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) is the biggest fantasy game in the world, with over six million players, 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 the best predictors of a dream team and the 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.

Keywords: FPL, fantasy, soccer, goals, assists, random forest, regression, predictors, dream team

1. Introduction

People feel good when their predictions turn out to be true. The feeling is even better when one wins a bet against his or her 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 ("Help"). 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. This motivated us 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 we believe this 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). 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. Similarly, Lago- Peñas 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 possessions, and venue (288). These research articles motivate us to ask the following research questions:

- What factors best predict FPoints?
- What factors best predict a Dream Team? [1]

2. Data

2.1 Introduction of Data

To answer our research questions, we got the data of all the soccer players playing GW 15, 16, 28, 29, 33, and 34 of Barclays Premier League (BPL) 2018-19 season.[2] The data was collected by OptaSports, an international sports analytics company and was made available on the official website of BPL (*Statistics Explained*). However, it could only

[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.

[2] BPL is the most competitive English soccer league.

be used through data scrapping. Vaastav Anand had cleaned the data and posted it on github.

A positive thing about the dataset was that there was no missing data. The reason may be that the data is from the official source which keeps track of everything. The target population for our analysis was the soccer players playing BPL for 2018-19 season. The sample was a good representation of the population because: it consisted of almost all the players' performances in multiple GWs of FPL and the data was collected in the middle of the season, which is less erratic than the beginning of the season, as the teams would have settled down.

2.2 Multiple Datasets

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	whether a player made it to the dream team in the future GW (1) or not (0)

2.3 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.

The 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* **(1)**

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 Cards, and Yellow Cards in Up-to Dataset.

The Only Dataset:

Let's take an example of a variable from Only Dataset: Assists mean number of Assists (explained in Table 1) of a player in *N* GW only. We recreated Assists as:

Assists = Original Assists – Average Assists from Up-to data **(2)**

This way we could determine how far from a player's average he is 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.

3. Methods

We used several statistical methods and machine learning algorithms to analyze the data.

3.1 Baseline Model

Firstly, we created baseline model for FPoints and Dream Team.

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 next game how he has been performing leading up to the game.

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.

3.2 Random Forest with K-fold Cross Validation Model

We used Random Forest with k-fold cross validation (RFk) to find the best predictors of FPoints and Dream Team. We used Random Forest because it is robust as it considers all the variables available. We used RFk because we wanted to make sure that our model is low on bias and variance (Gupta).

We trained our model by applying RFk on 80% of the combined data. We wanted to test on 20% of the combined data to find the Mean Squared Error (MSE) of our model. Later on, we did use our model to predict the best team for GW 35 of BPL.

We compared the MSE of the RFk model vs the baseline model to see how well the RFk model predicted FPoints and Dream Team.

3.3 Multiple Linear Regression Model

We used the top 10 predictors given by RFk model for FPoints and ran multiple linear regression models with FPoints as the target variable and its top 10 predictors as the response variables. 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 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 the model. Where appropriate, slopes were reported for the best predictors. Conditions of regression were also used to distinguish between the models.

This was done because the RFk model has several insignificant factors in it, even though it is robust. The RFk model is difficult to interpret as it includes every variable available in a model. 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 the RFk model. To check that, MSE of the best multiple linear regression model was calculated and compared against the MSE from the RFk and the baseline models.

4. Results

Using the methods above we have the following results.

4.1 Best Factors Predicting FPoints

Figure 1 shows the top 10 best factors predicting FPoints given by the RFk model. All the variables except *clean_sheets (-)* have positive correlation with FPoints.

Here predictive power means variable importance which is measured on a scale of 100. The variable importance is determined by the in-built mechanism of Random Forest tool from the caret package of the R programming language.

The variable importance of all the factors are surprisingly similar. For instance, there is not much difference between the variable importance of top 1 factor (*completed_passes*) and that of top 10th factor (*bonus_t*). We speculate that there is a lot of randomness involved predicting FPoints because we have experienced so when playing FPL.

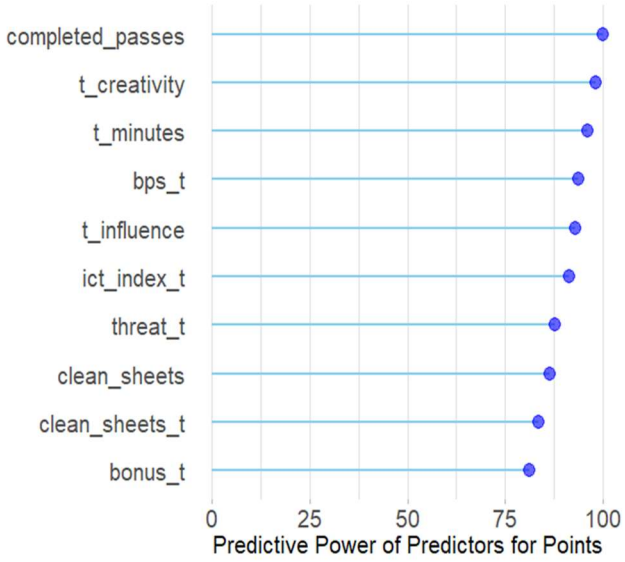

Figure 1: Top 10 factors predicting FPoints. Here variables with "t" are from the Up-to Dataset and variables without "t" are from the Only Dataset.

4.2 Best Factors Predicting Dream Team

Figure 2 shows the top 10 best factors predicting Dream Team given by the RFk model. All the variables have positive correlations with Dream Team.

The main difference between Figure 1 and Figure 2 is that we see 4 new variables appear in Figure 2: they are: *transfers_out*, *selected_by_percent*, *bonus*, and *transfers_balance*. Except *bonus*, the 3 variables are what we called the "*wisdom of the crowd*" because the FPL players' opinions seem to play a role in predicting Dream Team. For instance, *transfers_out* means how many FPL players transferred out a soccer player from their teams in a GW and the variable seems to be one of the best predictors of Dream Team.

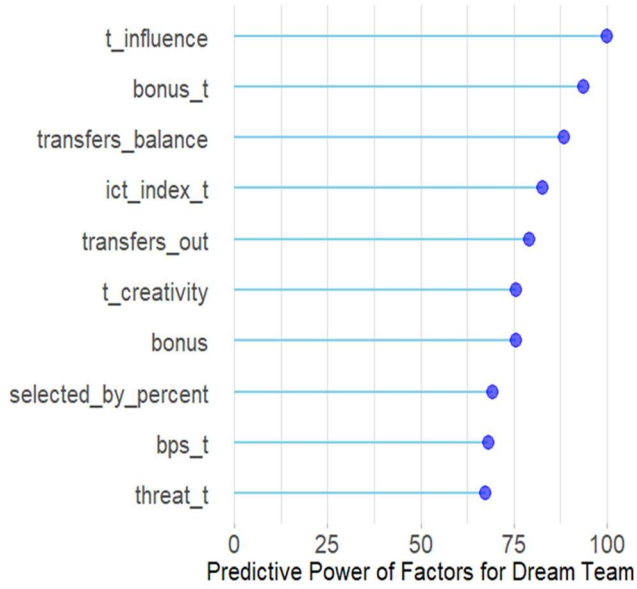

Figure 2: Top 10 factors predicting Dream Team.

4.3 Best Multiple Linear Regression Model for FPoints

The result of the methods used for the regression model (under Section 3.3) is 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:

$$FPoints = \beta 0 + Average\ Minutes \times \beta 1 + Average\ ICT\ Index \times \beta 2 + ICT\ Index \times \beta 3$$

Our first reaction was if Average ICT Index and ICT index are correlated, then there must be multicollinearity associated with the model. We ran the VIF (multicollinearity) test. From Table 1, you can see that each variable has a value less than 5. Thus, there is no problem of multicollinearity.

Table 2: VIF test values.

t_minutes	ict_index_t	ict_index
2.67	2.66	1.00

Table 3 shows the estimated parameter, standard error, and the corresponding p-values 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.

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

Terms	Estimated Parameter	Standard Error	p-value
Intercept, $\beta 0$	0.122	0.110	0.267
Average Minutes	0.024	0.004	$<2x10^{-16}$
Average ICT Index	0.221	0.0587	$<2x10^{-16}$
ICT Index	0.185	0.0281	$<2x10^{-16}$

From Table 3, the model has the following form:

$$FPoints = 0.122 + Average\ Minutes \times 0.024 + Average\ ICT\ Index \times 0.221 + ICT\ Index \times 0.185$$

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 regression. 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 same thing happened for square root transformation of FPoints.

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. 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. Even though the regression model has a low adjusted R-squared, it is an interesting model to work with as we will discuss in the next section.

4.4 Predictive Power of Our Models

Table 4: Mean-squared error of our models

	Points	Dream Team
Baseline	6.01	0.0280
Random Forest	**4.70**	0.0300
Regression	4.79	X

Table 4 represents the predictive power of our models in terms of MSE. We can see the RFk model predicts FPoints the best since it has the least MSE. However, an interesting thing to observe is that the multiple linear regression model, which has only 3 response variables, comes very close to challenging the RFk model, which has a whopping 63 response variables.

For Dream Team, we could not improve on the baseline model. We could have run a logistic regression for it but once we had the best model predicting FPoints we shifted our focus to apply the model to predict the recent GW of FPL.

Our results are editorialized in the following section.

5. Discussion

According to our results, the RFk model best predicts FPoints. If we observe Figure 1 carefully, most of the variables predicting FPoints are attacking variables like, *completed_passes*, *creativity*, and *threat*. Our results agree with Lago-Peñas et al. in that attacking variables determine the course of a game and in our case predicts players' future performances (288). So, our results recommend assembling more attacking players in our FPL team than defensive players to score more points. This is what Nick Cummings and "My Winning FPL Strategy" have been preaching. The results also agree with Matthews et al. because they suggested *minutes* would predict FPoints better, which we found in both the RFk model and the regression model.

When we looked at the best predictors from RFk model for Dream Team, we found both attacking variables 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 the same since they are highly correlated. The results agree with Goldstein et al. because they suggested to build a FPL team based on the wisdom of the crowd (488).

We narrowed down from 63 factors to 3 factors that best predict FPoints: *t_minutes*, *ict_index_t*, and *ict_index*. Using *ict_index* to build a FPL team seems both novice and well backed by research at the same time: it seems

to be well backed by research because *ict_index* indirectly includes attacking variables and *minutes*. It seems novice because we have not found any research that directly recommends us to use *ict_index* to build FPL team.

The strengths of our analysis are:

- **The predictive power of the RFk model predicting FPoints**: Since this model has the least MSE of all models, it can be used to predict FPoints.
- **Simplicity of the multiple linear regression model**: Our model of FPoints against *t_minutes*, *ict_index_t*, and *ict_index* is simple in that it has only three predictors and is easy to interpret. Also, it performs as well as the RFk model.
- **Linearity of the regression model**: The linearity of our regression model is okay with data points lying reasonably close to the regression line (see Figure 1 in the Appendix for the linearity of the model).
- **Representativeness**: The target population for our analysis was the soccer players playing BPL for 2018-19 season. Our sample actually includes the soccer players playing BPL for 2018-19 season.

The weaknesses of our analysis are:

- **Normality**: The residuals of the regression model predicting FPoints are not normally distributed because the data points do not lie on the normal line. (see Figure 2 in the Appendix for the normality of the model).
- **Constant variance**: The constant variance of the errors of the linear model is also problematic as the data points flare out when going towards the right of the residual plot (see Figure 1 in the Appendix for the constant variance of the model).
- **Independence**: The independence of our dataset is problematic. If a player scores against another player, the scorer gets positive points in FPL, but the defender gets negative points. So, each data is not independent of another.
- **Results unreliable for regression**: Since most conditions of regression (normality, constant variance, and independence) for the linear model predicting FPoints are not met, the p-values reported might not be accurate. Thus, the results are not reliable.
- RFk models have multicollinearity which is expected.
- We could not improve the baseline model for Dream Team.

6. Application

Being aware of the limitations of our research, we went ahead and applied what we found. We applied the RFk model predicting FPoints to predict the best team for GW 35 of FPL. Our team is shown in Figure 3.

Our performance is depicted in Figure 4: We got 67 points while the average was 49. We were ranked in the top 11% (723,122 out of 6,306,862), which is not impressive at all. However, an interesting incident happened in the GW 35. Ayoze Perez, a player who was selected by only 1.4% of the 6 million players, scored a hattrick and made it to the Dream Team in the GW. Our RFk model predicted that he would be a top scoring player so we had him in our team (as you can see in Figure 3 with a blue star).

This shows that our RFk model for FPoints needs to be tested and improved upon. 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:

7. Future Work

- Run (test) both our RFk and regression models over several GWs and see what our rank would be.

- Try numerous transformations for the regression model to predict FPoints to improve conditions of regression and make results reliable.
- Create a logistic regression model for a Dream Team and see how it performs.
- Introduce more GW data to increase the sample size.
- 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.

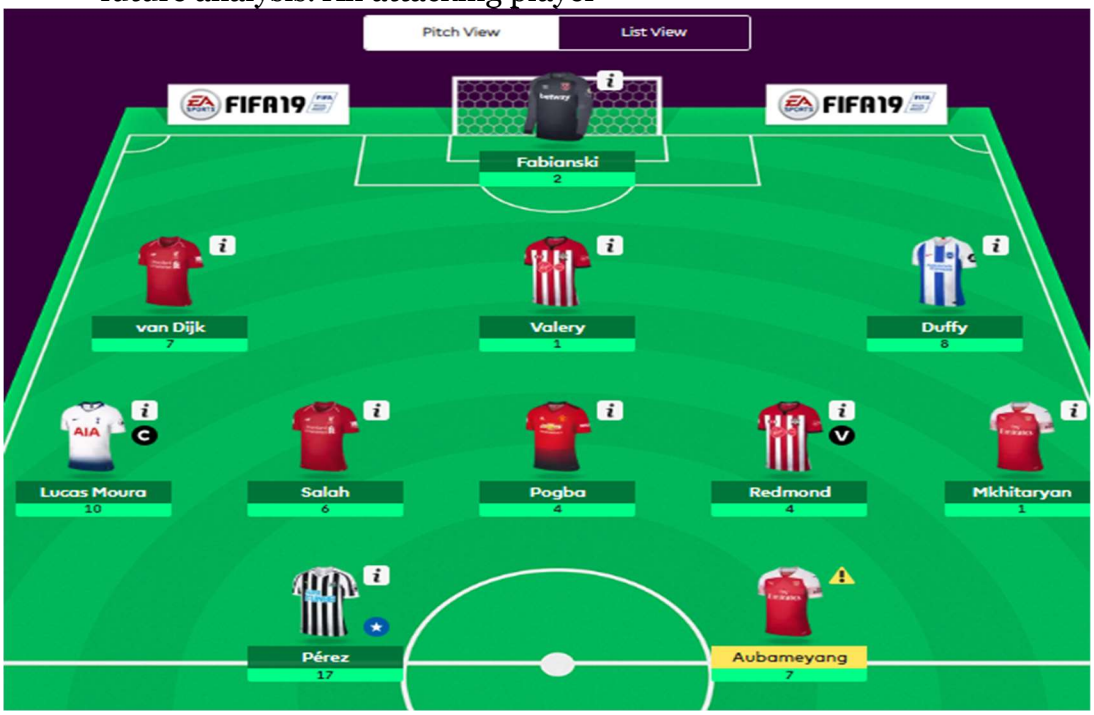

Figure 3: Our team assembled by applying RFk model to predict best team for GW 35.

Figure 4: Our points and rank for GW 35.

8. Conclusion

To summarize, we found that the three best predictors that predict FPoints are: Average Minutes, Average ICT Index, and ICT Index. Similarly, like Goldstein et al. suggested, wisdom of the crowd along with other factors is crucial to predict a Dream Team in FPL (488). Our research is something to be improved upon and we have shown in our paper how it could be done. Overall, we believe our research is a solid foundation for prediction related work in FPL.

References

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.

Daniel G. Goldstein, Randolph Preston McAfee, and Siddharth Suri. 2014. “The wisdom of smaller, smarter crowds.” In *Proceedings of the fifteenth ACM conference on Economics and computation* (EC '14). ACM, New York, NY, USA, 471-488. DOI: https://doi.org/10.1145/2600057.2602886

Gupta, Prashant. *Cross-Validation in Machine Learning*, Towards Data Science, towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f.

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

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.

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.

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.

Appendix

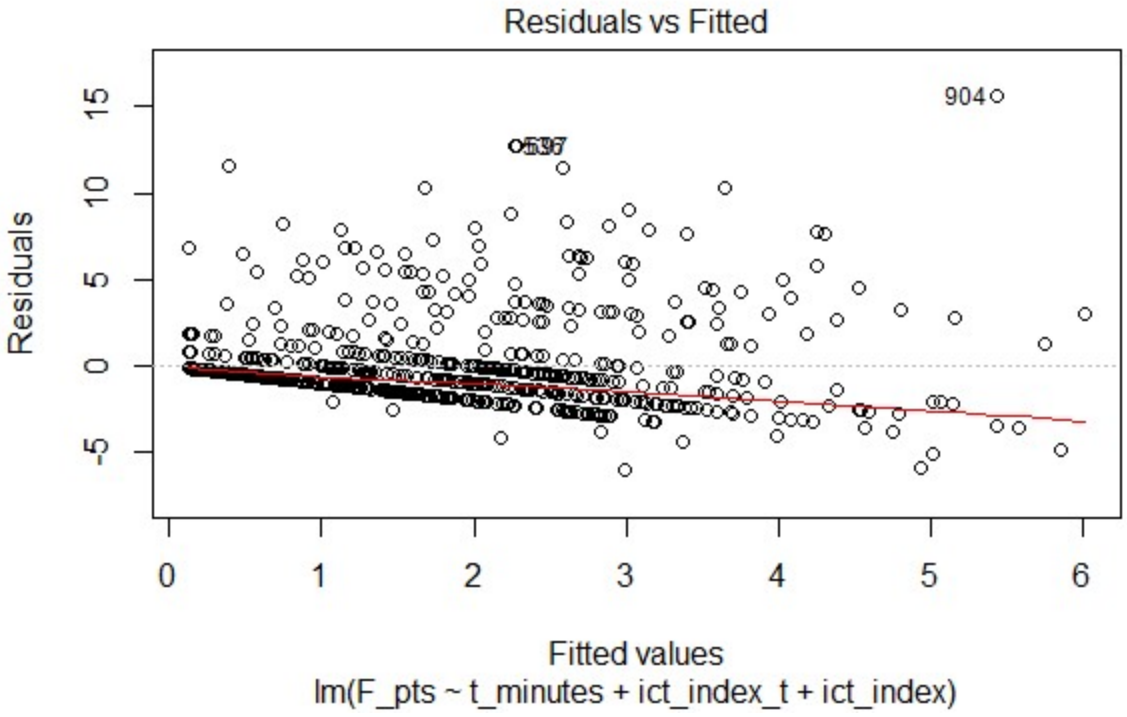

Figure 1: Checking the constant variance of the errors and linearity of the regression model predicting FPoints.

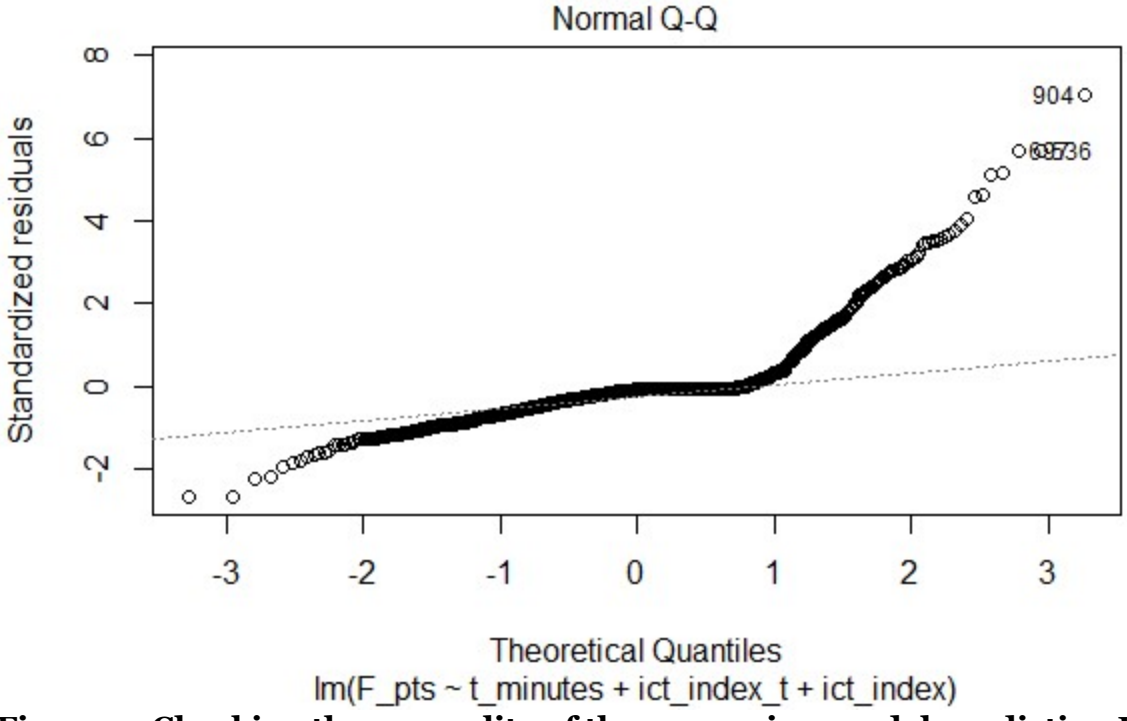

Figure 2: Checking the normality of the regression model predicting FPoints.