# ANALYSING THE CHARACTERISTICS AND VOTING PATTERNS OF THE REFERENDUM

SAVVAS PAFITIS

UNIVERSITY COLLEGE LONDON

2017

## Abstract

The purpose of this study is to carry out an exploratory analysis that will allow one to understand the social, economic and demographic characteristics that are associated with the voting outcome for a ward and provide a model that potentially predicts the proportion of ‘Leave’ votes in different wards. This will be done by fitting a statistical regression model to the data provided by Martin Rosenbaum. The report will focus on correctly identifying the variables that explain the greatest variation and introduce the idea of interaction covariates to allow a greater insight in their behaviour. The data provided includes information for 1070 unique wards with each ward having values for 45 variables that may or may not be relevant in understanding why people voted as they did.

ANALYSING THE CHARACTERISTICS AND VOTING PATTERNS
OF THE REFERENDUM

## I. STUDYING THE DATA

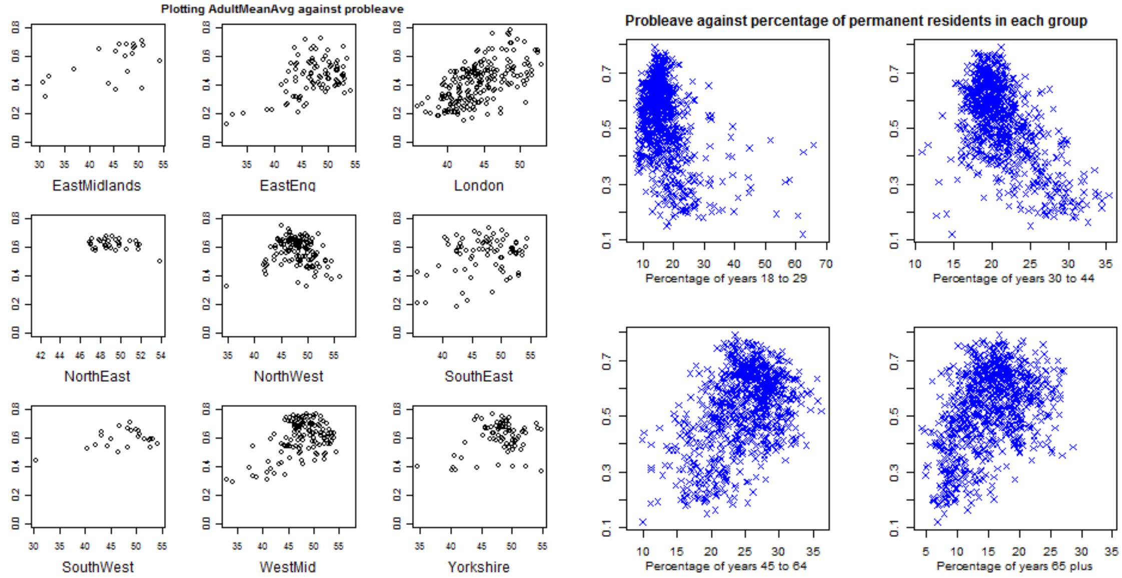
The first step is to clean up the data and study what the variables represent. This will allow the identification of key characteristics and establish a basis that will be viable in predicting the behaviour of the voters, while dismissing variables that appear to be redundant. Through the data, 4 potential groups of variables are identified and can be investigated further. The choice of groups was made by thinking about what socioeconomic factor each set has on the probability of voting 'Leave'. To quantify such probability a new variable, 'probleave', was introduced where it calculates the probability of voting leave in each ward.

$$probleave = \frac{\text{no. of leave votes}}{\text{total number of votes}}$$

### Age.

As the voting patterns are of interest, proceed to isolate the data that has information on people over the age of 18. The readings are subdivided into four groups to encapsulate their behaviour according to the corresponding age intervals. Such intervals were ages '18-29', '30-44', '45-64', '65+'. Plotting the relations of each interval and its corresponding probabilities, some correlation between each group and its associated probability to vote leave can be seen.

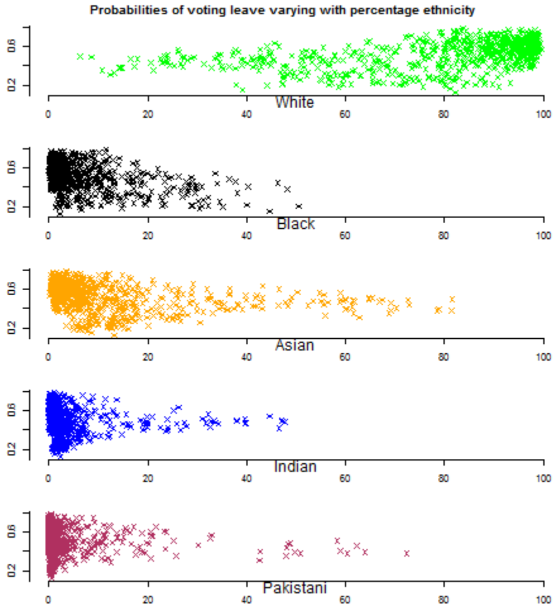
*Figure 1*



Referring to *Fig. 1,* the behaviour of wards with predominantly older voters[1], can be seen to be unlike that of wards involving younger voters. All regions show signs of positive correlation in varying degrees, between age and probability to vote 'Leave'. London shows a slight deviation from the norm[2], with a linear funnel shape[3] whereas other regions have clusters in their scatterplots. The shape of the London graph indicates a markedly different probability of voting leave between young and old. This could be explained by factors such as creative economies and a diverse, student-heavy population. This will be investigated further.

[1] We consider older voters to be those in age ranging from 45 and above.
[2] Refer to final model.
[3] Heteroscedasticity can be inferred from this.

*Figure 2*



***Ethnicity.***

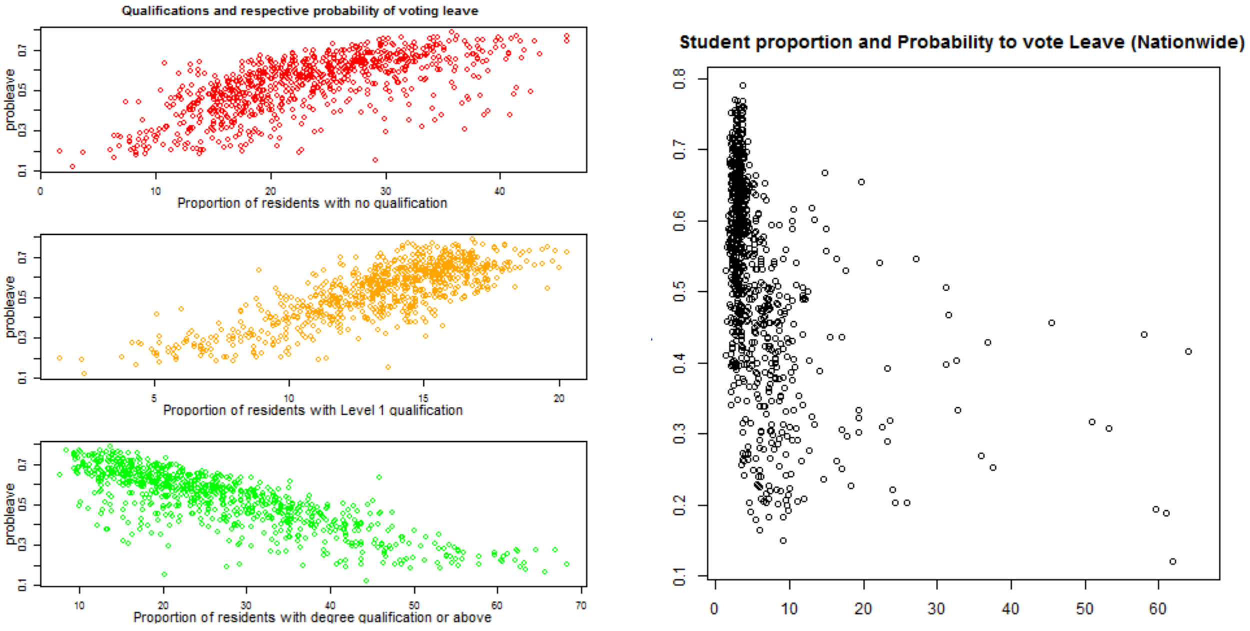
The data includes readings for voters self-identifying under 5 ethnic groups including ‘White’, ‘Black’, ‘Asian’, ‘Indian’, ‘Pakistani’. When plotting the respective probabilities*,* *Fig. 2*, shows significant variation in behaviour between white and non-white ethnicities, with majority white regions being likelier to have voted ‘Leave’. To investigate further, plots have been produced comparing areas with high and low percentage of each ethnicity. Majority white areas produced distinctly higher probability to vote Leave, while in diverse or majority non-white areas ethnicity played a far lesser role. These patterns align with studies that suggest that non-white groups are likelier to vote for the Labour Party, which tends to favour pro-immigration and antidiscrimination policies [4]. Both the Labour and Conservative parties experienced discord in the runup to the referendum, hence the slight long tail in *Fig. 2.* It is worth noting that the higher support for Remain among ethnic minorities could be linked to social class, as non-white groups tend to experience lower income and higher unemployment levels.

[4] BBC Bitesize – Voting Behaviour: *http://www.bbc.co.uk/education/guides/z8447hv/revision*

### *Education and Qualifications.*

Using the resources available, data on 3 unique variables is accessible that can aid in the analysis of the voting behaviour. Information on the percentage of the population with varying degrees of qualifications is collected. The three groups are people with no qualifications (NoQuals), those with GCSEs or below (L1Quals), and those with university-level qualifications (L4Quals_plus). As seen in *Fig.3*, areas with a large concentration of highly-qualified residents show a substantially lower likelihood of voting 'Leave'. This is heavily discussed in Rosenbaum's article, which suggested that people who had few qualifications or none tended to vote 'Leave' whereas more qualified people tended to vote 'Remain'. As for students, who are in the process of obtaining qualifications, *Fig.3* shows that they are highly likely to have voted 'Remain', a trend that holds nationally with little variation between regions.

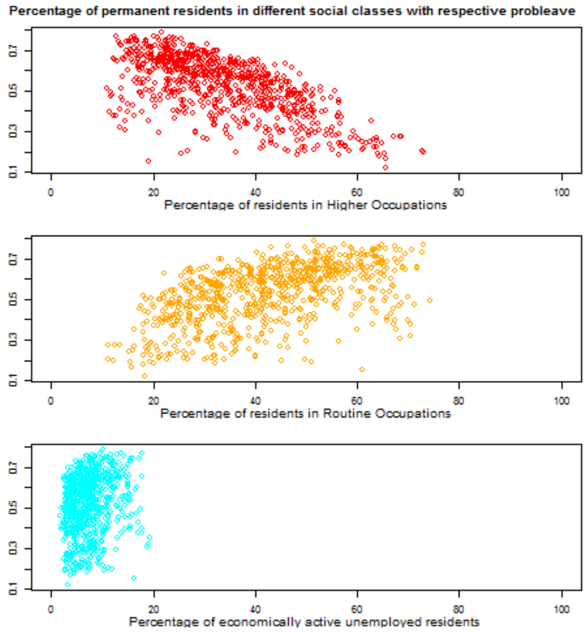
*Figure 3*



***Social Class.***

Social class is considered an important determinant of voting behaviour. This has been discussed in numerous papers, where some analysts argue that class is the most important factor affecting voter behaviour in the UK[5]. In the data, 11 possible variables are able to describe such behaviour, most of which are interlinked. Take for example the two variables for unemployment rate, Unemp and UnempRate_EA. Unemployed people are included in both variables, indicating the existence of aggregated information. To avoid this introducing errors in the predictions combinations that potentially are independent have been considered. It is also worth noting that unemployment rates could potentially be misleading as people tend to be unemployed for short periods of time.

*Figure 4*



[5] EU Referendum – Ashley Kirk and Daniel Dunford

Next, consider how voting behaviour is linked to people's work. *Fig.4* shows that residents in higher-level work behave differently to those in routine occupations and the unemployed. One way to explain the close link between social class and voting behaviour is by studying historic differences in party policies. The Conservatives have a tradition of favouring low taxes and reduced welfare support. These types of policies appeal to wealthier people in social classes A/B who are less reliant on the state. People with higher-level occupations are likely to possess more advanced skills and qualifications, which links to the correlation we previously observed: the lower likelihood of highly-qualified residents to have voted 'Leave'. As the red graph shows, areas with a higher percentage of highly skilled workers voted 'Remain'. Such areas are likely to have strong financial, manufacturing or service industries.

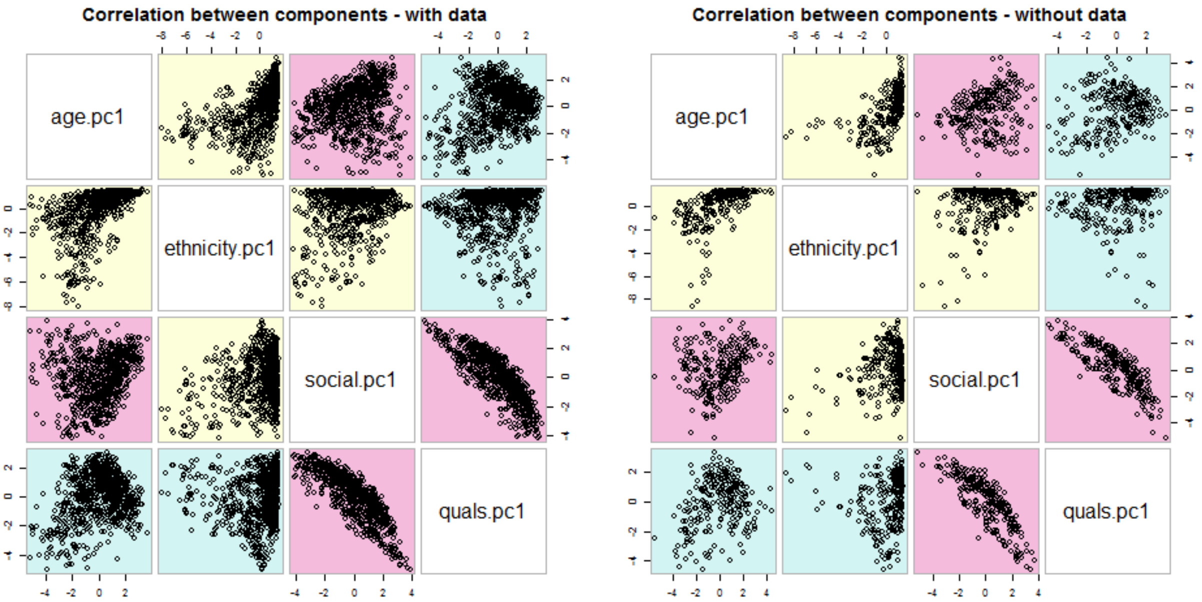
## II. BUILDING A MODEL

The aim is to achieve the greatest explained variation with the fewest variables. Hence principal component analysis has been conducted to find the linear combination of the set of variables that has maximum variance and remove its effect. Through this, mutually uncorrelated variables have been achieved for our analysis. Only interpretable principal components[6] have been used. Such components are as follows:

- age.pc1 – Majority of old residents
- age.pc2 – Majority of middle aged residents
- ethnicity.pc1 – Majority of white residents
- ethnicity.pc2 – Majority of black residents
- quals.pc1 – No/Low Qualifications
- social.pc1 – High level occupation

[6] Other principal components have been omitted due to lack of interpretation.

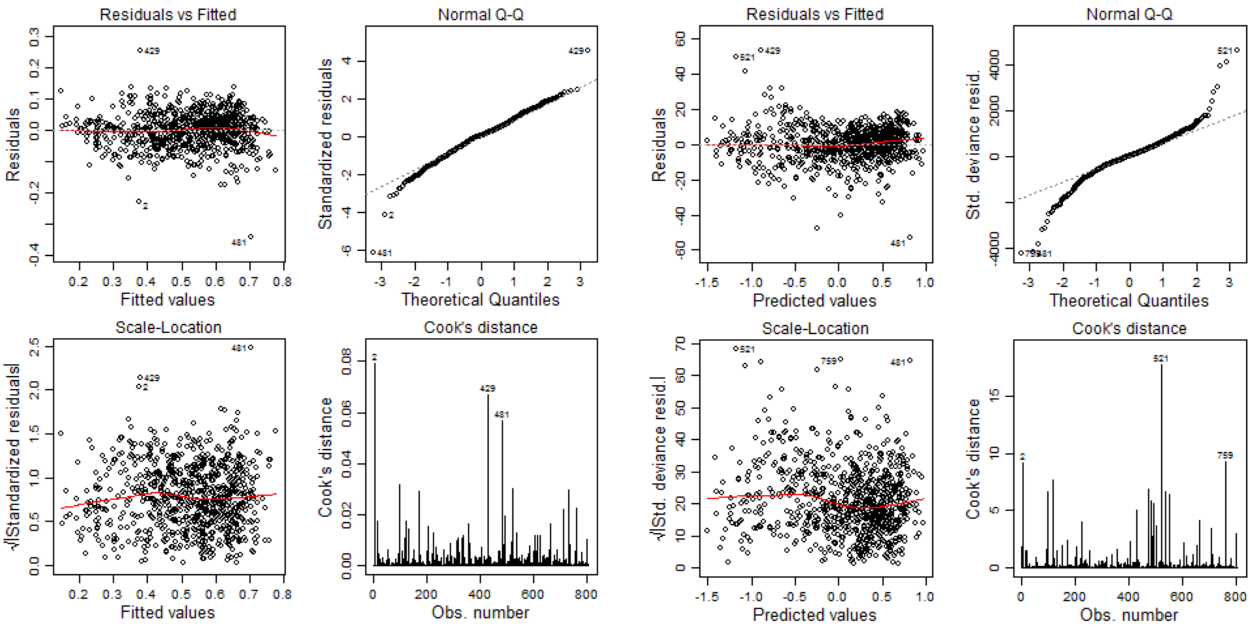
*Figure 5*



Using the finalised principal components, two pair plots have been produced, *Fig.5*, to identify the form of these covariates in respect to their siblings. This was done with respect to two data frames, one including and one excluding ‘Leave’ count values. Although one plot shows more bold clusters in their patterns this is due to frequency rather than change in behaviour. Remember, one data frame has 803 observation and the other has 267. This provides some evidence that our principal components will behave well when predicting the data without the ‘Leave’ counts. Intuitively the red sections indicate strong correlation and yellow and blue tints indicate moderate and low correlation respectively. Some results are expected, such as the high linearity in social.pc1 and quals.pc1 which naturally are related as one measures highly skilled workers and other measures residents in higher level occupations.

Two different types of models have been considered: linear and general linear. Each model was formulated by gradually introducing covariates in a step-like method, allowing for greater control in refining each iteration as the effect of each covariate can be identified. The idea of interaction covariates has been rigorously examined and yielded positive results when applied. Particularly, both types of models fit the data much more closely, yielding improvements in R-squared values and in residual standard deviation. The general linear model initially used the binomial distribution, but its results were not on par with the linear one. Not only did outliers have great impact on its performance but it also was experiencing lack of fit between few wards that the linear one had no substantial problems with. This can be seen in *Fig.5,* in the ill-fitting Normal Q-Q plot but more so in the scale of both Cook's Distance and the Residual plot. The Gaussian family has then been tested with positive results indicating again that the linear model is the proper model to use. The linear model can be seen in the left part of *Fig.5* and the general linear on the right.

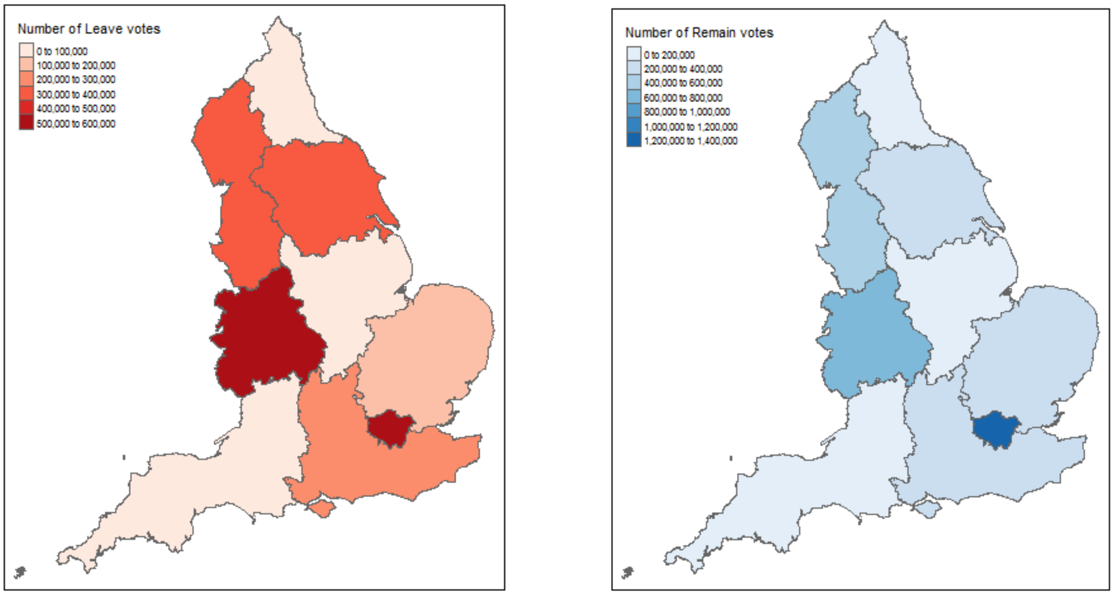
*Figure 6*



## III. FINAL MODEL

To properly describe the final model, first discuss the case of London. London produced interesting results from the start and it was prudent to study it separately. Some reformulation of the data has been carried out such that information on regions is accessible rather than wards. This was done by combining all the wards into their respective regions. This allows the production of a map[7] indicating 'Leave' and 'Remain' votes for each region, *Fig.6*.

*Figure 7*



London is seen to be influential in both the 'Leave' and 'Remain' camps, but voted more heavily to remain more than other regions. The 'Remain' campaign dominated in London leading to some wards to have 'Remain' majorities of over 75%. This cannot be said for most other regions. It is

[7] Spatial dataframe obtained from *http://www.natureonthemap.naturalengland.org.uk/*

thus sensible to introduce a final variable that takes into consideration whether the ward in question is in London or not. This was done by introducing a new binary variable, London, to allow its use in the model.

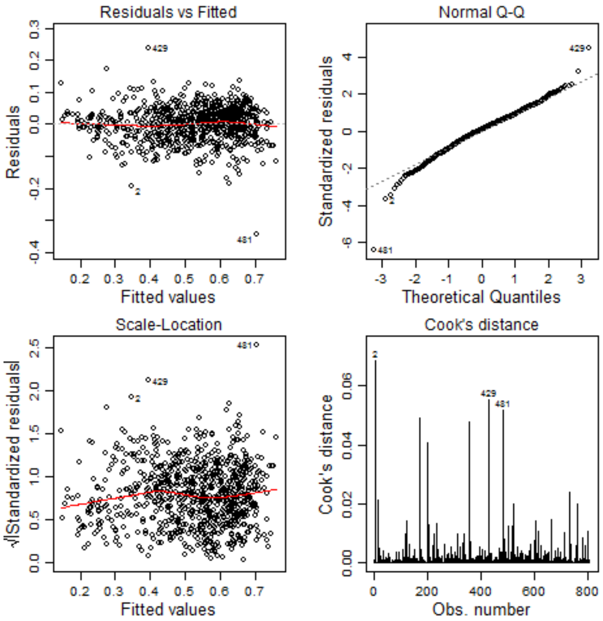
The final model is expressed as:

$$
\begin{aligned}
probleave = {} & 0.538 + 0.010age.pc1 + 0.002(age.pc1)^2 + 0.04ethnicity.pc1 - 0.001(ethnicity.pc1)^2 - 0.003ethnicity.pc2 \\
& + 0.089quals.pc1 + 0.022social.pc1 - 0.003ethnicity.pc1\colon ethnicity.pc2 - 0.004social.pc1\colon ethnicity.pc1 \\
& - 0.013London\colon ethnicity.pc1 - 0.042London\colon ethnicity.pc2
\end{aligned}
$$

| Coefficients | Estimates |
|---|---:|
| (Intercept) | 0.5281571 |
| age.pc1 | 0.0099769 |
| ethnicity.pc1 | 0.0045843 |
| ethnicity.pc2 | -0.0032972 |
| quals.pc1 | 0.0893079 |
| social.pc1 | 0.0221180 |
| I(age.pc1^2) | 0.0021385 |
| I(ethnicity.pc1^2) | -0.0005432 |
| ethnicity.pc1:ethnicity.pc2 | -0.0028878 |
| ethnicity.pc1:social.pc1 | -0.0041544 |
| ethnicity.pc1:London | -0.0134650 |
| ethnicity.pc2:London | -0.0417610 |

Using the table above one can see the effect each covariate had on the probability of voting ‘Leave’. Note that the socioeconomic qualification factor is the most influential — a one unit increase in quals.pc1, which measures the proportion of lowly qualified residents, is accompanied by an almost 0.09 increase in the probability of voting ‘Leave’. The second most influential factor is social class, social.pc1, which indicates the percentage of residents in high-level occupations. On the other hand, the main factor that reduces the probability of voting ‘Leave’ is the interaction covariate, the percentage of black residents living in London. It leads to a probability decrease of 0.04 per unit increase. *Fig.7* shows the general fitness of the final model.

*Figure 8*



Although our model is protected from outliers, inevitably it experiences a few. These outliers were previously tested by removing them from the data and inspecting the suitability of the model, but the results were marginally different. It is better scientific practice not to modify the data. If one refers to the wards that do behave differently, a substantial percentage of them are in the Greater London area. It seems that our new variable, London, is not sufficient to eliminate such behaviour. It is thus sensible to note that the model may not predict the behaviour of London wards with as much accuracy as other regions. Considering that London is an amalgamation of people, ideas, values and social standards, it is to be expected that voting and social patterns are unlike those in the rest of England.