Quantitative economics Project

Introduction

For my assignment, I will investigate the performance of players which I believe is dependent on Creativity, influence, and threat as the main independent variables. My data is in the form of a time series that has over 40 observations meaning I will be able to adopt more degrees of freedom and may also adapt my table to best suit the result. Below, I have listed 14 observations as proof. All the data found I have provided can be found on the Github website where they collect and compile a variety of data based on Fantasy football statistics of the English Premier League.

Theory

$TP = \beta 0 + \beta 1(G+A) + \beta 2(I) + \beta(C) + \varepsilon$

For my theoretical model I have included my dependant variable and some of the independent I will use in this project because if G+A increase total points will increase but if influence or create increase then is increase the likelihood of total point increasing.

Data

Whilst my motive is to see whether a players performance is affected by different variable and it is mainly focusing on the individuals players statistics

not as a whole team performance and this is where there may be some discrepancies and drawback as football is a team game and you have to rely on team mates just as much as the players own skill and performance whilst I am measuring Total points as my dependant variable and my performance indicator so the higher the point the high the performance but I will also look at several other variable such as Goal+Assists, Influence, creativity, threat and the amount of minutes played. I have put a players goals and assist statistic together as this represents their contribution to a team as a whole and not to their own individual benefit whilst I have players from all positions besides the goal keeper and it would be ideal to put these to factors together as defenders now create and score and are vital to teams performance just as much as the midfielders and forwards, whilst influence is the first measurement - this evaluates the degree to which that player has made an impact on a match, or matches over the season. It considers events and actions that could directly or indirectly effect the match outcome. At the very top level these are decisive actions like goals and assists. However, the Influence score also processes significant defensive actions to analyse the effectiveness of defenders and goalkeepers. Creativity assesses player performance in terms of producing goal scoring opportunities for others. It can be used as a guide to identify the players most likely to supply assists. While this analyses frequency of passing and crossing, it also considers pitch location and the incisiveness of the final ball. Threat is the third measure, producing a value that examines a player's threat on goal; it therefore gauges those individuals most likely to score goals. While attempts are the key action, the Index looks at pitch location, giving greater weight to actions that are regarded as the best openings to register a goal. All three of these scores are then combined to create an overall ICT Index

score. That then offers a single figure that presents a view on that player as an FPL asset.

Results

reg1=lm(log(tp)~ga, data = FPLDataoriginal)

summary(reg1)

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.255199	0.076942	55.304	2.00E-16	***
G+A	0.037131	0.006554	5.665	1.51E-06	***
Signif. Codes	0 '***'	0.001'**'	0.01'*'	0.05'.'	0.1' '
Multiple R-squared: 0.4514, Adjusted R-squared: 0.4374 p-value: 1.511e-06					

From the regression we see that as G+A increases by 1 Total point(TP) increases by 0.03 which is expected as there are many players and it has a positive relationship, also as our t value is greater than 2 and our p value is less than 1% or very close zero this is statistically significant, however we need to add more variables to make this test more accurate and complete.

However we know that Influence, Creativity and Threat are all correlated because they all are similar and can be put together as a form of an ICT index as stated form Github, therefore, I will have to choose one or two from the three options and because minutes is a key variable and does not correlate I will also keep that in.

reg2=lm(log(tp)~Minutes+ga+Creativity+Threat, data = FPLDataoriginal)

summary(reg2)

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.71E+00	1.49E-01	24.986	2.00E-16	***
Minutes	2.42E-04	6.02E-05	4.016	0.000287	***
ga	3.73E-02	1.26E-02	2.972	0.005243	**
Creativity	2.60E-04	1.10E-04	2.36	0.023797	*
Threat	-1.07E-04	1.96E-04	-0.546	0.588538	

From the regression we can see it can still be improved as the results so that ga is less significant compared to reg 1 but minutes has the most significance because the more minutes a player plays they are guaranteed more point because as minutes go up by 1 unit TP increase slightly by 0.00242. Also we must also conclude to improve this regression we must remove threat as it has a negative impact on TP and is not statistically significant with a t value below 2 in absolute terms and a very high p value at 58%.

reg3=lm(log(tp)~Minutes+ga+Creativity, data = FPLDataoriginal)

summary(reg3)

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.70E+00	1.44E-01	25.748	2.00E-16	***
Minutes	2.45E-04	5.93E-05	4.124	0.000202	***
ga	3.13E-02	5.84E-03	5.35	4.76E-06	***
Creativity	2.78E-04	1.05E-04	2.657	0.011578	*

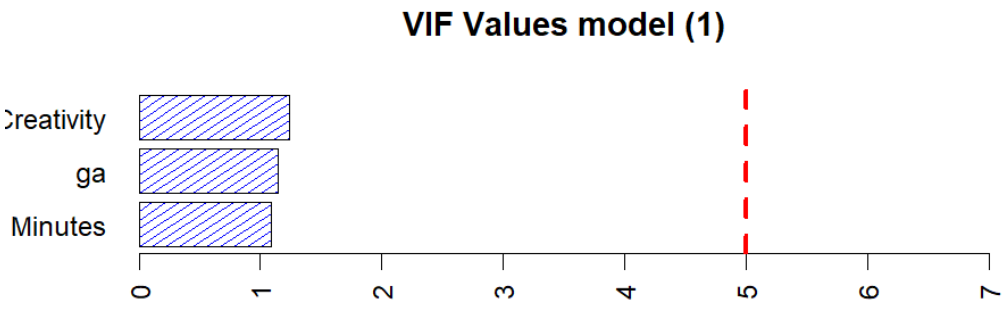
Signif. Codes	0 ‘***’	0.001‘**’	0.01‘*’	0.05‘.’	0.1‘ ’
Multiple R-squared: 0.64 Adjusted R-squared: 0.6108 p-value: 2.49E-08					

After removing threat from the regression the result are much more complete and all have a level of significance because GA has gained in significance as p value is below 1% and very close to zero and we can see that these independent variables are useful as the adjusted r^2 increased.

I am now happy with my regression 3 and can now move onto diagnositcs.

Multicollinearity

```
#obtain vif

reg3vif=vif(reg3)

##plot vif

barplot(reg3vif, main = "VIF Values model (1)", horiz =TRUE, col = "blue",

      xlim=c(0,7), density = 20, las=2)

abline(v = 5, lwd = 3, lty = 2, col="red")
```

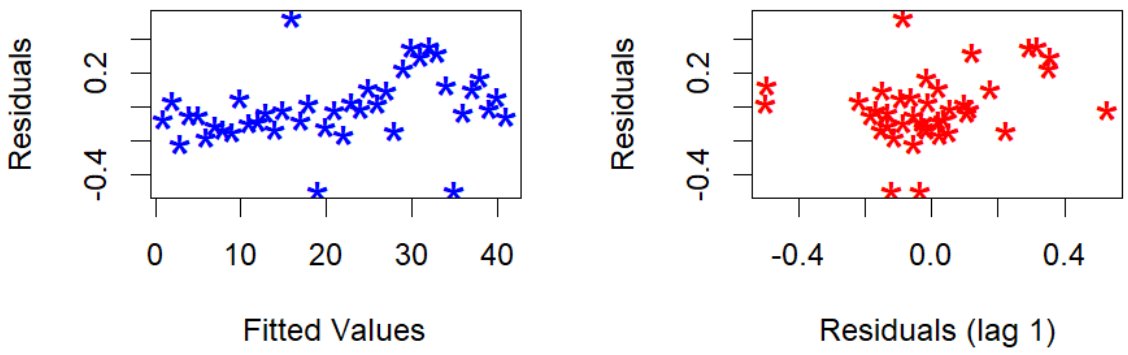


Test statistic is smaller than 5 - no obvious multicollinearity

problem in our data so we can keep all variables.

Autocorrelation

```
errors3=residuals(reg3)
# plot errors
par(mfrow=c(1,2))
plot(errors3, xlab="Fitted Values", ylab="Residuals", pch="*", cex=2,
col="blue")
plot(errors3[-41]~errors4[-1], xlab="Residuals (lag 1)", ylab="Residuals",
    pch="*", cex=2, col="red")
```



Looking at the residuals plot it is evident that there is no serial correlation as we want a scattered correlation for our errors.

DW test

```
# perform DW test
durbinWatsonTest(reg3)
# correct for potential autocorelation with NW appraoch
NWVcov=NeweyWest(reg3, lag=1, prewhite=FALSE, adjust=TRUE)
coeftest(reg3, vcov=NWVcov)
```

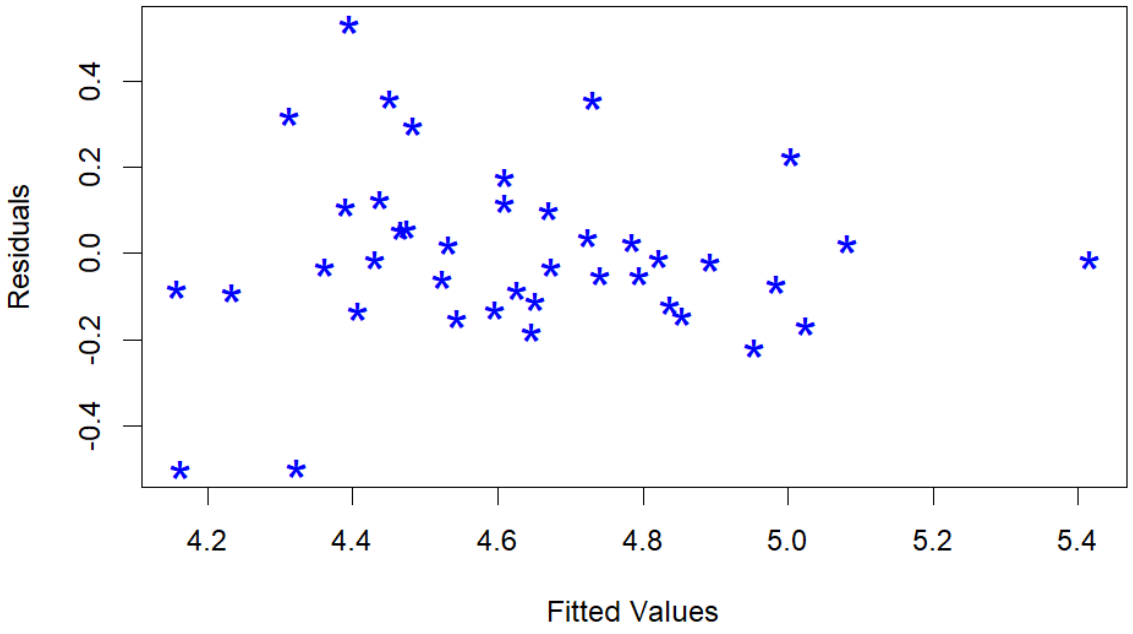
t test of coefficients		
	Estimate	
(Intercept)	3.69E+00	
Minutes	2.45E-04	
ga	3.13E-02	
Creativity	2.78E-04	
	Std. error	
(Intercept)	1.96E-01	
Minutes	8.97E-05	
ga	6.15E-03	
Creativity	7.53E-05	
	t value	Pr(>|t|)
(Intercept)	18.8617	2.20E-16
Minutes	2.7266	0.009722
ga	5.0814	1.10E-05
Creativity	3.6886	0.000721
(Intercept)	***	
Minutes	**	
ga	***	
Creativity	***	

Results have improved such as creativity which it more statistically significant as it now is below the 1% and is very close to zero but also there are less chances of finding errors as they are all close to zero because creativity will impact Total points more than minutes because all players will play but those who are more creative tend to score more points as they may complete forward passes leading to a goal or have a high number of successful dribbles this shows that creativity can be more important than minutes

Heteroskedasticity

```
##obtain fitted values
fitted4=fitted(reg3)
# perform visual diagnostics
par(mfrow=c(1,1))
```

```
plot(errors3~fitted4, xlab="Fitted Values", ylab="Residuals", pch="*", cex=2,
    col="blue")
```



No evidence of Heteroskedasticity as there is no increasing/decreasing shaped bell whilst it is scattered so there wont be any need to run any further tests because heteroskedasticity is bad for regression and because it isn't evident this is a positive sign.

Conclusion

In conclusion the results from all the test supported my theory that key performance indicators such as creativity and Goals+Assists do affect total point and they have a positive relationship as backed up by the results, but also minutes played had an big impact which I didn't expect as I thought it would be how they perform on the field not how long they are playing for, but evidently the longer the players are on the field the more likely they could get

chance to have an impact which is always increasing their points, whilst I had to remove threat as a variable which I thought would have a positive impact as it means you pose a danger to the opposition and highly likely to score or assist, however this was not the case and it did not have an impact on performance which resulted a slight decrease in Total points, one thing that can be done differently and for further analysis is I could do three separate test on players from each position meaning; defender midfielder and attacker, and compare them to each other and how each performance differs from all positions and which gain most points from which variables, also one thing to consider is that because football is a team sport it is also heavily reliant on the team doing well and so we see players from the same teams getting most of the points, but because my focus was on the premier league players this meant that most results were unpredictable and the points distribution always differ every week, meaning that I was able to get players from different clubs and positions to make it a fair test.