**Estimating the Market value of Soccer players based on their club and playing position**

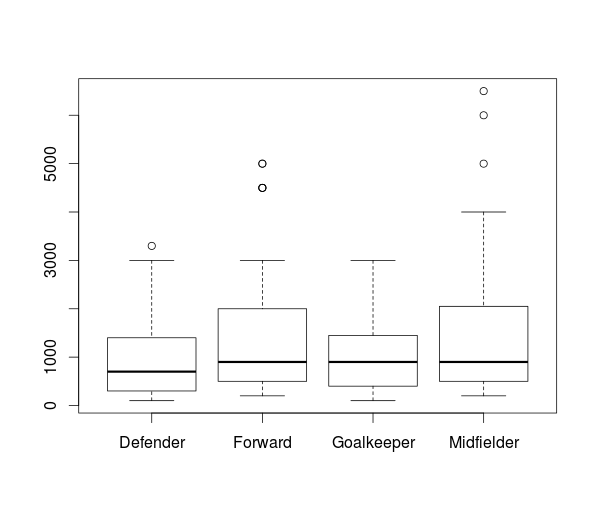
In this analysis I am trying to estimate the market price of the soccer players based on their playing position and the club which they represent.I have collected data from <http://www.transfermarkt.com> .The data contains of 198 players who are playing for different clubs in English Premiere League(First Division league in England).The Playing position of any player are classified into below four positions

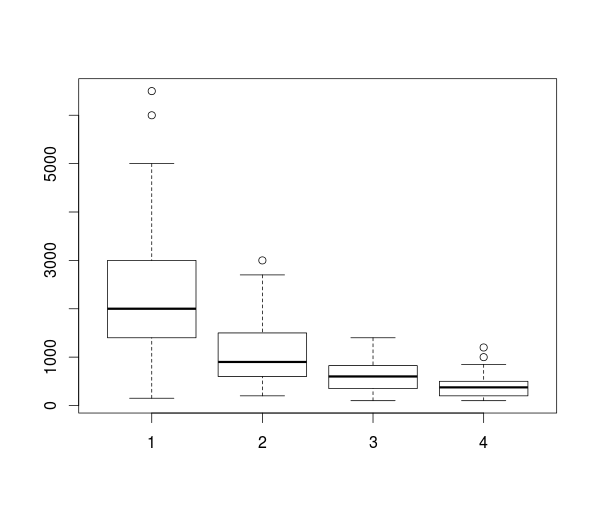
1. Forward
2. Midfielder
3. Defender
4. GoalKeeper

Any football team will generally have 1-goalkeeper,3-5 Defenders,4-5 Midfielders and 1-3 Forwards.The Player can belong to 4 categories of teams (Team-1, Team-2, Team-3, Team-4 ).The Team categories are formed based on the market value of the team,The teams with higher market values fall in team-1 and lower in team-4.

Based on these two categorical predictor variables we will try to estimate the target variable which is the market price of the player,which is considered to be in units of Million Euros.

The analysis will give the main effects and interaction effects of teams and position of the player.Do the teams value players more who play at a particular position.Below are the side to side box plots of Teams and Positions





We can see from the boxplot midfielders and forwards have higher market price then defenders or goalkeepers. The variability is higher in the price value of midfielders than others. The Teams side by side box plot clearly shows the players belonging to Team-1 category has higher Market price and variability followed by the other teams.

*POS  
TEAM Defender Forward Goalkeeper Midfielder  
 1 20 10 4 31  
 2 16 11 5 22  
 3 10 9 4 16  
 4 12 8 2 18*

The above table shows that we don't have any gaps in the data and we will can test for the interaction between these two groups.

Trying to fit in an Anova model with Price as target variable and team and position as prediction variables we get the below results

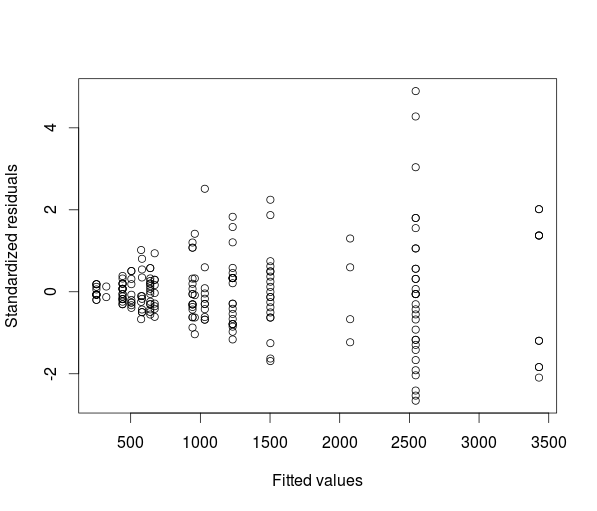
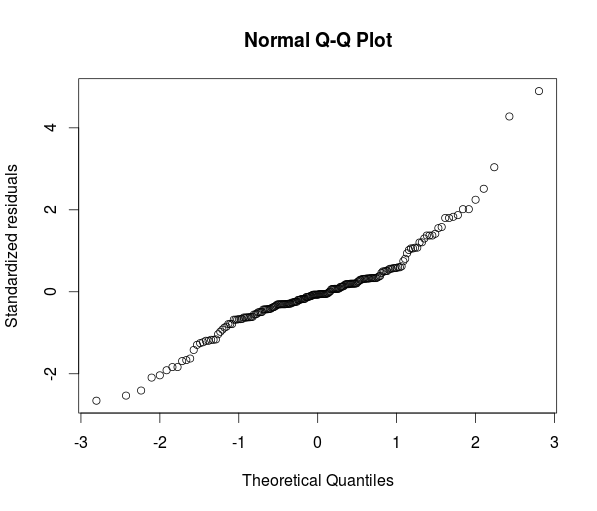
*Anova(fit,type = 3)  
Anova Table (Type III tests)  
  
Response: PRICE  
 Sum Sq Df F value Pr(>F)   
(Intercept) 45150125 1 66.9823 4.628e-14 \*\*\*  
POS 27490197 3 13.5943 4.836e-08 \*\*\*  
TEAM 13837200 3 6.8427 0.0002151 \*\*\*  
POS:TEAM 15296786 9 2.5215 0.0095638 \*\*   
Residuals 122678994 182   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Call:  
lm(formula = PRICE ~ POS + TEAM + POS \* TEAM)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2145.2 -343.0 -54.2 266.0 3954.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1502.5 183.6 8.184 4.63e-14 \*\*\*  
POSForward 1927.5 318.0 6.062 7.58e-09 \*\*\*  
POSGoalkeeper 572.5 449.7 1.273 0.20460   
POSMidfielder 1042.7 235.5 4.428 1.64e-05 \*\*\*  
TEAM2 -558.8 275.4 -2.029 0.04391 \*   
TEAM3 -997.5 318.0 -3.137 0.00199 \*\*   
TEAM4 -1248.3 299.8 -4.164 4.82e-05 \*\*\*  
POSForward:TEAM2 -1839.4 452.2 -4.067 7.07e-05 \*\*\*  
POSGoalkeeper:TEAM2 -556.2 615.8 -0.903 0.36753   
POSMidfielder:TEAM2 -754.6 358.1 -2.107 0.03645 \*   
POSForward:TEAM3 -1760.3 493.4 -3.568 0.00046 \*\*\*  
POSGoalkeeper:TEAM3 -502.5 661.9 -0.759 0.44874   
POSMidfielder:TEAM3 -907.0 406.2 -2.233 0.02676 \*   
POSForward:TEAM4 -1600.4 491.5 -3.256 0.00135 \*\*   
POSGoalkeeper:TEAM4 -501.7 771.6 -0.650 0.51642   
POSMidfielder:TEAM4 -855.2 386.1 -2.215 0.02801 \*   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 821 on 182 degrees of freedom  
Multiple R-squared: 0.5521, Adjusted R-squared: 0.5152   
F-statistic: 14.96 on 15 and 182 DF, p-value: < 2.2e-16*

The P-values signify that the two main effects as well as the interaction effects are statistically significant.

The R-squared value is 55%.The standard error of 821,says that the Market price of the player can be estimated at confidence interval of 95% between[-1600,1600]Million Euros.

Performing Diagnosis we get the below plots.

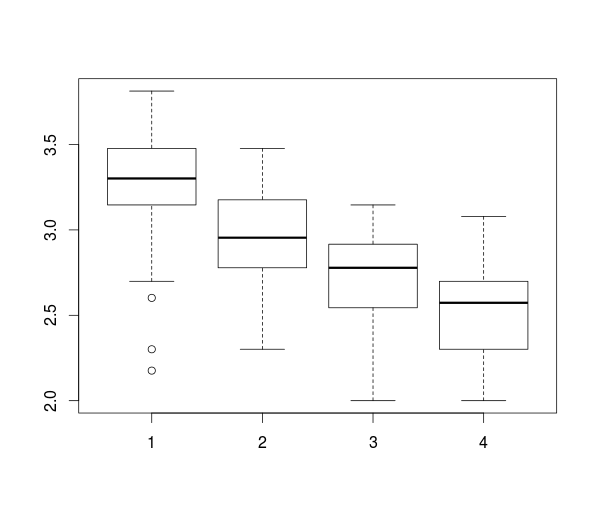


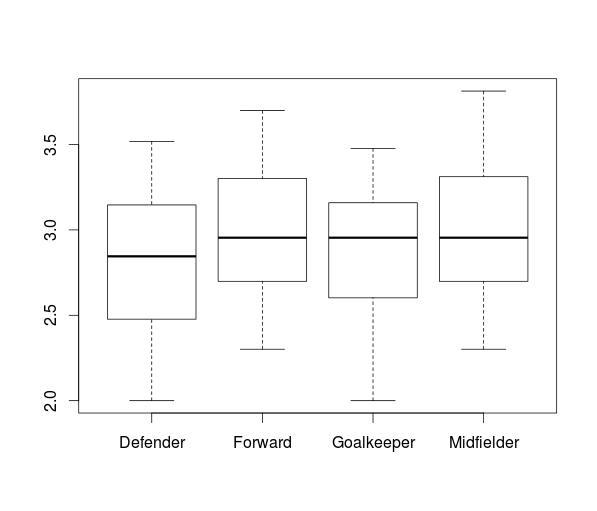
We see a problem of non-constant variance as well as the residuals are not normally distributed.We can check for non-constant variance by performing levene’s test by using absolute standard residuals.

*Anova Table (Type III tests)  
  
Response: abs(fit.diag$std.res)  
 Sum Sq Df F value Pr(>F)   
(Intercept) 10.425 1 28.2602 3.081e-07 \*\*\*  
POS 7.613 3 6.8792 0.0002052 \*\*\*  
TEAM 3.001 3 2.7117 0.0464222 \*   
POS:TEAM 4.483 9 1.3501 0.2140618   
Residuals 67.140 182   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

The main effects show that there is non-constant variance in this model.We try to address the effect of non-constant variance by taking logged value of prices.

Below are the side to side box plots after taking logged price values.





We can clearly see much lesser variability in Market price after taking the logged values.

Performing the Anova on the two main effects and interaction effects we get the below results.

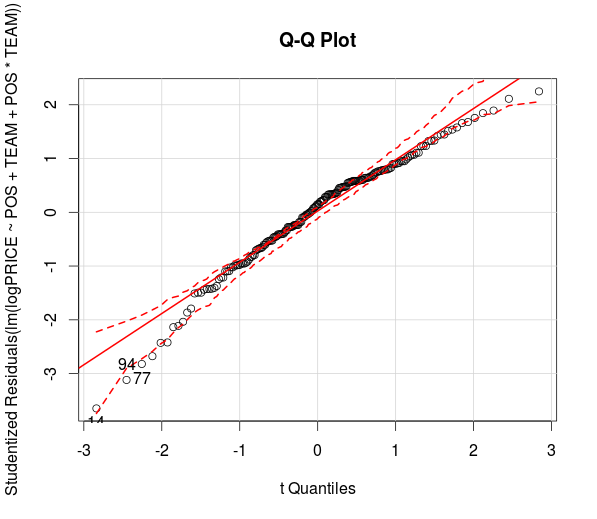
*Anova(fit,type=3)  
Anova Table (Type III tests)  
  
Response: logPRICE  
 Sum Sq Df F value Pr(>F)   
(Intercept) 190.623 1 2717.6121 < 2.2e-16 \*\*\*  
POS 1.289 3 6.1252 0.0005436 \*\*\*  
TEAM 4.407 3 20.9433 1.059e-11 \*\*\*  
POS:TEAM 0.620 9 0.9820 0.4564485   
Residuals 12.766 182   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Call:  
lm(formula = logPRICE ~ POS + TEAM + POS \* TEAM)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.91117 -0.15324 0.03859 0.16817 0.52817   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.08726 0.05922 52.131 < 2e-16 \*\*\*  
POSForward 0.41382 0.10257 4.034 8.05e-05 \*\*\*  
POSGoalkeeper 0.20233 0.14506 1.395 0.16478   
POSMidfielder 0.23597 0.07596 3.106 0.00220 \*\*   
TEAM2 -0.17169 0.08883 -1.933 0.05482 .   
TEAM3 -0.44256 0.10257 -4.315 2.62e-05 \*\*\*  
TEAM4 -0.72622 0.09671 -7.509 2.57e-12 \*\*\*  
POSForward:TEAM2 -0.38044 0.14588 -2.608 0.00987 \*\*   
POSGoalkeeper:TEAM2 -0.25101 0.19863 -1.264 0.20797   
POSMidfielder:TEAM2 -0.12697 0.11551 -1.099 0.27313   
POSForward:TEAM3 -0.29507 0.15915 -1.854 0.06536 .   
POSGoalkeeper:TEAM3 -0.24328 0.21353 -1.139 0.25605   
POSMidfielder:TEAM3 -0.12333 0.13103 -0.941 0.34782   
POSForward:TEAM4 -0.10476 0.15854 -0.661 0.50957   
POSGoalkeeper:TEAM4 -0.06337 0.24892 -0.255 0.79932   
POSMidfielder:TEAM4 0.01577 0.12455 0.127 0.89937   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.2648 on 182 degrees of freedom  
Multiple R-squared: 0.5832, Adjusted R-squared: 0.5488   
F-statistic: 16.98 on 15 and 182 DF, p-value: < 2.2e-16*

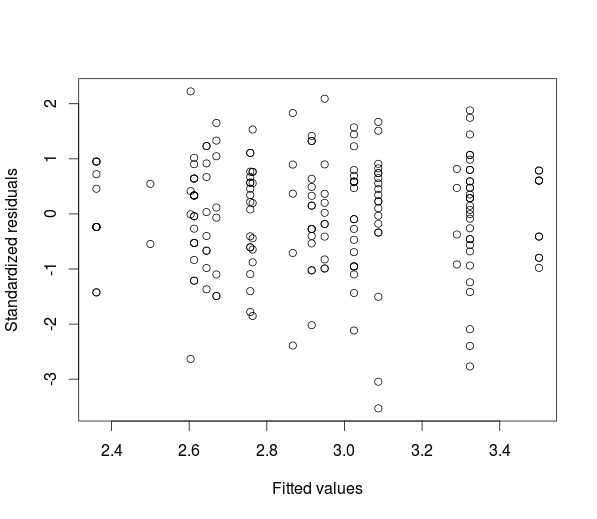
*> model.tables(fit,"mean")  
Tables of means  
Grand mean  
   
2.931668   
  
 POS   
 Defender Forward Goalkeeper Midfielder  
 2.813 2.992 2.861 2.997  
rep 58.000 38.000 15.000 87.000  
  
 TEAM   
 1 2 3 4  
 3.276 2.965 2.711 2.541  
rep 65.000 54.000 39.000 40.000  
  
 POS:TEAM   
 TEAM  
POS 1 2 3 4   
 Defender 3.087 2.916 2.645 2.361  
 rep 20.000 16.000 10.000 12.000  
 Forward 3.501 2.949 2.763 2.670  
 rep 10.000 11.000 9.000 8.000  
 Goalkeeper 3.290 2.867 2.604 2.500  
 rep 4.000 5.000 4.000 2.000  
 Midfielder 3.323 3.025 2.757 2.613  
 rep 31.000 22.000 16.000 18.000*

We can still see that the main effect is statistically significant but the Interaction effect becomes statistically insignificant after taking the logged values.The mean values shown fro each group above are in multiplicative factor of 10.For forwards and midfielders its 10^3 which is around 1000Million Euros.

Performing Diagnosis on this logged model we get the below results.



The points 14,94, and 77 are observed to be outliers.They are Hector Berlin,Steven Gerrard and Demicellies respectively.All the three players belong to Team-1 but have less market value than expected.This because these players have passed their best and are approaching their retirement age.



We can see the non-constant variance has been not seen after taking logged values.but still we are not sure about the normality of residuals from the qq-plot graph.

We can test the non-constant of variance by performing levene’s test.The results are shown below.

*> levene <- aov(abs(fit.diag$std.res) ~ POS+TEAM+POS\*TEAM)  
> Anova(levene,type=3)  
Anova Table (Type III tests)  
  
Response: abs(fit.diag$std.res)  
 Sum Sq Df F value Pr(>F)   
(Intercept) 16.073 1 42.9854 5.517e-10 \*\*\*  
POS 0.487 3 0.4338 0.7291   
TEAM 0.463 3 0.4124 0.7443   
POS:TEAM 3.442 9 1.0228 0.4234   
Residuals 68.054 182   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

The P-values shows that there is lack of non-constant variance .We need to confirm it again by performing leven’s test by removing the interaction effect.We get the below results.

*> Anova(levene,type=3)  
Anova Table (Type III tests)  
  
Response: abs(fit.diag$std.res)  
 Sum Sq Df F value Pr(>F)   
(Intercept) 25.212 1 67.3532 3.303e-14 \*\*\*  
POS 0.691 3 0.6153 0.6059   
TEAM 0.519 3 0.4618 0.7093   
Residuals 71.496 191   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

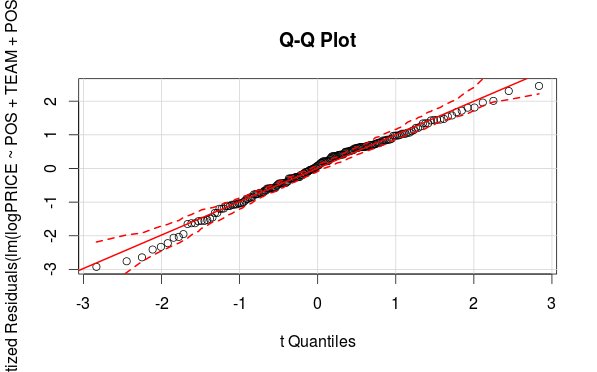
The p-value of greater than 0.05 shows that the non-constant variance is not present in this model.

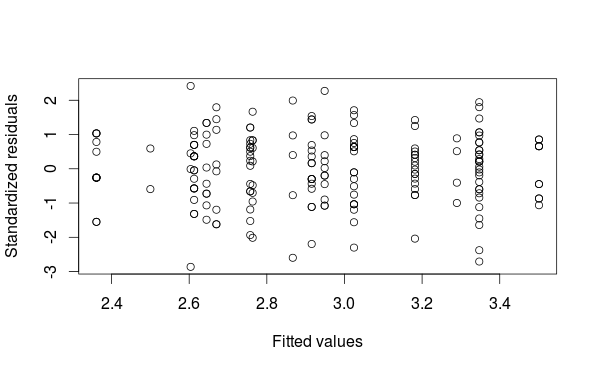
Removing those three players,and running the same model with logged prices we get the below results.

*Anova(fit,type=3)  
Anova Table (Type III tests)  
  
Response: logPRICE  
 Sum Sq Df F value Pr(>F)   
(Intercept) 182.202 1 3070.9242 < 2.2e-16 \*\*\*  
POS 0.699 3 3.9264 0.009572 \*\*   
TEAM 5.323 3 29.9079 1.01e-15 \*\*\*  
POS:TEAM 0.424 9 0.7934 0.622751   
Residuals 10.620 179   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Call:  
lm(formula = logPRICE ~ POS + TEAM + POS \* TEAM)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.64829 -0.15527 0.02082 0.16449 0.52817   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.18156 0.05741 55.416 < 2e-16 \*\*\*  
POSForward 0.31952 0.09607 3.326 0.00107 \*\*   
POSGoalkeeper 0.10803 0.13464 0.802 0.42342   
POSMidfielder 0.16570 0.07262 2.282 0.02368 \*   
TEAM2 -0.26599 0.08369 -3.178 0.00175 \*\*   
TEAM3 -0.53686 0.09607 -5.588 8.42e-08 \*\*\*  
TEAM4 -0.82052 0.09078 -9.039 2.56e-16 \*\*\*  
POSForward:TEAM2 -0.28614 0.13539 -2.113 0.03595 \*   
POSGoalkeeper:TEAM2 -0.15671 0.18358 -0.854 0.39447   
POSMidfielder:TEAM2 -0.05671 0.10807 -0.525 0.60043   
POSForward:TEAM3 -0.20077 0.14750 -1.361 0.17516   
POSGoalkeeper:TEAM3 -0.14898 0.19722 -0.755 0.45099   
POSMidfielder:TEAM3 -0.05307 0.12213 -0.435 0.66441   
POSForward:TEAM4 -0.01047 0.14694 -0.071 0.94330   
POSGoalkeeper:TEAM4 0.03093 0.22965 0.135 0.89303   
POSMidfielder:TEAM4 0.08603 0.11625 0.740 0.46024   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.2436 on 179 degrees of freedom  
Multiple R-squared: 0.6404, Adjusted R-squared: 0.6103   
F-statistic: 21.25 on 15 and 179 DF, p-value: < 2.2e-16*

The R-square value has increased to 64% and the standard error of .24 is in multiplicative factor of 10.Performing the Diagnosis on this model we get the below graphs.





We can say,that the residuals are normally distributed as they are tracking the line with 95% confidence interval.Performing Levene’s test to identify if there is any non-constant variance we get the below results.

*> levene <- aov(abs(fit.diag$std.res) ~ POS+TEAM+POS\*TEAM)  
> Anova(levene,type=3)  
Anova Table (Type III tests)  
  
Response: abs(fit.diag$std.res)  
 Sum Sq Df F value Pr(>F)   
(Intercept) 6.021 1 17.6107 4.262e-05 \*\*\*  
POS 0.837 3 0.8158 0.4867   
TEAM 0.725 3 0.7064 0.5494   
POS:TEAM 3.640 9 1.1828 0.3085   
Residuals 61.202 179   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
> levene <- aov(abs(fit.diag$std.res) ~ POS+TEAM)  
> Anova(levene,type=3)  
Anova Table (Type III tests)  
  
Response: abs(fit.diag$std.res)  
 Sum Sq Df F value Pr(>F)   
(Intercept) 15.185 1 44.0259 3.36e-10 \*\*\*  
POS 1.508 3 1.4576 0.2276   
TEAM 0.840 3 0.8119 0.4887   
Residuals 64.842 188   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

The p-values suggests the lack of non-constant variance with or without interaction effect.Hence we don't need to take weights and perform Weighted least Squares to handle non -constant variance.

Generating a model only on two main effects and ignoring the interactive effect we get the below output.

*> fit= aov(logPRICE ~ POS + TEAM)  
> Anova(fit,type=3)  
Anova Table (Type III tests)  
  
Response: logPRICE  
 Sum Sq Df F value Pr(>F)   
(Intercept) 357.75 1 6089.8615 < 2.2e-16 \*\*\*  
POS 1.21 3 6.8629 0.0002065 \*\*\*  
TEAM 17.34 3 98.4102 < 2.2e-16 \*\*\*  
Residuals 11.04 188   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

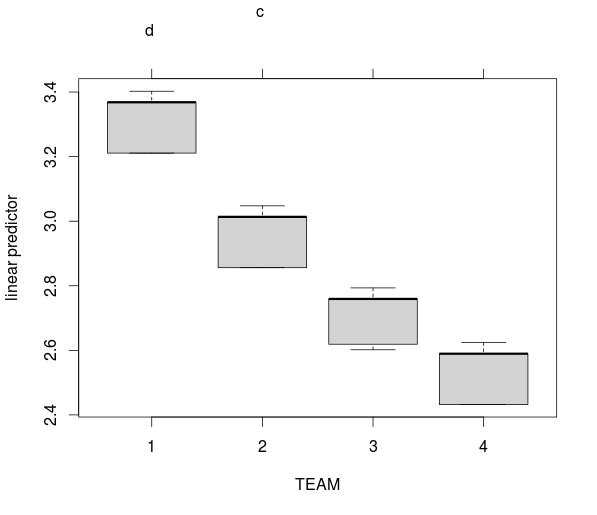
*> summary(lm(logPRICE ~ POS + TEAM))  
  
Call:  
lm(formula = logPRICE ~ POS + TEAM)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.66941 -0.15224 0.01243 0.16310 0.47797   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.21096 0.04115 78.038 < 2e-16 \*\*\*  
POSForward 0.19164 0.05102 3.756 0.00023 \*\*\*  
POSGoalkeeper 0.03418 0.07066 0.484 0.62912   
POSMidfielder 0.15742 0.04164 3.780 0.00021 \*\*\*  
TEAM2 -0.35503 0.04523 -7.850 3.07e-13 \*\*\*  
TEAM3 -0.60916 0.04972 -12.252 < 2e-16 \*\*\*  
TEAM4 -0.77875 0.04919 -15.830 < 2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.2424 on 188 degrees of freedom  
Multiple R-squared: 0.6261, Adjusted R-squared: 0.6142   
F-statistic: 52.47 on 6 and 188 DF, p-value: < 2.2e-16*

*> model.tables(fit, "mean")  
Tables of means  
Grand mean  
   
2.940467   
  
 POS   
 Defender Forward Goalkeeper Midfielder  
 2.834 2.992 2.861 3.001  
rep 56.000 38.000 15.000 86.000  
  
 TEAM   
 1 2 3 4  
 3.319 2.966 2.713 2.541  
rep 62.000 54.000 39.000 40.000*

The above means indicate that keeping the Team constant.A player playing in Forward position is estimated to cost 10^(3-2.8)=1.58 times of a Defender if they belong to same team.Similarly keeping the Position of the player constant the players in team 1 are estimated to cost 10^(3.3-3)=2 times more than players in team-2 if they play at same position.

Performing multiple comparisons for the Teams we get the below results

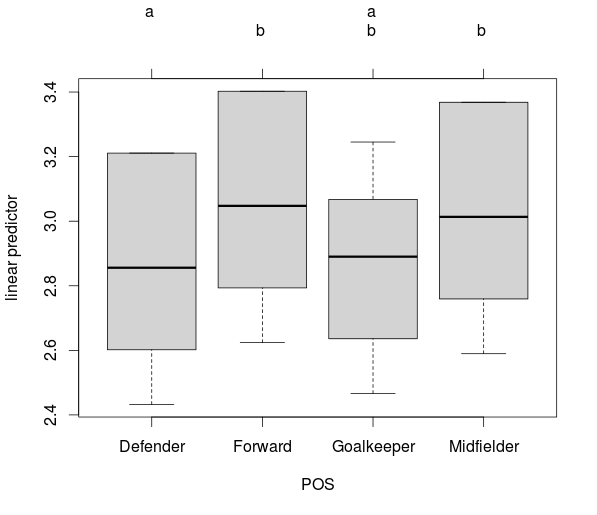
*summary(glht(fit, linfct=mcp(TEAM="Tukey")))  
  
 Simultaneous Tests for General Linear Hypotheses  
  
Multiple Comparisons of Means: Tukey Contrasts  
  
  
Fit: aov(formula = logPRICE ~ POS + TEAM)  
  
Linear Hypotheses:  
 Estimate Std. Error t value Pr(>|t|)   
2 - 1 == 0 -0.35503 0.04523 -7.850 <0.001 \*\*\*  
3 - 1 == 0 -0.60916 0.04972 -12.252 <0.001 \*\*\*  
4 - 1 == 0 -0.77875 0.04919 -15.830 <0.001 \*\*\*  
3 - 2 == 0 -0.25414 0.05096 -4.987 <0.001 \*\*\*  
4 - 2 == 0 -0.42372 0.05065 -8.367 <0.001 \*\*\*  
4 - 3 == 0 -0.16959 0.05468 -3.101 0.0118 \*   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
(Adjusted p values reported -- single-step method)*



We can clearly see that there are 4 distinct groups found in this method by using Tukey’s test for Teams.

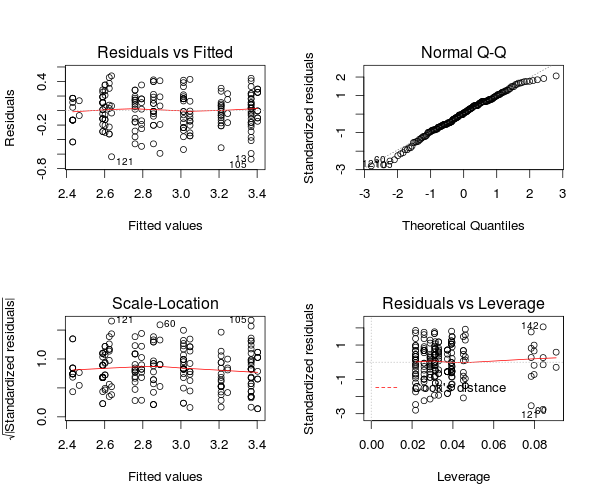
Performing multiple comparisons for the positions we get the below results.

*summary(glht(fit, linfct=mcp(POS="Tukey")))  
  
 Simultaneous Tests for General Linear Hypotheses  
  
Multiple Comparisons of Means: Tukey Contrasts  
  
  
Fit: aov(formula = logPRICE ~ POS + TEAM)  
  
Linear Hypotheses:  
 Estimate Std. Error t value Pr(>|t|)   
Forward - Defender == 0 0.19164 0.05102 3.756 0.00121 \*\*  
Goalkeeper - Defender == 0 0.03418 0.07066 0.484 0.96112   
Midfielder - Defender == 0 0.15742 0.04164 3.780 0.00104 \*\*  
Goalkeeper - Forward == 0 -0.15746 0.07399 -2.128 0.14158   
Midfielder - Forward == 0 -0.03422 0.04734 -0.723 0.88378   
Midfielder - Goalkeeper == 0 0.12324 0.06805 1.811 0.26217   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
(Adjusted p values reported -- single-step method)*



We can clearly see that the market prices of both forwards and midfielders is same and is statistically different from The Defenders based on the Tukey’s test results.

Below are the Diagnostics for this model



We can still see a right tail for normality test.But the non-constant variance has been handled by taking logged price.