**Does Past Help Predict Future in Fantasy Premier League?**

Shivam K C, Duc Trinh

Department of Mathematics and Computer Science, College of Wooster, Wooster, Ohio

***Abstract***

Fantasy Premier League (FPL), the biggest fantasy game in the world with over six million players, is an online game where people assemble a team of real-life soccer players and score points based on the actual performance of these players. Since the points heavily depend on unpredictable factors like goals and assists, the points become harder to get. We used the very little research (related to FPL) available to find factors that best predict points a player gets in FPL. After using several methods including random forest and regression, we have found the best predictors and have gotten a glimpse of how to win FPL.

***Introduction***

People feel good when their predictions turn out to be true. The feeling is even better when one wins a bet against his friends. This feeling is driving not only the purpose of our research, but also FPL. FPL is an online fantasy game where people assemble a team of real-life soccer players and score points based on the actual performance of these players. The points are based on the playing time, the number of goals, assists, fouls, etc. of the real life game. For instance, a player gets four points in FPL for scoring a goal and two points for playing 60 minutes in a match. The goal of FPL is to get more points than other participants. Since the points heavily depend on unpredictable factors like goals and assists, the points become harder to get. That is why the main goal of the research is to find a model that best predicts the points (FPoints) a soccer player gets in a given game week (GW) of FPL.

Either some FPL players have found the model and they are keeping quiet or there is not much work done in this field. The reason why is that the articles that we researched are not directly related to FPL. An exception is “The Wisdom of Smaller, Smarter Crowds” by Goldstein et al., which recommends making new teams based on popular players (488). Basically, we used other articles to help us guess and narrow down the variables that can predict the FPoints better. According to Matthews et al., the number of minutes a soccer player plays is crucial (cite). Similarly, Lago-Peas et al. found that in a soccer game the variables that discriminate between winning, drawing and losing teams were the total shots, shots on goal, crosses, crosses against, ball possession and venue (288). These research articles motivate us to ask the following research questions:

* What factors best predict FPoints?
* What factors best predict Dream Team?[[1]](#footnote-1)

***Data***

The data were collected of all the soccer players playing game week (GW) 15, 17, and 34 of Barclays Premier League (BPL) for 2018-19 season by OptaSports, an international sports analytics company (Statistics Explained). GW 15 and 17 data were used for training models and GW 34 data was used for testing. The data was made available in the official website of the Premier League. However, it could only be used through data scrapping. Vaastav Anand had cleaned the data and posted it posted on github (cite). A positive thing about the dataset was that there were no missing data. The reason may be that the data is from the official source which keeps track of everything. The target population for the analysis was the soccer players playing BPL for 2018-19 season. The sample was a good representative of the population because: it consisted of almost all the players' performances in three different GWs of FPL and the data were collected in the middle of the season, which is less erratic than the beginning of the season, as the teams would have settled down.

The sample could be made better by adding more GWs of data as one would have more data points to train a model on.

For each GW there were two datasets: a dataset with information of players from GW 1 to *Nth* GW (Up-to Data) and a dataset with information of players for *Nth* GW only (Only Data). Here *N* is the specific GW we are looking at and we are predicting *N+1* GW. For example, if N is 33, then the Up-to Data is the dataset from GW 1 to GW 33 and the Only Data is the dataset for only GW 33 in order to predict GW 34. We combined the two datasets and got a dataset which included 1163 samples and 64 variables. Other notable information of the dataset is summarized in Table 1.

**Table 1.** The descriptive summaries of key variables. All variables are numerical except Dream Team, which is binary.

|  |  |
| --- | --- |
|  | Definition |
| FPoints (points) | points a soccer player gets in future GW |
| BPS | utilizes a range of statistics supplied by Opta that capture actions on the pitch, to create a performance score for every player |
| Bonus (points) | players with the top three BPS in a given match receive bonus points - three points to the highest-scoring player, two to the second best and one to the third |
| Threat | a value that examines a player's threat on goal |
| Influence | evaluates the degree to which that player has made an impact on a single match or throughout the season |
| Selected by Percent (%) | percent of people that selected the player |
| Minutes | minutes a soccer player has played in the season |
| ICT Index | a single score for a soccer player for three key areas – Influence, Creativity and Threat |
| Creativity | assesses player performance in terms of producing goalscoring opportunities for others |
| Completed Passes | the number of passes a player completed in last GW |
| Clean Sheets | clean sheet is an event when a player’s team does not concede a goal; this variable is a number of such events |
| Transfers Balance | it is the difference of number of FPL players that transferred in the player and number of FPL players that transferred out the player in a GW |
| Dream Team (Yes-1 or No-0) | whether a player made it to the dream team in the future GW or not |

***Preprocessing of Data***

We had to alter the combined dataset so that it would allow us to increase the number of observations by enabling us to make comparisons among players across different GWs.

*Up-to Dataset:*

Let’s take an example of a variable from Up-to Dataset: total BPS (explained in Table 1) of a player is the player’s total BPS from GW 1 to the *Nth* GW. Instead of total BPS, we used average BPS:

Average BPS = Total BPS up to Nth GW / N

We did this to create average values for total Points, Goals Scored, Assists, Minutes, Goals Conceded, Creativity, Influence, Threat, Bonus, BPS, ICT Index, Clean Sheets, Red Card and Yellow Cards in Up-to Dataset.

*Only Dataset:*

Let’s take an example of a variable from Only Dataset: Assists mean number of Assists (explained in Table 3) of a player in N GW only. We recreated Assists as:

Assists = Original Assists – Average Assists from Up-to data

This way we could determine how far from a player's average is he performing. We did this to recreate Assists, Bonus, BPS, Clean Sheets, Creativity, Goals Conceded, Goals Scored, ICT Index, Influence, Minutes, Red Cards, Yellow Cards, and Threat in Only Dataset.

The initial and final number of observations in the analysis are reported below:

* Initial observations in GW15 (both Upto and Only) is 568
* Initial observations in GW17 (both Upto and Only) is 605
* The final observations in combined dataset is 1163

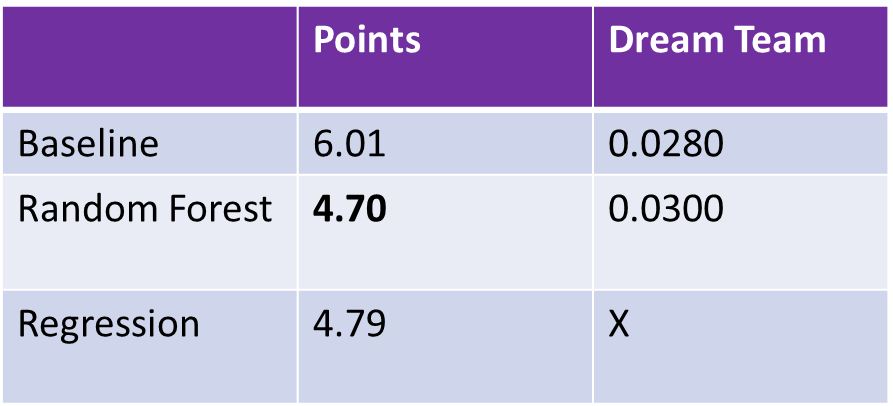
***Materials and Methods***

Below is our thought process for using statistical methods and machine learning algorithms to answer our research question:

* We created baseline model for FPoints and Dream Team (email)
  + For FPoints, we used Average Points of a player from Up to Dataset for the baseline model because it is a simple model which says that a player will perform in the future how he has been performing. (change this)
  + For Dream Team, we used the mean of the Dream Team column because it shows the likelihood of a player being in a Dream Team.
* We used Random Forest with k-fold cross validation to train a model on the combined data and to predict FPoints and Dream Team of a new GW data.
  + We used this because it is robust in making predictions as it takes numerous predictors into account. It is also appropriate for classification problems. [literature?] Also, we wanted to make sure that our results were not a fluke.
  + We compared the Mean Squared Error (MSE) of the model vs the baseline model to see how well the model predicted FPoints and Dream Team.
* We used the top ten predictors given by Random Forest for FPoints and ran multiple linear regression models with FPoints as the target variable against its top 10 predictors. Correlations were calculated and scatterplots were created to explore the association between each of the response variables with FPoints. Predictors were narrowed down to using summary wise regression. Simplicity and adjusted-R squared were used to find to the best model. The t-tests of the coefficients were examined to determine the significance of each variable in the model. Also, F-statistic was used to determine the usefulness of a model. Where appropriate, confidence intervals and slopes were reported for the best predictors. Conditions of inference were also used to distinguish between the models. This was done because the Random Forest model has several insignificant factors in it, even though it is robust. It is complicated to include every variable available in a model. So, we wanted to simplify and see if we could create a multiple linear regression model that predicted FPoints better than or at least at the same level as Random Forest. To check that, MSE of the best multiple linear regression model was calculated and compared against the MSE from random forest and baseline model.

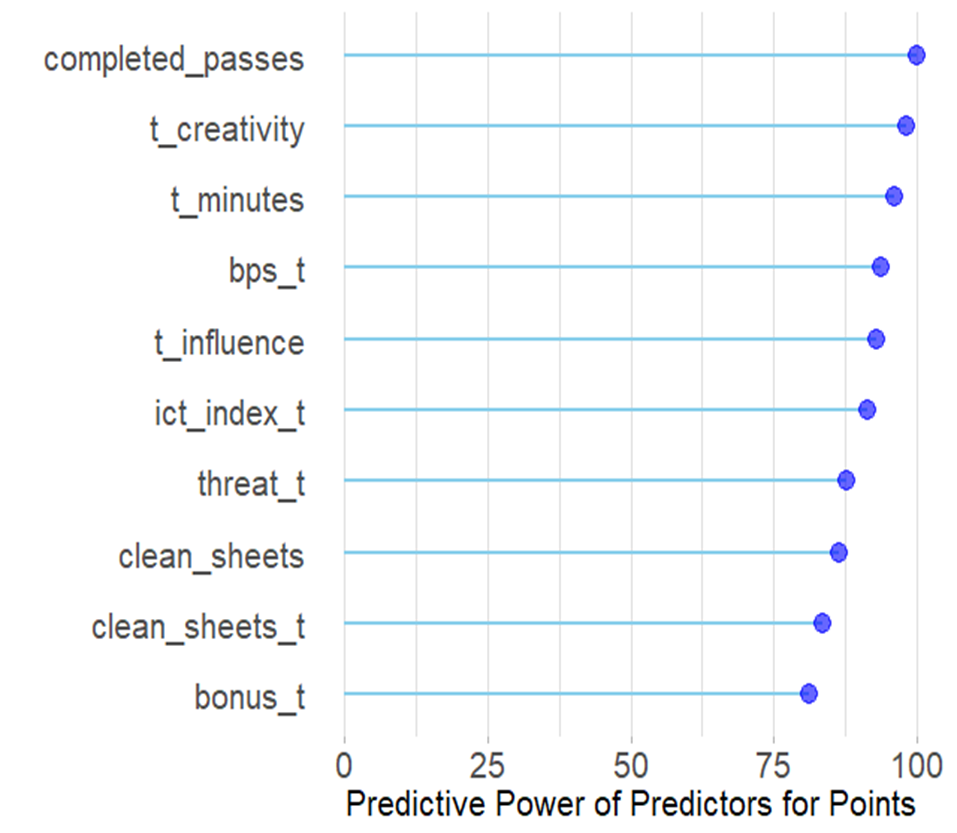
***Results***

***Predictive Power of our Models***



The above table represents the predictive power of our models in terms of MSE. We can see the Random Forest (with k-fold cross validation) model predicts FPoints the best since it has the least MSE. Multiple linear regression model comes very close to challenging the Random Forest model with MSE being only higher by 0.02%.

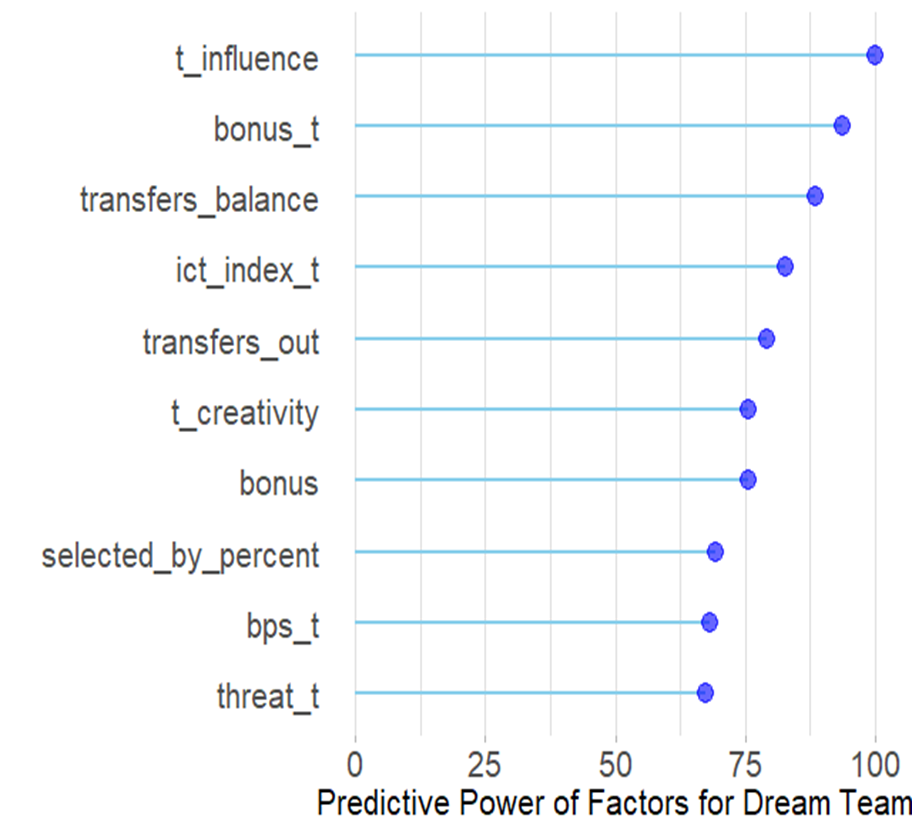
For Dream Team, we could not improve on the baseline model, but we did take a close look at its best predictors from the Random Forest model just like we did for FPoints.



***Best Factors Predicting FPoints***

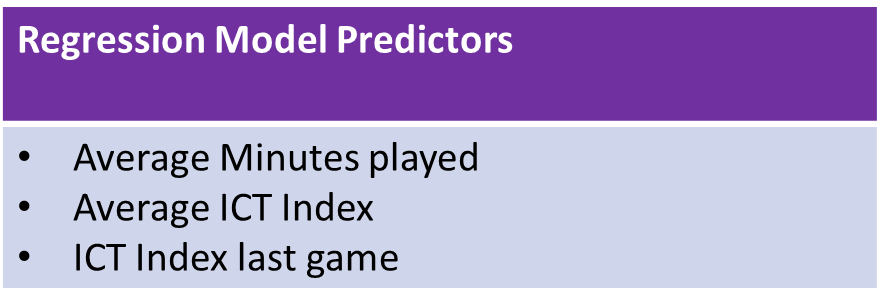
Here variables with “t” are from the Up-to dataset and variables without “t” are from the Only Data.Here we see the top ten best factors predicting FPoints given by the Random Forest (RF) model. Most of them (60%) are attacking factors and factors measuring players’ influence in the game: for instance, Completed Passes, Creativity, Influence, ICT Index, and Threat. Only couple of them (20%) are defensive factors like Clean Sheets.

***Best Factors Predicting Dream Team***



All the predictors above with “t” in it are Average of that variable (Up to Dataset) and without “t” are just the variables of recent GW (Only Dataset).Here we see the top ten best factors predicting Dream Team given by the Random Forest (RF) model. We find that the best factors for FPoints and Dream Team do not match: 30% of the best factors for Dream Team are factors measuring the popularity (Transfers balance, Selected by Percent, and Transfers out) of the players among those who play FPL. That’s 0% for FPoints. 40% of the factors are factors measuring players’ influence in the game: for instance, Bonus, Creativity, Influence, and ICT Index. The rest 30% are attacking factors. Obviously, we are aware of multicollinearity in the Random Forest model. For instance, Transfers Balance, Selected by Percent, and Transfers Out all appear together in a model when they are highly correlated with each other. However, this is not the case with the following regression model.

***Multiple Linear Regression Model***



The results of the methods used for regression model (under Methods Regression) are a model that uses the factors of Average Minutes, Average ICT Index and ICT Index to indicate what a player’s FPoints is expected to be in the next game week. The model is a linear regression on the FPoints variable of the following form:

*FPoints = β0 +Average Minutes⋅β1 + Average ICT Index⋅β2 + ICT Index⋅β3*

Our first reaction was if Average ICT Index and ICT index are correlated, then there must be multicollinearity problem related to the model. We ran the VIF (multicollinearity) test. From the table, you can see that each number are less than 5. Thus, there is no problem of multicollinearity.

t\_minutes ict\_index\_t ict\_index

2.674505 2.669178 1.008134

Table 3 shows the estimated parameter, 95% confidence interval, standard error, and the corresponding p-value for each factor. The coefficients for the three variables are found to be significant at the p < 0.001 level of significance. Therefore, Average Minutes, Average ICT Index and ICT Index are significantly important predictors for FPoints in the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Terms | Estimated Parameter | Standard Error | p-value | 95% Confidence Interval |
| Intercept, β0 | 0.122 | 0.110 | 0.267 | (-0.085, 0.345) |
| Average Minutes | 0.024 | 0.004 | <2x10-16 | (0.012, 0.028) |
| Average ICT Index | 0.221 | 0.0587 | <2x10-16 | (0.195, 0.429) |
| ICT Index | 0.185 | 0.0281 | <2x10-16 | (0.129, 0.240) |

**Table 3.** The summary of the statistical results for the constructed model.

From Table 3, the model has the following form:

*FPoints = 0.122 +Average Minutes⋅0.024 + Average ICT Index⋅0.221 + ICT Index⋅0.185*

From the model there are several expectations indicated by the coefficients associated with each variable:

* If Average Minutes increases by one minute, FPoints will increase by 0.024 points holding other variables constant.
* If Average ICT Index increases by one unit, FPoints will increase by 0.221 points holding other variables constant.
* If ICT Index increases by one unit, FPoints will increase by 0.185 points holding other variables constant.

The model meets only some conditions of inference. Its linearity is good. However, constant variance and normality are problematic. The data fulfills representativeness. However, each data is not independent of another in that if a player scores against another player, the scorer gets positive points, but the defender gets negative points. To see if constant variance and normality of the model would improve, natural log of FPoints was plotted against Average Minutes, Average ICT Index and ICT Index. However, linearity, constant variance, normality and adjusted R-squared got worse and thus was rejected. The final model itself was found to have an adjusted R-squared of 0.222, meaning that it was found to ‘explain’ approximately 22.2% of the variation of FPoints within the data set. From Table 3, we can see that our predictors are statistically significant as their p-values are approximately equal to zero, however the intercept seems to be statistically insignificant as its p-value of 0.267 is greater than zero. Similarly, we can see that our model is useful because the F-statistic of 89.5 is significantly greater than 1. Additionally, we can make the following statements from our model:

* We are 95% confident that if Average Minutes increases by one minute, the average FPoints will increase between 0.012 and 0.028 points holding other variables constant.
* We are 95% confident that if Average ICT Index increases by one unit, the average FPoints will increase between 0.195 and 0.429 points holding other variables constant.
* We are 95% confident that if ICT Index increases by one unit, the average FPoints will increase between 0.129 and 0.240 points holding other variables constant.

We see here the average increase in FPoints is not huge when the factors are considered individually. But we know that together the 3 factors can ‘explain’ 22.2% of the variation of FPoints within the data set, which is almost the same adjusted R-squared the RF model has: 23.3%. In other words, the 3 factors model predicts almost as good as the 63 factors model.

The results are editorialized in the following section.

***Discussion***

According to the results, Rf model best predict FPoints. The ten best predictors in the Rf model are shown in figure…Most of them are attacking factors like, Completed Passes, Creativity, and Threat. Our results agree with Lago-Peñas et al. in that attacking variables determine the course of a game (288): Threat is an attacking variable in the Rf model that best predicts FPoints. So, the results recommend assembling more attacking players in your FPL team than defensive players to score more points. That is why this book recommends having an attacking formation like 3-5-2 or 3-4-3 to win (citation). The results also agree with Matthews et al because they suggested Minutes to better predict FPoints, which we found in both Rf model and regression model.

When we looked at the best predictors from Rf model for Dream Team, we found attacking factors and player’s popularity factors coming into play. Consequently, the best factors for Dream Team and FPoints differed even though we expect them to be same since they are highly correlated. The result agrees with Goldstein et al. (488) because they suggested Selected by Percent to better predict Dream Team, which we found.

We narrowed down from 63 factors to three factors best predicting FPoints using linear regression model. Average Minutes, Average ICT Index, and ICT Index were the three best factors. Using ICT Index to build your team sounds novice because there were no research suggesting that.

The strengths of our analysis are:

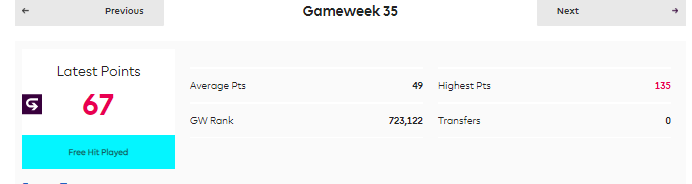
* The predictive power of Rf model predicting FPoints: Since this model has the least MSE of all models, it can be used to predict FPoints.
* Simplicity of the linear model: Our model of FPoints against Average Minutes, Average ICT Index, and ICT Index is simple in that it has only three predictors and the model is easy to interpret.
* Linearity: The linearity of our linear model is really good with data points lying closer to the regression line (see in the Appendix for the linearity of the model).
* Representativeness: The target population for the analysis was the soccer players playing BPL for 2018-19 season. The sample is the soccer players playing BPL for 2018-19 season.

The weaknesses of our analysis are:

* Normality: The normality of the linear model is problematic as most data points do not lie on the line in the normal Q-Q plot. (see in the Appendix for the normality of the model).
* Constant variance: The constant variance of the linear model is also problematic as the data points flare out when going towards right of the residual plot (see in the Appendix for the constant variance of the model).
* We could not improve the baseline model for Dream Team.
* Independence: The independence of my dataset is problematic. If a player scores pass another player, the scorer gets positive points in FPL, but the defender gets negative points. So, each data is not independent of another.
* Random forest models have multicollinearity which is expected.
* Results unreliable: Since most conditions of inference for the linear model are not met, the p-values reported might not be accurate. Thus, the results are not reliable.

***Application***

We were able to apply our findings and play FPL for GW 35. We used the RF model to predict FPoints and Dream Team and assembled a team based on the predictions. The result:

We ranked in the top 11 percent (723,122 out of 6,306,862) which is not impressive. However, an interesting thing happened in the GW. Ayoze Perez, a player who was selected by only 1.4% of the 6 million players, scored a hattrick and made through the Dream Team. Our Rf model predicted that he would be a top scoring player so we had him in our team as you can see in figure…. This does not mean that our Rf model predicted well. Neither does it mean that it did bad. This means we need to build on it. We definitely have a long way to go but this is a good starting place. So, to improve our analysis and build on our research, we suggest the following:

***Future work***

* Introduce more GW data.
* Consider difficulty of games in the future analysis. An attacking player playing against a weaker team might perform better than an attacking player playing against a strong team.
* Definitely use the predictive power of Rf to build your FPL team.
* Try numerous transformations for the linear model to predict FPoints to improve conditions of inference and make results reliable.
* Create a logistic regression model for Dream Team and see how it performs. (We were unable to do it because our primary focus was FPoints.)

***References:***

Lago-Peñas, Carlos et al. “Game-Related Statistics That Discriminated Winning, Drawing and

Losing Teams from the Spanish Soccer League.” *Journal of Sports Science &*

*Medicine* 9.2 (2010): 288–293.

Tim Matthews, Sarvapali D. Ramchurn, and Georgios Chalkiadakis. 2012. Competing with

humans at fantasy football: team formation in large partially-observable domains.

In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial*

*Intelligence* (AAAI'12). AAAI Press 1394-1400.

Anand, Vaastav. “Fantasy-Premier-League/data/2018-19/”, <https://github.com/vaastav/Fantasy-Premier-League/tree/ab719ddfa5c9921fc38cbe6592dd22eaf31b82b5/data/2018-19>

Cummings, Nick. *Mastering the Fantasy premier league: Transfer Hub guide to playing FPL*. 2016.

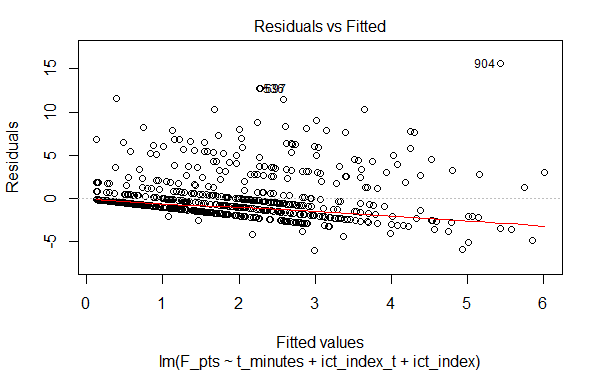
*My Winning FPL Strategy*, 20 July 2018, fplfanatic.blog/2018/07/20/fpl-strategy/.

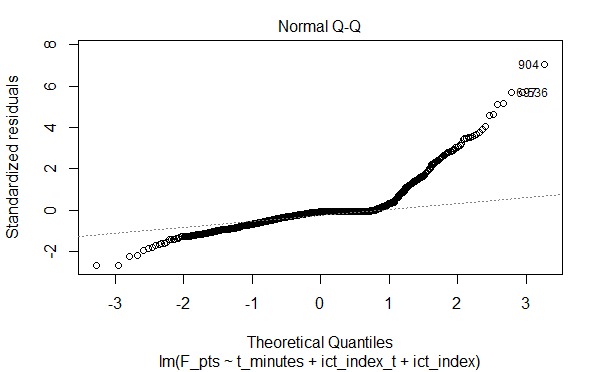
*Statistics Explained*, Barclays Premier League, www.premierleague.com/stats/clarification.

Accessed 20 April. 2019.

"Help." *Fantasy Premier League*, fantasy.premierleague.com/help/.

***Appendix***





1. A Dream Team is a team of players with the highest points in their respective positions: Goal Keepers, Defenders, Midfielders, Forwards. Obviously FPoints and Dream Team are highly correlated, so we do not use one to predict another. [↑](#footnote-ref-1)