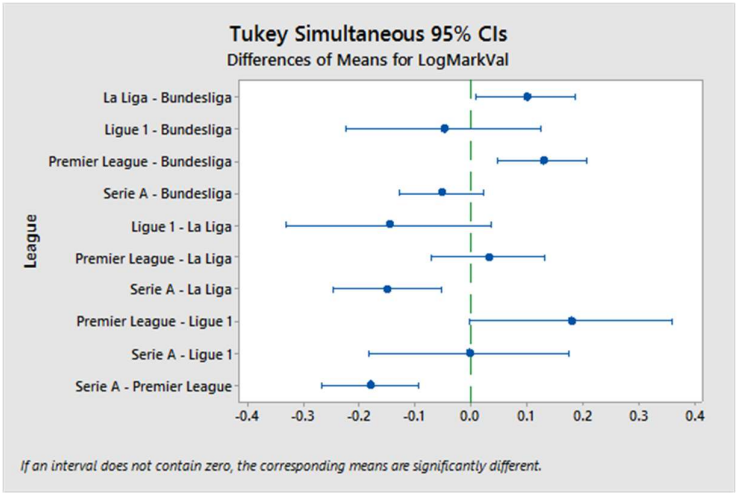
Position	N	Mean	Grouping		
Striker	13	1.75469	A		
Midfielder	28	1.62742		B	
Defender	13	1.54200			C

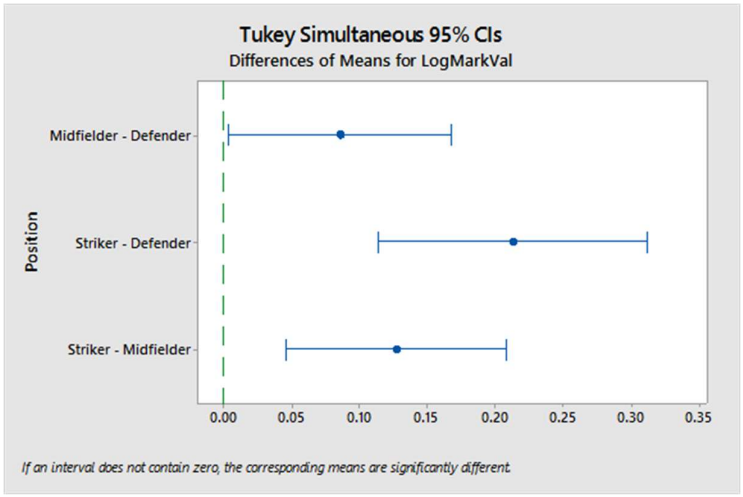
Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of Position Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
Midfielder - Defender	0.0854	0.0336	(0.0034, 0.1675)	2.54	0.039
Striker - Defender	0.2127	0.0404	(0.1141, 0.3113)	5.26	0.000
Striker - Midfielder	0.1273	0.0334	(0.0459, 0.2087)	3.81	0.001

Individual confidence level = 98.06%





We can clearly see the grouping in both the main effects. For leagues, there are two distinct groups with relation to logged market value. For positions, all three fall under different groups. Premier League and La Liga have relatively higher means for the logged market

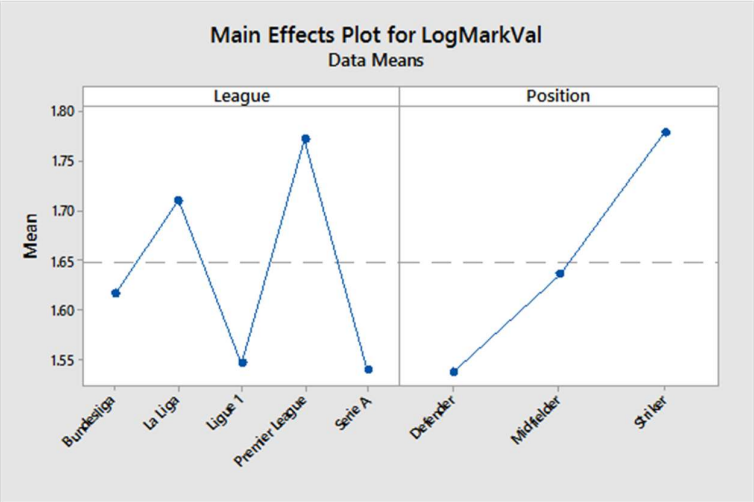
value of the players, while Bundesliga, Serie A, and Ligue 1 have relatively lower means for the logged market value of the players. All the pairwise comparisons that are statistically significant are between the leagues from these two groups. As for the position main effect, all the pairwise comparisons are statistically significant, indicating that the difference of means is significant between each group.

Means

Term	Fitted Mean	SE Mean
League		
Bundesliga	1.6171	0.0157
La Liga	1.7104	0.0261
Ligue 1	1.5482	0.0513
Premier League	1.7577	0.0178
Serie A	1.5594	0.0173
Position		
Defender	1.5581	0.0203
Midfielder	1.6358	0.0153
Striker	1.7218	0.0205

The fitted logged market value for players in Bundesliga is 1.6171, which equates to $10^{1.6171}$ = £41.4 million. For players in La Liga, it is 1.7104, which equates to $10^{1.7104}$ = £51.33 million. For players in Ligue 1, it is 1.5482, which equates to $10^{1.5482}$ = £35.33 million. For players in Premier League, it is 1.7577, which equates to $10^{1.7577}$ = £57.2 million, and for players in Serie A, it is 1.5594, which equates to $10^{1.5594}$ = £36.26 million.

The fitted logged market value for the players who play as strikers is 1.7218, which equates to $10^{1.7218}$ = £52.7 million. For Midfielders, it is 1.6358, which equates to $10^{1.6358}$ = £43.23 million. And for the Defenders, it is 1.5581, which equates to £36.14 million. Here are the main effect plots to visualize the main effects.



We can clearly see that Premier League (UK) and La Liga (Spain) have an average of player's market value that's more than the average market value of the players in Serie A (Italy), Ligue 1 (France), and Bundesliga. It is also evident from the plots that Strikers are valued the most, followed by the Midfielders, which are followed by the Defenders.

Premier League and La Liga are, in fact, the most talked about leagues in the soccer world. With the viewership of 2.7 billion, Premier League is the most watched League in the world. It has produced players like David Beckham, Cristiano Ronaldo, Alan Shearer etc. that are considered among the world's best players. And naturally, the more viewership a league has, the more fans its players get. As for La Liga, it contains the two most successful teams in Europe: FC Barcelona and Real Madrid FC. These two have won the most number of Champions League titles than any other European club. And this is why these clubs are really famous in the soccer world, attracting famous soccer stars like Kaka, Ronaldo, Ronaldinho,

Zidane, Messi etc. in the past. As for the Position, it is not surprising to see Striker over Midfielders and Midfielders over Defenders. This is because the player who scores most goals is usually a striker and since goals win matches, goals also win that player value and fans. So the closer the player is positioned to the opponent's goal area, the higher his market value. There is an interesting thing to note that might also be a limitation in my project: a player's market value might also have a relation with the 'hype' around that player, which could be estimated from the number of fans that player has. The more popular a player is, the higher his value is regardless of the position he plays at. As we noticed earlier that a few players have a really high market value, making them appear as outliers with respect to their league and position, it turns out they all have one of the largest fan bases (from Twitter). For the future, I would try to include the variable for the number of fans to see how that plays out as an effect in the presence of League and Position effect. This project was just a basic attempt to see how the market value of a player is evaluated considering his position and the league he plays in. However, in reality, a lot of factors come into play, making this problem a complicated one. To better understand the market value of the players, not only do I look forward to more predictors but also to a larger dataset.

Works Cited

"Comparing Europe's Top 5 Leagues - Which is currently the best?" *Sportskeeda.com – Get Latest Sports news & updates*, 8 Mar. 2017

"Most valuable players." *Transfermarkt*, www.transfermarkt.co.uk/spieler-statistik/wertvollstespieler/marktwertetop.

Settimi, Christina. "Neymar Signs With PSG, Set To Become World's Highest-Paid Soccer Player Ahead Of Messi And Ronaldo." *Forbes*, Forbes Magazine, 4 Aug. 2017,