Course Assignmemt 1 2023-24

Advanced Data Analysis with R

Nikolaos Papadopoulos, Postgraduate student in Statistics AUEB

2024-01-05

Contents

1. Introduction

The *'Super League Greece'* championship is the name of the football league that is being governed by the Hellenic Football Federation (HFF) founded on the 14th of November 1926. HFF is an active member in the FIFA and UEFA affiliation and almost annually sends the best Greek teams to the well known European Championships to compete with the big names of Europe in football.

An interesting year in the *'Super League Greece'* was the one that took place during the season of 2013-2014 and was the 78th championship in football league in which 18 teams were clashing for the first place. The champion team of that year turned to be *'Olympiakos'* with a tremendous 17 difference of points from the second PAOK.

The first place was immediately qualified for the Champions League group stage, while the 2nd to 5th place had to clash again in the play-offs to claim the second space for the Champions League. The rest three teams after the play-offs would be qualified for the Europa League play-off rounds. At the same time, the last two teams (17th and 18th) would be relegated to the today's *'Super League 2 Greece'* which was then called *'Football League'*. The team in the 16th place would be qualified to the relegation play-off, clashing with the second team of the *'Football League'* which would try to climb to the 1st Division.

The first part of this assignment will be focused on the part that has to do completely with the data that came out of the football matches. These data contain mainly the names of the teams, the the goals scored until half time and full time, the outcomes of the half and full times that correspond to the matches. In a more visual way, the following table is an index of the 'results' data. In this one, the first 3 matches appear in the list.

Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR
G1	17/08/13	Atromitos	Ergotelis	2	2	D	1	2	A
G1	17/08/13	PAOK	Xanthi	3	0	H	1	0	H
G1	17/08/13	Platanias	Veria	2	2	D	2	0	H

The second part of the assignment will be exclusively on the odds that some of the most popular bookmakers make for the public to bet on the matches. Those odds provide useful information before the beginning of the matches and can be also analysed. Those, have the following form;

B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA	LBH	LBD	LBA
1.57	3.5	7	1.57	3.5	6.75	1.55	3.7	5.4	1.62	3.25	6.0
1.44	4.0	8	1.45	3.9	8.00	1.57	3.6	5.4	1.45	3.60	7.5
2.00	3.2	4	2.00	3.2	3.90	2.10	3.0	3.4	1.95	3.00	3.9

Standing Board of Super Leauge 2013-2014 *(If not found try to run the html file)*

2. Descriptive analysis and exploratory data analysis

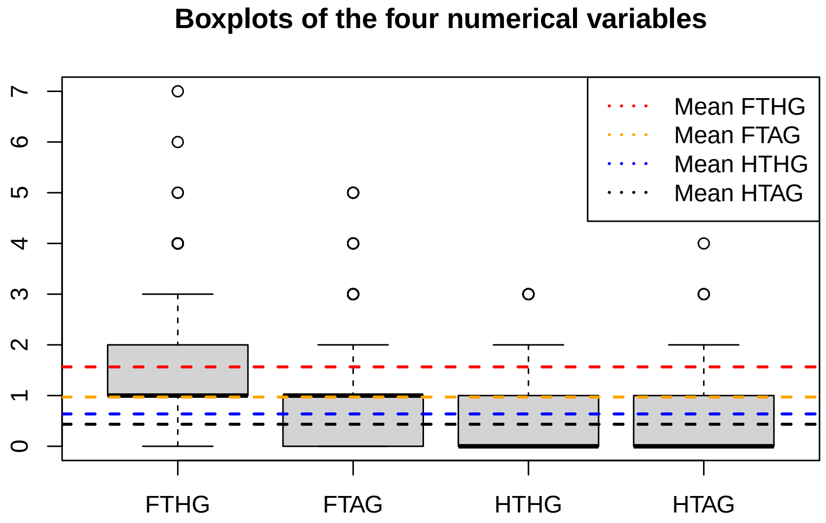
2.1. Results' data descriptive analysis

In this part of the analysis, some elementary characteristics of the four variables 'Final Time Home Goals' (from now on FTHG), 'Final Time Away Goals' (from now on FTAG), 'Half Time Home Goals' (from now on HTHG) and 'Half Time Away Goals' (from now on HTAG) will be printed in order to retrieve information relative to their distributions.

First of all, the descriptive statistics is the most efficient way to start with. In the following table, we can see various meters that provide information about these variables.

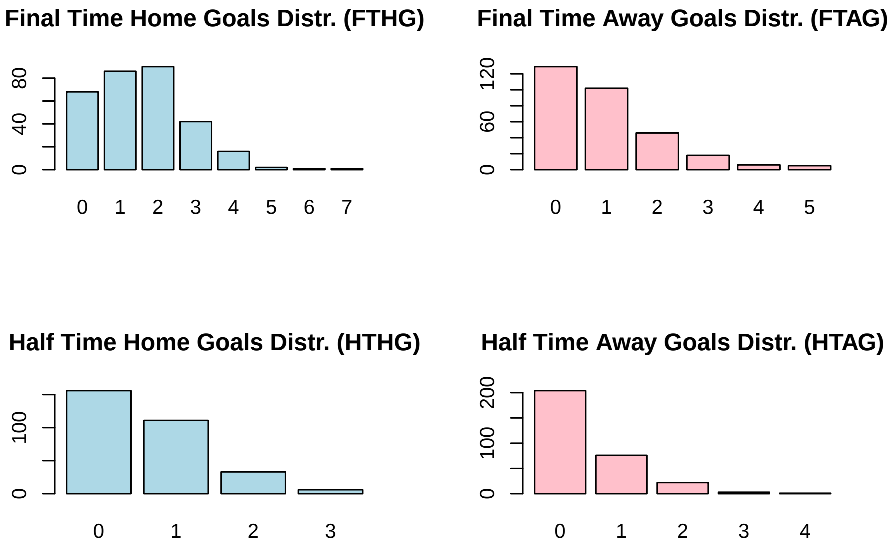
	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
FTHG	1.565	1.235	1	1.472	1.483	0	7	7	0.719	0.790	0.071
FTAG	0.971	1.117	1	0.780	1.483	0	5	5	1.351	1.813	0.064
HTHG	0.637	0.753	0	0.524	0.000	0	3	3	0.978	0.346	0.043
HTAG	0.435	0.699	0	0.293	0.000	0	4	4	1.694	2.991	0.040

As one may see, almost all indexes are different. The ones that are most important to compare are the first two rows and the last two. The reason is because in this way we can compare how better the Home team does against the Away team. As we can see in both cases (Full and Half Time) the Home teams are on average superior to the Away teams.



In this graph, the four variables FTHM, FTAG, HTHG, HTAG are all fitted in a boxplot to make the comparison easier.

Lastly, the most efficient way to visualize the distributions of the aforementioned variables is the following;


Final Time Home Goals Distr. (FTHG)

Final Time Away Goals Distr. (FTAG)

Half Time Home Goals Distr. (HTHG)

Half Time Away Goals Distr. (HTAG)

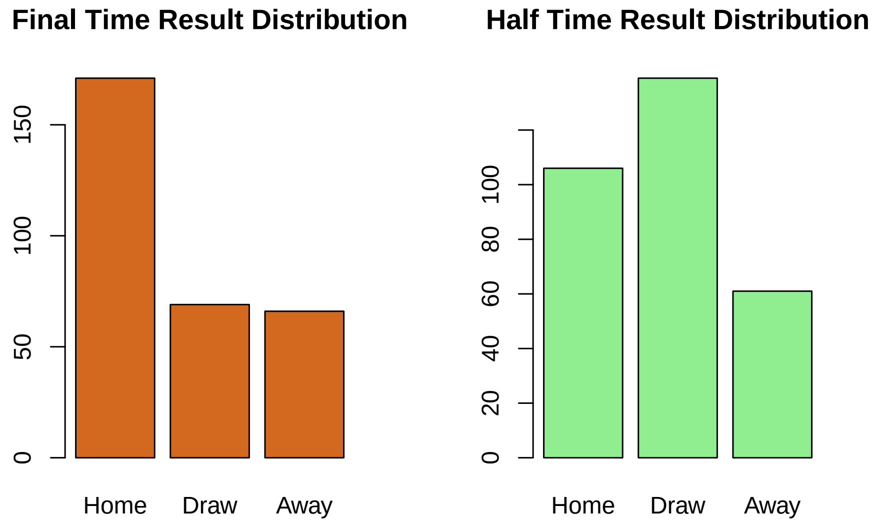
The above plot has on the left side, with the light blue colour, the distributions of the goals that the 'Home' teams scored. The difference between these two graphs is that the above corresponds to the goals achieved during the whole games, while the below corresponds to the goals achieved during the first half of the games.

Respectively for the two graphs appearing on the right side, with the only difference that these goals refer to the goals achieved by the 'Away' teams.

Some interesting facts that one could retrieve from these barplots are the following;

1. The fact that the distribution of the FTHG with the mod being 2, is different from the one of the FTAG with the mod being 0! This somewhat proves the old saying that being in your own field, plus the support of the fans help the 'Home' teams fight for a better outcome.

2. On the HTAG distribution number 4 is an actual bar, and corresponds to the match 'Ergotelis - Olympiakos' where Olympiakos crashed Ergotelis on half time with 0-4… The funny part is that the match ended 1-4. The odds from this match were of course in favor of Olympiakos with a mod equal to 1.33.



2.2. Results' data descriptive analysis

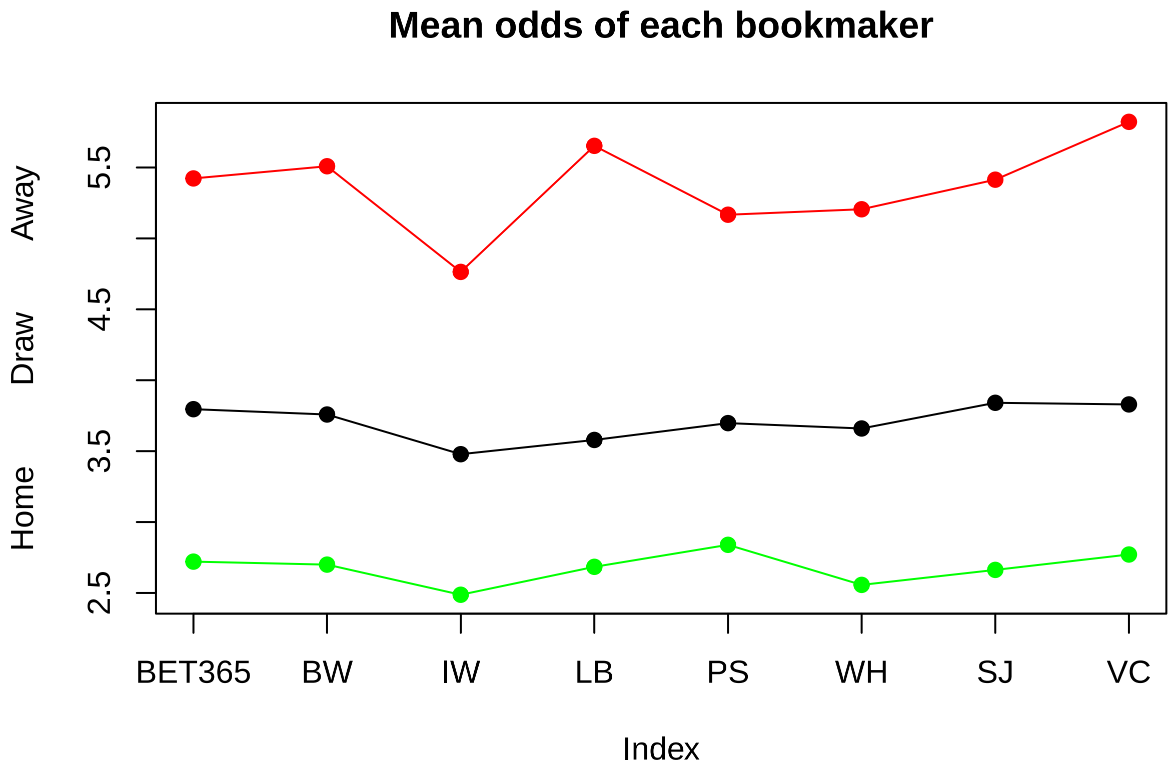
In this part of the descriptive analysis we will see the main characteristics of the odds which are numerical variables and can provide a lot of information for their continuous distributions.

Primarily a table can be printed to indicate the summary statistics.

	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
B365H	2.720887	2.535424	2.0	2.173064	0.637518	1.05	19.0	17.95	4.110274	19.771434	0.1481210
B365D	3.795973	1.309177	3.3	3.481660	0.296520	2.88	12.0	9.12	3.107018	11.280483	0.0764829
B365A	5.422867	4.813332	4.0	4.443149	2.223900	1.14	34.0	32.86	2.835081	9.981077	0.2811979
BWH	2.699802	2.529575	2.0	2.167325	0.667170	1.05	21.0	19.95	4.295756	21.738965	0.1453203
BWD	3.758218	1.278712	3.3	3.462510	0.444780	1.25	11.5	10.25	2.842077	9.550945	0.0734601
BWA	5.508878	4.834531	4.1	4.521687	2.298030	1.11	31.0	29.89	2.697978	8.821184	0.2777365

In short, if one analytically observe the whole table, he will notice that all the values that correspond to equal characteristics (for example if he compares all the Home odds) he will not be able to see distinguishable difference. And due to the large amount of observations, it is almost sure that there is no difference.

Below we can see a plot that will prove in a way what has been said. This is a simple plot that has with a green line the mean of the odds per bookmaker for every spot (H/D/A).

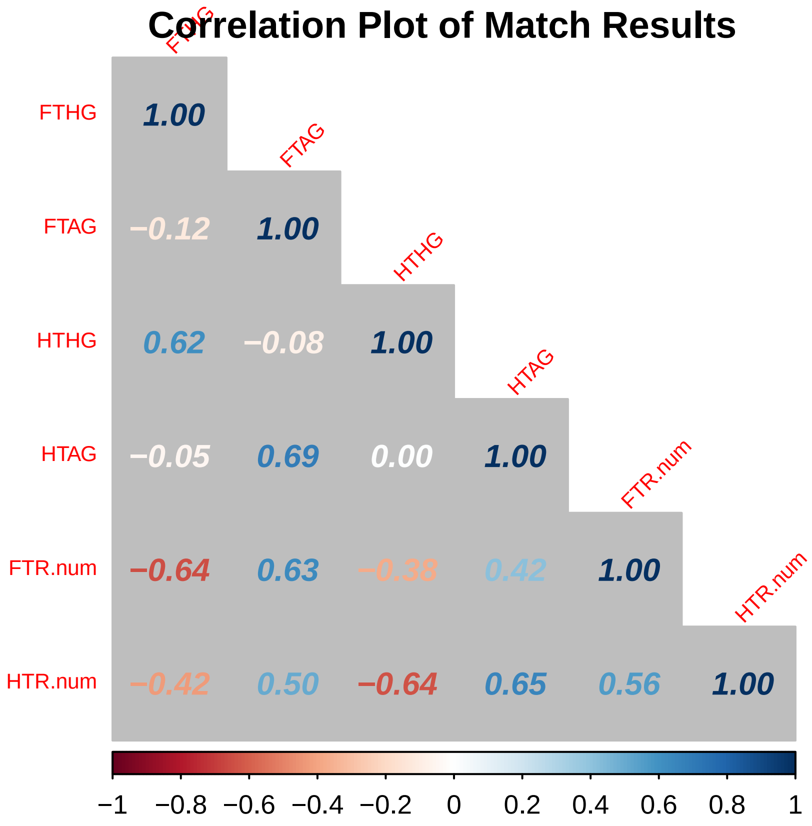


The only observable difference is in the Away mean odds. This can be explained due to the fact that there have been matches in which the Home team is much better than the Away team and small changes in the Home odds can create big changes in the Away odds. (For example a possible odd group is 1.17-7.00-12.00 and another one is 1.22-6.16-9.25 so a change of 0.05 in home creates a 2.75 difference in Away.) The same happens sometimes if the opposite happens - the Away team is much better than the home team, but from historical data we know that sometimes the *'david and goliath'* phenomenon happens often. So the bookmakers know this and are more conservative with their suggesting odds.

3. Pairwise comparisons

3.1 Results' data correlation

The pairwise comparisons have always been an excellent introduction step to dive into the correlation between the variables of the data. For the data that concern the matches' statistics, it is natural to be suspecting the existence of correlation between the number of goals (FTHG, FTAG, HTHG, HTAG) and of course the outcome (FTR, HTR).

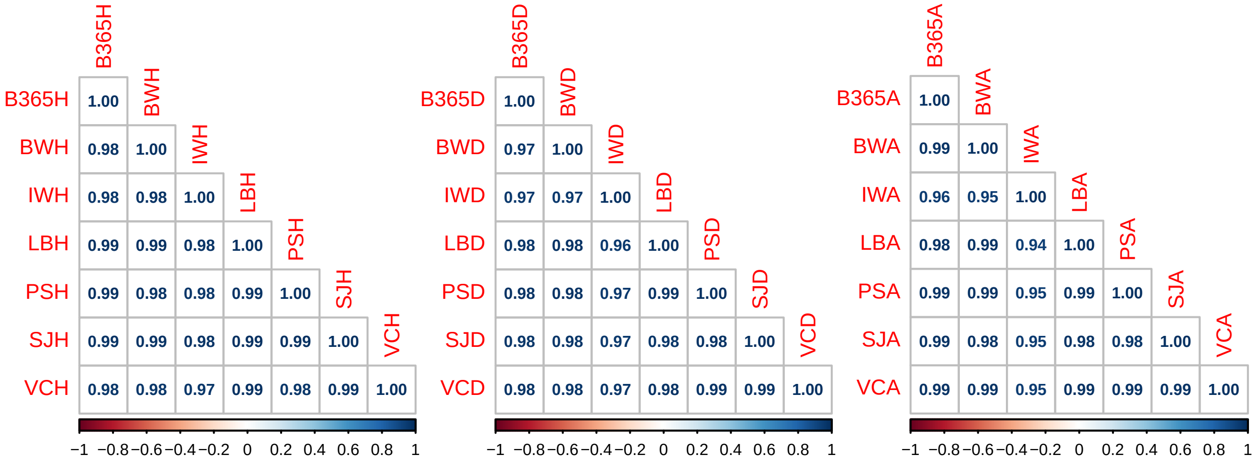


The first thing that can be said about this correlation plot is that the two categorical variables that became numeric so that they can fit into the analysis show huge correlation with all the variables. Interestingly the correlation between the two categorical variables have a correlation index of 0.56 which indicates that the Half Time Result is not strongly correlated to the Final Time Result.

Another interesting note is that some negative correlations appear too between FTR-FTHG and HTR-HTHG. This can be explained due to the way we have numbered the variables, because the higher number of goals by the Home team, the more likely it is the result to be in favor of Home, which is indicated with 1, which is the lowest value (H:1, D:2, A:3). The same applies with the half time and for the away goals which show high correlation too.

3.2 Bookmakers' odds' correlation

For the data that concern the matches' odds, we expect the correlation to be extremely strong - one could say even equal to 1 - because the odds represent probabilities of event in the match, and these probabilities mostly derive from past data and all these same data are analysed by the bookies individually.



	B365H	BWH	IWH	LBH	PSH	SJH	VCH
B365H	1.00						
BWH	0.98	1.00					
IWH	0.98	0.98	1.00				
LBH	0.99	0.99	0.98	1.00			
PSH	0.99	0.98	0.98	0.99	1.00		
SJH	0.99	0.99	0.98	0.99	0.99	1.00	
VCH	0.98	0.98	0.97	0.99	0.98	0.99	1.00

	B365D	BWD	IWD	LBD	PSD	SJD	VCD
B365D	1.00						
BWD	0.97	1.00					
IWD	0.97	0.97	1.00				
LBD	0.98	0.98	0.96	1.00			
PSD	0.98	0.98	0.97	0.99	1.00		
SJD	0.98	0.98	0.97	0.98	0.98	1.00	
VCD	0.98	0.98	0.97	0.98	0.99	0.99	1.00

	B365A	BWA	IWA	LBA	PSA	SJA	VCA
B365A	1.00						
BWA	0.99	1.00					
IWA	0.96	0.95	1.00				
LBA	0.98	0.99	0.94	1.00			
PSA	0.99	0.99	0.95	0.99	1.00		
SJA	0.99	0.98	0.95	0.98	0.98	1.00	
VCA	0.99	0.99	0.95	0.99	0.99	0.99	1.00

As we can see, indeed almost all values in this 7x7 box are almost one.

Another very significant reason why the bookmakers should pay attention and have similar odds with other platforms is the fact that the players can take advantage of it and have a guaranteed win by strategically placing large amounts of money in a specific match and betting all 3 possible outcomes in different platforms. This is called **arbitrage** and more information can be found *here*

```
## [1] "LBH VCD VCA"
```

The above names are the names that indicate the highest number if we sum by rows all the corrplots and keep the greatest number. This was done because we want to keep the bookmaker who has the highest accordance with the others and since VC appeared twice, we are going to continue the analysis with this bookmaker.

4. Predictive or Descriptive models

The next station of the analysis is of course the predictive and descriptive models that can be produced by almost any kind of data and especially the ones that are at some considerable degree correlated.

As it was mentioned earlier, for the sake of this assignment, the ‘VC’ bookmaker will stay in the data as representative of the rest bookmakers. The question that the data are going to answer in the following lines is whether or not it is possible that one can find a model which will accurately predict the odds of the bookmaker VC.

4.1. VCA odds

In order to do this, the initial data frame will be separated and only the meaningful variables will be kept, such as the FTHG, FTAG, HTHG, FTR, HTR. A linear model is in most cases the easiest way to continue with the analysis since we are dealing with numeric, continuous variables and not categories. To start with, the prediction of VCA will go first.

However, in this regression it is apparent that there are many errors. The most important one and the one that should be checked in advance is the fact that some variables contain the same information twice (for example the final time home goals contain the half time home goals of the same match) and this is a common case of multicollinearity.

The variables that are considered to be causing the multicollinearity to the model are the following: (Use of a strict criterion: vif < 4)

```
## [1] "HTHG" "HTAG" "FTR"  "HTR"
```

so running again the corrected model will give the following result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.51355	0.56886	7.93442	0.00000
FTHG	1.45339	0.24199	6.00609	0.00000
FTAG	-0.99025	0.26800	-3.69489	0.00026

The only two variables that seem to answer well the above model are the ones that are connected to the final time goals for each team. This makes a lot of sense, since by the time we know the final result of the match we can assume odds relevant to the final time result.

However, the R-squared adjusted value in our model is 0.1504834 means that approximately 17% of the variability in the dependent variable is explained by the independent variables in the model, after adjusting for the number of predictors. In other words, the independent variables in the model collectively account for 17% of the variation observed in the dependent variable, and the remaining 83% of the variation is not explained at all.

This can be rationally explained by the fact that even if we know the outcome of the game, we are not guaranteed to predict on point the Away odds of the match. An example to make this clear is the following… Imagine the national football team of France playing with the national football team of Gibraltar. The last match that these two teams had ended 14-0 in favour of France. Apparently, the odds for France would be very close to 1. However, imagine if the match had Gibraltar for winner. Then our model would never have predicted this, and it is very likely that it would give odds for Gibraltar even less than 2!

Also another reason that this model is quite faulty is that bookmakers don't use these variables to create the odds since they are unknown until that time. Instead, they use numerous football data from previous matches of both teams (wins/losses, goals scored, player strength, missing players, time of goals, corners, fouls, clean sheets, etc.), analyse them and come up with odds that also contain a little fee called **rake**. For more on rake, click *here*

So… This is a very bad model and in most cases it shouldn't be an option for predictions, but for the shake of this assignment we are going to continue with it. The next thing we have to do is to check the main regression assumptions if they are satisfied.

In order to check for the normality of the residuals, we are using three statistical tests, the *Shapiro test*, the *Kolmogorov-Smirnov test* and the *Lilliefors test*, and in all three the p-value is less than 0.05 so we reject the null hypothesis which assumes normality of the residuals.

Secondly in order to check for autocorrelation of the residuals, we are using the Durbin Watson test. The p-value is greater than 0.05, so there is no strong indication to reject the null hypothesis, so we can assume no autocorrelation in the residuals of the model.

Lastly, to check for homoskedasticity, we run the Breusch Pagan test. The p-value is less than 0.05, so we reject the null hypothesis which assumes homoskedasticity and there is heteroskedasticity.

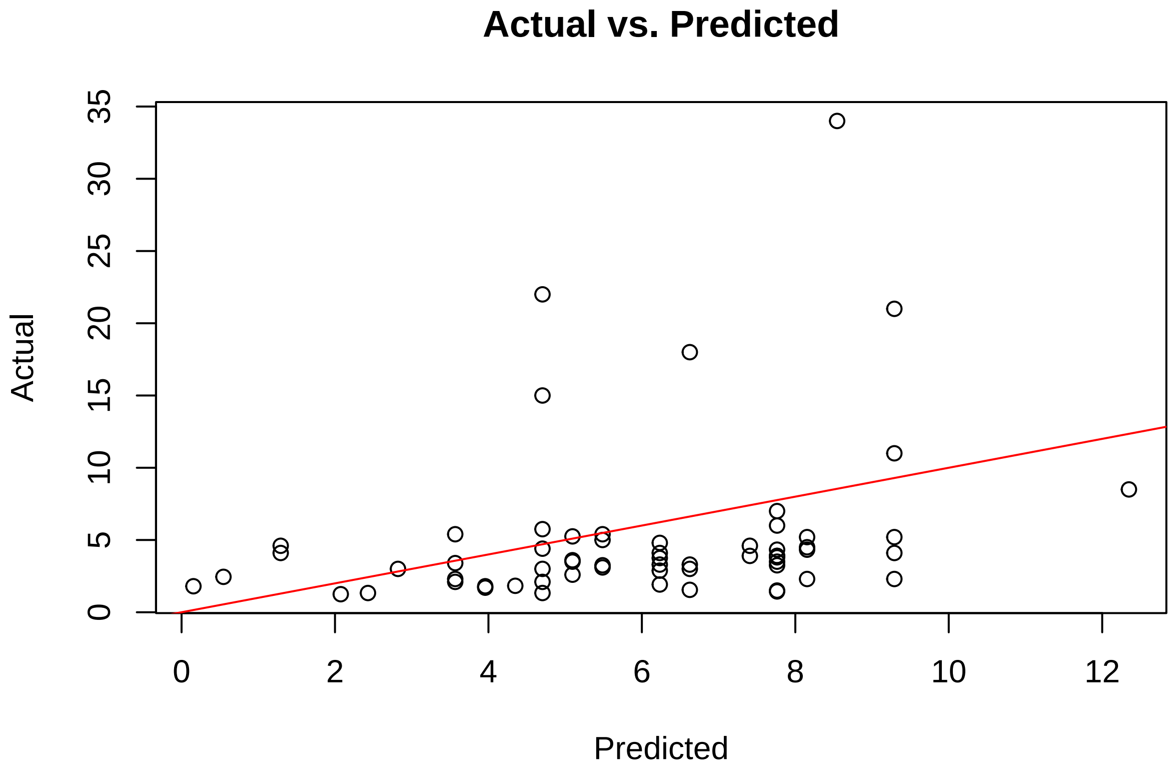
Now if we want to make predictions for the above model, we have to split the data into train and test data.

```
## [1] "Mean Squared Error: 30.572"
```

The process with which this training was done, was by taking the 80% of the data randomly to train or more commonly run the model, to find the estimates, etc and by taking the rest 20% of the data to test if these data fitting well in our regression.

The mean squared error of this testing is 30.572 which is a measure of the average squared difference between predicted values and actual values in the regression.

With the following plot we can see that there are values who are very close to the line but also some others that could be considered as outliers. The reason is similar as previously, as in football data there can be such odds values in very uneven matches.



4.2. VCD odds

Lastly, the same analysis will follow for the rest two odds, VCD and VCH. Starting with VCD, the model is the following:

where the variables that cause the multicollinearity are:

```
## [1] "HTHG" "HTAG" "FTR"  "HTR"
```

and after the multicollinearity fix, it becomes this:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.41379	0.14751	23.14270	0.00000
FTHG	0.24933	0.06275	3.97337	0.00009
FTAG	0.02582	0.06950	0.37157	0.71047

As we can see, only the FTHG is considered to be a significant variable for the model. And if run it in a simple linear regression we will see that again the R-square adjusted is extremely low.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.4432818	0.1241532	27.734134	0.00e+00
FTHG	0.2466015	0.0622303	3.962723	9.25e-05

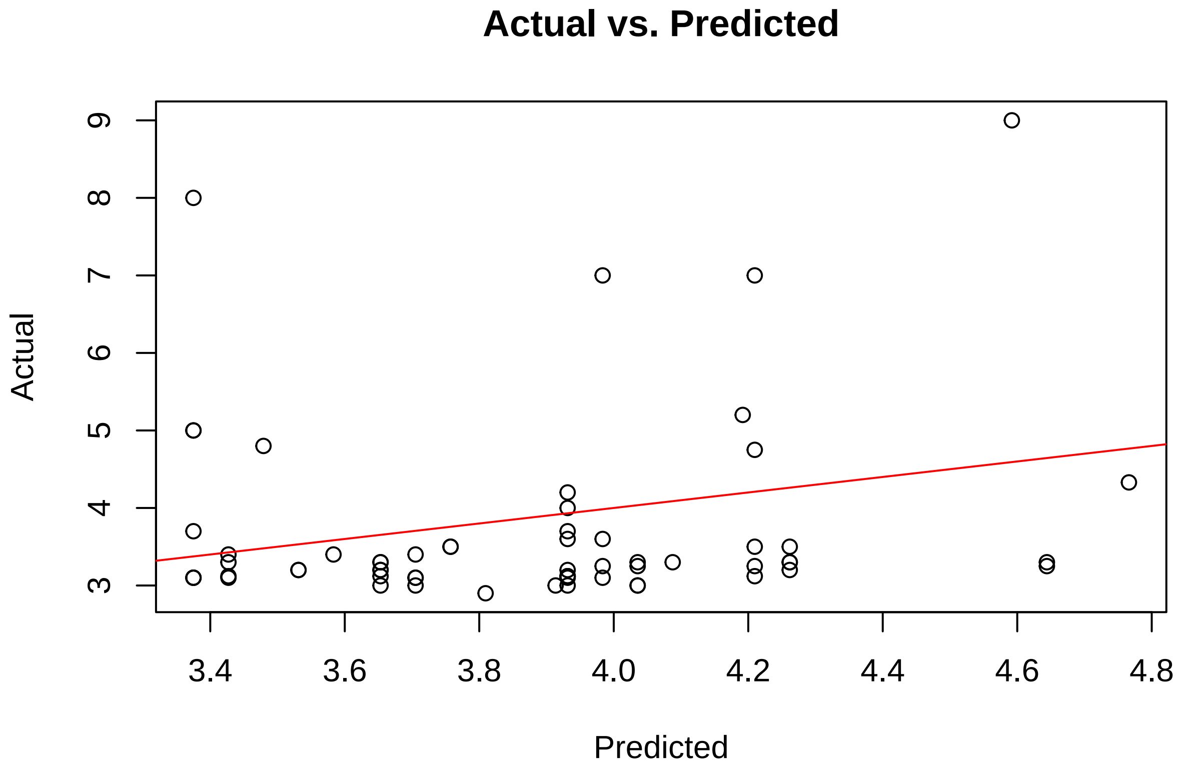
```
## [1] "R-square adjusted: 0.046"
```

Now after having done the necessary hypothesis testing for the assumption of the model, now all three assumptions are rejected for significance level a = 0.08.

Next we can do the training of the model right as before.

```
## [1] "Mean Squared Error: 1.47"
```

And then we can get the plot:



4.3. VCH odds

The last is the VCH, the model is the following:

Once more, the variables that cause the multicollinearity are the same:

```
## [1] "HTHG" "HTAG" "FTR"  "HTR"
```

and after the multicollinearity fix, it becomes this:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.47310	0.26735	9.25058	0e+00
FTHG	-0.51160	0.11373	-4.49851	1e-05
FTAG	1.12504	0.12595	8.93215	0e+00

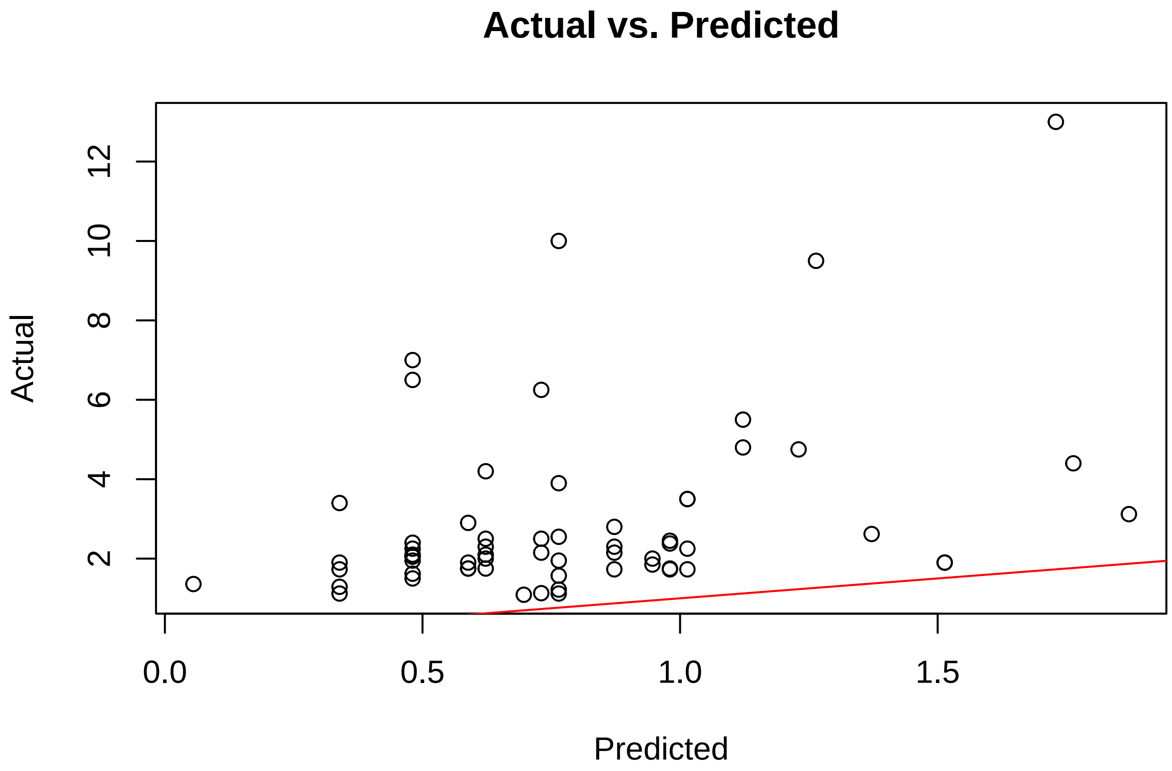
Now both variables are considered significant.

In this assumptions' analysis the shapiro test shows that the residuals are normally distributed after a log transformation on the residuals. The DW test shows no autocorrelation and the BP test shows heteroskedasticity. So, we are going to work with the transformed model.

Now the training gives the mean square error to be:

```
## [1] "Mean Squared Error: 9.12"
```

And for the last time the plot:


Actual vs. Predicted
Actual
2
4
6
8
10
12
0.0
0.5
1.0
1.5
Predicted

5. Further analysis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-33.0495124	1810417.1	-0.0000183	0.9999854
FTHG	53.8879750	111108.4	0.0004850	0.9996130
FTAG	-53.7524883	123545.8	-0.0004351	0.9996529
HTHG	0.0797165	102796.9	0.0000008	0.9999994
HTAG	-0.3012154	115940.7	-0.0000026	0.9999979
B365H	-3.8657984	909884.1	-0.0000042	0.9999966
B365D	-0.8089678	747954.9	-0.0000011	0.9999991
B365A	0.0134586	220445.2	0.0000001	1.0000000
BWH	-2.8705459	811770.9	-0.0000035	0.9999972
BWD	-1.9159883	606266.6	-0.0000032	0.9999975
BWA	-0.0280703	293991.3	-0.0000001	0.9999999
IWH	-2.7545227	600450.9	-0.0000046	0.9999963
IWD	-1.0626513	646016.4	-0.0000016	0.9999987
IWA	-0.2484837	200909.3	-0.0000012	0.9999990
LBH	-1.7651227	1053601.3	-0.0000017	0.9999987
LBD	0.2268159	1048546.4	0.0000002	0.9999998
LBA	-1.0212571	322154.0	-0.0000032	0.9999975
PSH	3.8632572	1218784.0	0.0000032	0.9999975
PSD	0.3358289	899455.4	0.0000004	0.9999997
PSA	-0.8357773	273499.2	-0.0000031	0.9999976
WHH	0.4325920	542938.7	0.0000008	0.9999994
WHD	1.2308164	517459.9	0.0000024	0.9999981
WHA	-0.4976765	169962.6	-0.0000029	0.9999977
SJH	0.8644956	946165.2	0.0000009	0.9999993
SJD	5.5881867	1045355.3	0.0000053	0.9999957
SJA	1.9854890	311443.1	0.0000064	0.9999949
VCH	4.6676809	1419004.9	0.0000033	0.9999974
VCD	-0.5338095	985429.8	-0.0000005	0.9999996
VCA	0.2589542	248800.3	0.0000010	0.9999992

```
## [1] "Accuracy: 0.8226"
```

6. Conclusions and Discussion

The analysis presented in this report focused on exploring and understanding the data related to the 2013-2014 Super League Greece football championship. The descriptive analysis of match results highlighted the dominance of home teams, both in terms of average goals scored and outcomes. The distribution of goals scored by home teams showed a distinct pattern compared to away teams, reinforcing the commonly observed home advantage in football. Additionally, the distribution of halftime and full-time results provided a snapshot of the competitiveness of the league.

Exploring the odds offered by bookmakers revealed interesting patterns. Correlation analyses demonstrated strong positive correlations among odds provided by different bookmakers, indicating a consensus in the market.

Further, predictive models were attempted to estimate the bookmakers' odds based on match statistics. However, the models exhibited limitations, with low R-squared values and various assumptions not being met. This highlighted the complexity of predicting odds solely based on match outcomes and goal statistics.

As a continuation, a logistic regression model was introduced to predict the probability of a home team winning a match. This model utilized both match statistics and bookmakers' odds as features. While the logistic regression model provided a framework for predicting match outcomes, its accuracy demonstrated the challenges of such predictions in the context of football.

In conclusion, the analysis offered valuable insights into the Super League Greece 2013-2014 season. The dominance of home teams, patterns in goal distributions, and the consensus among bookmakers were evident. The predictive modeling attempts shed light on the difficulty of accurately predicting bookmakers' odds based solely on match statistics. Further research and consideration of additional variables, such as team performance indicators and player statistics, may enhance the predictive capabilities of such models. Additionally, understanding the limitations and uncertainties in football predictions is crucial for making informed decisions in sports betting.