STAT0023 ICA2

Group AW

Introduction

This report aims to investigate the factors that influence the proportion of 'Leave' votes in the Brexit referendum. The analysis aims to identify key variables that may help to build a statistical model to explain and predict voting behaviour.

The dataset includes 45 variables with different categories: *ID, AreaType, RegionName, NVotes, Leave, Postals, Residents, Household*, Average age (*MeanAge* and *AdultMeanAge*), different age groups (16 groups from 0 to 90 and above), Ethnicities (5 ethnics), Housing type (4 types), Qualifications of residents (3 levels), *Students*, Unemployment (*Unemp* and *UnempRate_EA*), Occupation type (4 types), *Density*, Deprivation (*Deprived* and *MultiDepriv*), and Social Grades (3 types) for each of the 1070 electoral wards. At the outset, we considered the demographic, economic, and social factors that may impact voting behaviour.

A new variable *Leave_p* was derived from the proportion of leave votes from total votes: *Leave_p = Leave / NVotes*.

In the exploratory analysis, we examined the data distribution and relationships between variables using graphical and statistical methods. This allows us to identify significant variables, understand the data distribution, and determine any irrelevant variables to *Leave_p* deduct from the dataset. By reducing the number of variables, we aim to create a sensible statistical model that explains and predicts voting patterns in the electoral wards.

Exploratory Analysis

We grouped the same type of variables into each category for ease of analysis.

To begin the exploratory analysis, we created a correlation chart as shown in Fig1. The results indicate that qualification levels have the strongest correlation with *Leave_p*, highlighting the strength of the linear relationship. Additionally, we observed that occupation-related variables also showed a significant correlation with *Leave_p*. Notably, individuals from higher social grades showed a positive correlation, whereas private renting households showed a negative correlation with *Leave_p*, both with similar strength of correlation around 0.2. White and Black both show some correlations with *Leave_p* but in the opposite direction.

To simplify the analysis of contextual data, we initially focused on age-related variables by grouping them into four categories: non-voting age (0-17) as *age0_17*, young adults (18-29) as *age18_29*, middle age (30-64) as *age30_64*, and elderly (65 and above) as *age65_above*. Hence, it will become easier to deal with the important group.

Each graph provides valuable insights into the relationship between the variables in each category and *Leave_p* respectively:

A. *L4Quals_plus* has a strong positive correlation to *Leave_p*, whereas both *NoQuals* and *L1Quals* show a similar trend of strong negative correlation.
B. *C1C2DE* and *C2DE* have a positive correlation while DE has a weak negative correlation with *Leave_p*. Also, the higher the social grade, the higher *Leave_p*.
C. The trend of *HighOccup* contrasts solidly with *RoutineOccupOrLTU*, while the unemployed groups remain at a low *Leave_p*.
D. The proportion of *White* is heavily scattered towards a high *Leave_p* whereas other ethnicities are concentrated towards a low *Leave_p*.
E. Similar to ethnicity, *Owned* has a higher voting leave proportion while other housing types don't.
F. *AdultMeanAge* and *MeanAge* have a positive correlation
G. The variables showing different dimensions of deprivation also have a positive correlation.
H. *Age65_above* has the highest proportion that votes for leave among other age groups.
I. Lastly, postal voters appear more likely to have backed Remain than those who voted in a polling station

The dataset given contains many attributes to handle and each category is highly dimensional, with some closely correlated variables, therefore, each category is then tackled by using PCA (Principal components analysis). This transforms them into a lower-dimensional space while retaining as much of the variability in the data as possible by grouping them in a sensible context. We aimed for new variables with higher than 80% variability when doing PCA for each sector without losing too much data. Thus, those variables are condensed into 12 new variables;

- *Ethinicity_white* (white dominant ward, the lack of ethnic minorities),
- *Ethinicity_black* (black majority ward),
- *Education_low* (lower qualification wards),
- *SocialGrade* (overall social grade),
- *HousingType* (Mostly owned housing in a ward),
- *HousingType2* (Mostly Private rent in a ward),
- *Occupation* (Employed as a majority),
- *RoutineEmploy* (Majority of permanent residents in 'routine' occupations)
- *Wealth* (low rate of Deprivation),
- *AverageAge* (mean age),
- *Age1* (Young - Middle age Adults), *Age2* (All Adults), and *Age3* (Young people in Education).

We've also created a new variable called *Residents_p* which describes the proportion of residents in a household: *Residents_p = Residents / Households*.

A new dataset *VotingData2* is created with variables; *ID, AreaType, RegionName, NVotes, Leave, Postals, Residents, Households, Students, Density, age0_17, age18_29, age30_64, age65_above, Age1, Age2, Age3, AverageAge, Education_low, Ethnicity_black, Ethnicity_white, HousingType, HousingType2, Occupation, RoutineEmploy, SocialGrade, Wealth, Residents_p, Leave_p*.

Moreover, the new data set *VotingData* has been created with only the first 803 wards as we will be focusing on this to build a statistical model.

Fig2 produced the correlation with *Leave_p* for all these variables.

Now we explore more on how regions and area types can be clustered based on their characteristics for further analysis. We did this by splitting the data into subsets by calculating means for all covariates in the initial dataset. These means are then standardised to N(0,1) to be used for Distance Matrix Computation and produce a clustering tree as shown in Fig3 and Fig4. The red lines are where we separate each type.
For the region, we put them into 3 groups as follows:

1. East Midlands, London
2. East of England, South East, South West
3. North East, North West, West Midlands, Yorkshire and The Humber

For the area type, E06 and E08 are Group 1, E07 is Group 2 and E09 is Group 3.
New variables are produced for region and area types called *NewGroup* and *NewArea* respectively.

Model building

We start with choosing the family for the model, and the Central Limit Theorem (CLT) is the reason why the proportion of votes from each ward can be approximately normally distributed. The CLT states that as the sample size increases, the distribution of the sample mean will approach a normal distribution, even if the population distribution is not normal. In the case of the proportion of votes from each ward, there are thousands of votes in each ward, which means that the sample size is large enough for the CLT to apply. Therefore, the distribution of the proportion of votes from each ward can be approximately normal. Thus, we used 'family = Gaussian()' in our model.

We then construct *Model1* with these covariates: *NewGroup, NewArea, Students, Density, Age1, Age2, Age3, AverageAge, Education_low, Ethnicity_black, Ethnicity_white, HousingType, HousingType2, Occupation, RoutineEmploy, SocialGrade, Wealth, Residents_p* against *Leave_p*. We then compare the deviances of the model without each of three of the weakest correlated variables (*Occupation, Residents_p and Age3*) from Fig2. By producing *Model1a* (without *Occupation*), *Model1b* (without *Residents_p*) and *Model1c* (without *Age3*) and compute the difference in deviances with *Model1*. If the difference in deviances is significant, it suggests that the full model has a better level of fit to the data. All

three tests are significant ($p<0.05$), hence, we decided to move forward with the same model, which we now call *Model2*.

Interaction

We then investigated interactions that we can add to our model. According to an article, people who have higher education were more likely to vote leave in low-skilled areas compared to high-skilled areas (JRF, 2016). Indeed, Fig5 shows that the slopes differ with social grade level, having larger differences for the wards with higher qualifications. Therefore, we produce *Model2a* that adds an interaction variable between *Education_low* and *SocialGrade*. The interaction term of this model being significant supports this theory indeed. However, some terms became insignificant by adding the interaction term, so we conduct a chi-squared test to justify whether to keep it. The p-value (0.02675) being lower than 5% suggests that there is enough evidence that a model with interaction is a better model at a 5% significance level.

Additionally, inner cities with high numbers of ethnic minorities voted for Remain (BBC News, 2017). Indeed, Fig6 indicates that *NewArea 1 and 3* have more diversity in a ward and show a lower proportion of *Leave_p* compared to *NewArea 2*. So, we construct *Model2b* with an interaction of *Ethnicity_white* and *NewArea*. By conducting the chi-squared test with *Model2a*, a very small p-value (3.702e-14) suggests we should keep this interaction.

Furthermore, the correlation between ethnicity and the *Leave_p* varied with different regions (BBC News, 2017). It is shown in Fig7 that ethnicity in *NewGroup 3* seems to have almost no effect on *Leave_p*, while the correlation between ethnicity and *Leave_p* is more visible in *NewGroup 1* and *2*. Adding the interaction term *Ethnicity_white:NewGroup* to *Model2c*, the p-value for the chi-squared test was large (0.5875), meaning that there is not enough evidence that adding an interaction *Ethnicity_white:NewGroup* improved the model. So, we will not include this interaction in our model.

Therefore, we construct *Model3* by adding these two interactions to *Model2*: *Education_low:SocialGrade, Ethnicity_white:NewArea.*

Finally, we look for the best link to use in our model. For the Gaussian family, there are 3 links, “identity”, “inverse”, and “log”, so we create *Model3a*, *Model3b*, and *Model3c* using these links respectively. To compare these models, we compare AIC values. This is an indicator which shows how well the model fits the data and the lower the value, the better fit the model. Among our models, *Model3a* gave the lowest AIC, meaning that *Model3a* using identity link is the best fit for the data.
Therefore, we create a final model *Model4* with covariates: *NewGroup, NewArea, Students, Density, Age1, Age2, Age3, AverageAge, Education_low, Ethnicity_black, Ethnicity_white, HousingType, HousingType2, Occupation, RoutineEmploy, SocialGrade, Wealth, Residents_p, Education_low:SocialGrade, Ethnicity_white:NewArea* using Gaussian family “identity” link.

Model Checking

Now we assess the fit for our model by plotting observed values against fitted values. In Fig8, although there are several points further from the y=x line, most of the points are situated near the line indicating that the model closely fits the data. However, we see that more points are concentrated where observed *Leave_p* is higher. This can be explained by the *Leave_p* of the first 803 wards as being positively skewed.

Our generalised linear model has the following assumptions;
Linearity- The relationship between the predictors and the response variable is linear.
Homoscedasticity - Variance of residuals is constant
Normality- The error components are normally distributed.
Independence- Residual errors are independent of each other.

First, to check linearity, we look at the residuals vs fitted values plot in Fig9. The points are equally spread out along the red line where residuals are equal to zero. Since the line is fairly straight and horizontal, linearity can be assumed. In Fig9 and the scale-location plot in Fig11, the red line is relatively horizontal and the distance between points is evenly distributed along the line, suggesting that residual variance is constant. So, homoscedasticity seems reasonable. Regarding normality, the normal QQ plot in Fig10 reveals that although some points at the ends deviate from the Q-Q line, the majority of points are reasonably concentrated about the line, suggesting that normality is likely satisfied. In Fig9's residuals versus fitted values plot and Fig11's scale-location plot, the points are randomly dispersed around the red line, indicating that the independence of the errors is probably sufficient. Nevertheless, we cannot easily assess other forms of dependence.

In Gaussian distribution, overdispersion does not need to be considered as the variance of the response variable is constant and does not depend on the mean while overdispersion occurs when the variance of the response is greater than what's assumed by the model.

Conclusion

In summary, our model shows that the 22 factors had an influence on the proportion of leave votes within a ward. The most significant covariate from the model is *Education_low*, with a negative estimate, stating that wards with lower education levels would show the most significant increase in the leave votes compared to the wards with higher education levels of residents. The statistically significant variables (p-value < 0.05) in the following has a positive correlation with *Leave_p*: *NewGroup3, NewArea3, Age1, Age2, AverageAge, Ethnicity_white, RoutineEmploy, Wealth, and Education_low:SocialGrade*. This means that each unit increase in each variable would statistically increase the proportion of leave votes in a ward. The following significant variables are those with a negative correlation with *Leave_p*: *Students, Density, Age3, PostalsP, Ethnicity_black, SocialGrade, and NewArea3:Ethnicity_white.*

In terms of age, having a larger proportion of adults in a ward would lead to an increased leave vote compared to other age groups.

Factors such as fewer residents per household and a higher proportion of owned households also increased the proportion of leave votes, though their influence was less significant compared to other factors.

Moreover, the proportion of leave votes was higher in *NewGroup3* (corresponding to the northwest region of the UK) compared to other regions. Inner cities also exhibited higher leave vote proportions than other types of areas.

What is more, the correlation between low education and increased leave vote proportions is stronger in the wards with fewer households in higher social grades. Furthermore, the influence of the proportion of white individuals on leave vote proportions is weaker in E07 and E09 compared to E06 and E08.

Limitation

Limited scope: The data set includes variables that may be relevant to the analysis. At the outset, the report considered factors that may have an impact on voting behaviour. However, it is important to note that there may be other variables that were not included in the dataset that could also have an impact on voting behaviour, eg. information on Brexit provided to each region.

Confounding variables refer to variables that are related to both the independent and dependent variables, making it difficult to determine the true causal relationship between them. There may be other variables that are related to both the independent variable (voting behaviour) and the dependent variables (demographic, economic, and social factors) that we did not consider in our analysis. For example, political affiliation or ideological beliefs could be confounding variables that influence both voting behaviour and the demographic, economic, and social factors we analyzed. Failure to account for such confounding variables could lead to biased or inaccurate results.

Assumptions of statistical models: We assume that CLT is applied for this model which is debatable as the response variable may not be exactly a normal distribution, causing errors in the predictions.

Using PCA: it transforms the original variables into new variables (principal components), which we may interpret the meaning wrongly. Also, PCA gives more weight to variables with high variances, which may not always be the most important variables for understanding the data. Additionally, the number of principal components we chose to retain is where it explains more than 80% of the variation in (standardised) variables and the accuracy decreases.

Reference

Matthew G, Oliver H (2016) *Brexit vote explained: poverty, low skills and lack of opportunities,* JRF. Available at; https://www.jrf.org.uk/report/brexit-vote-explained-poverty-low-skills-and-lack-opportunities [accessed: 21 April 2023]

BBC News (2017) *Local voting figures shed new light on EU referendum* Available at; https://www.bbc.co.uk/news/uk-politics-38762034 [accessed: 21 April 2023]

Graph and plots

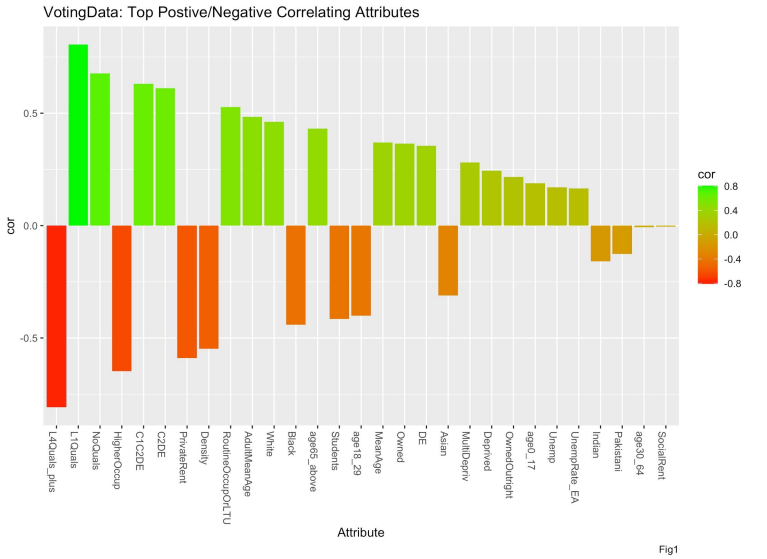


Fig1

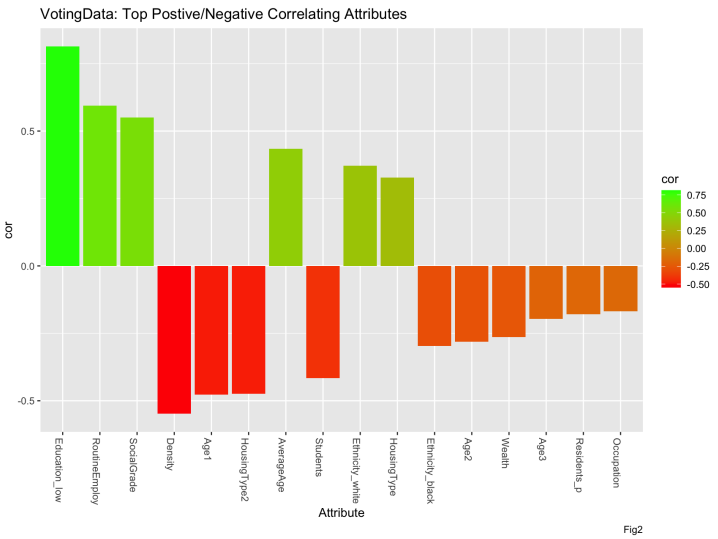
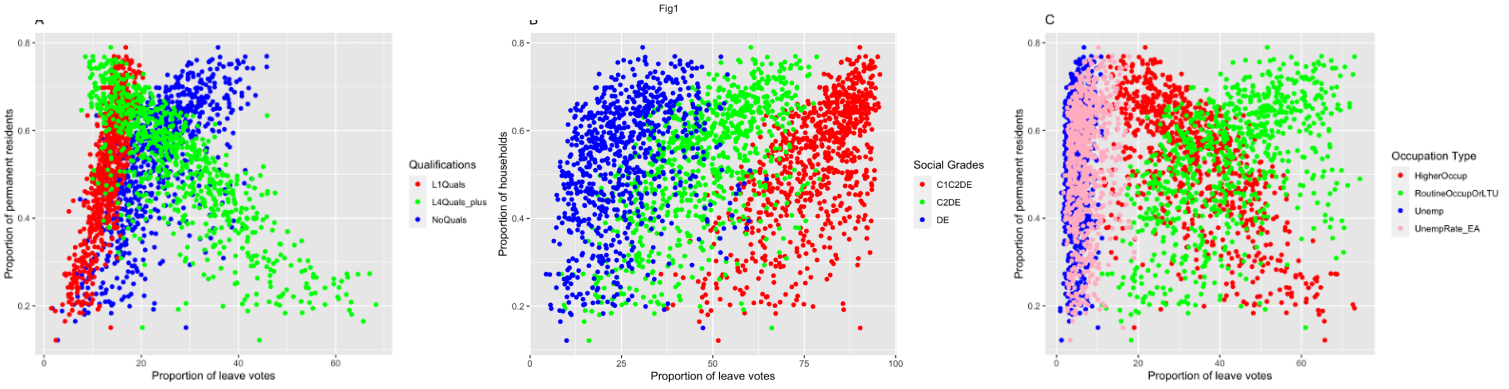
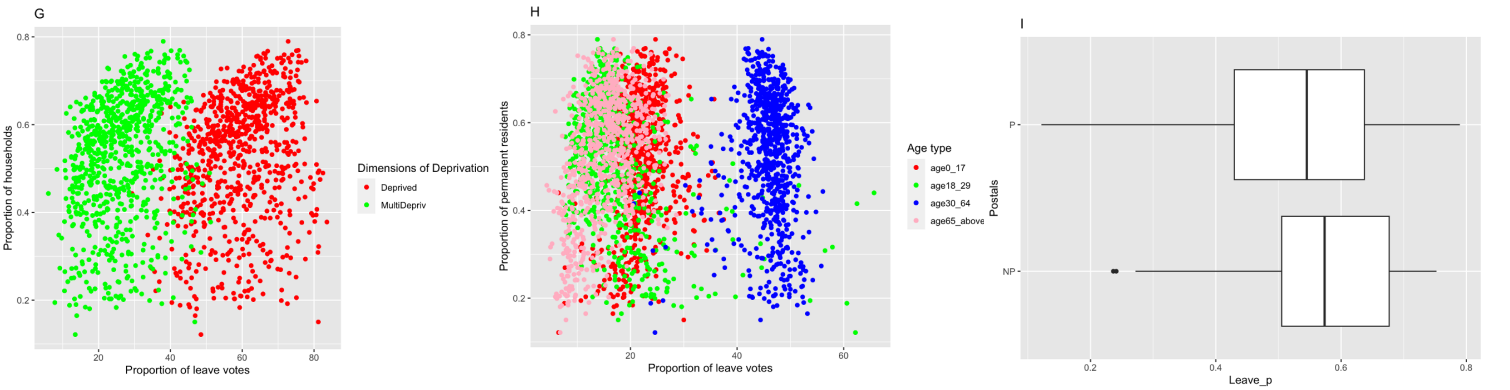
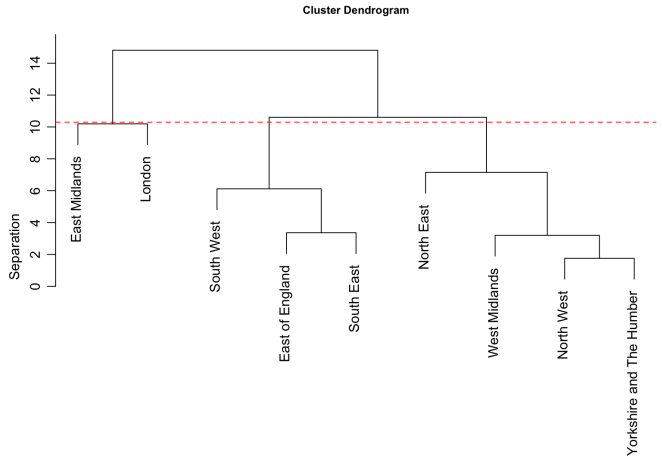


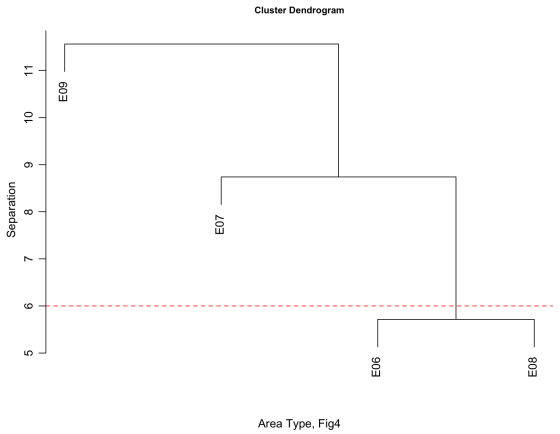
Fig2







Region name, Fig3



Area Type, Fig4

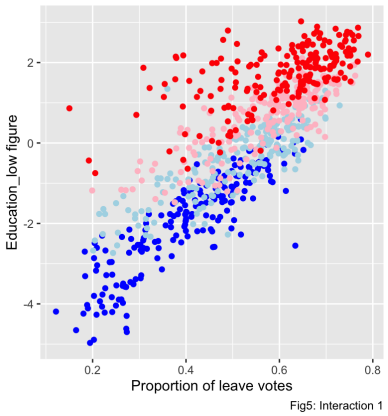


Fig5: Interaction 1

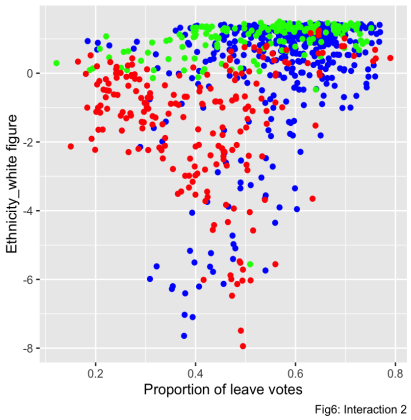


Fig6: Interaction 2

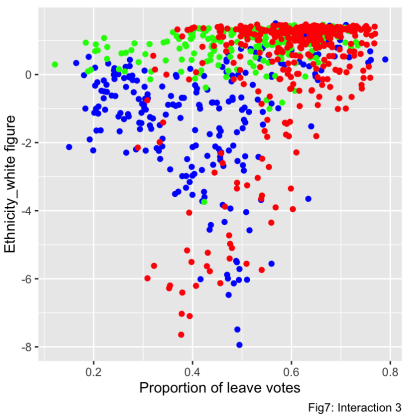


Fig7: Interaction 3

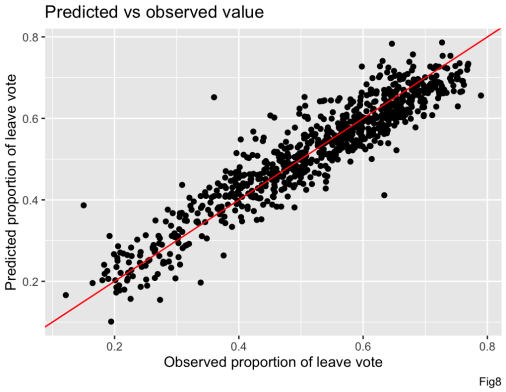


Fig8

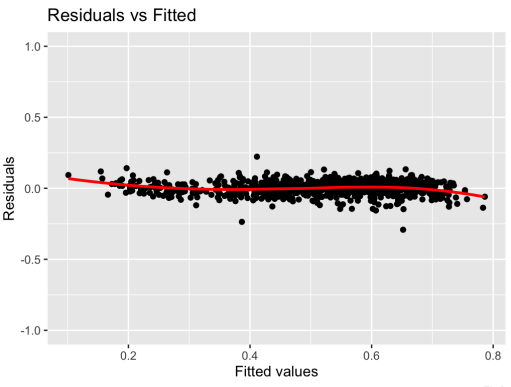


Fig9

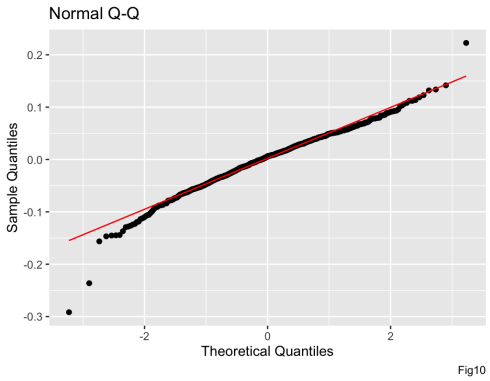


Fig10

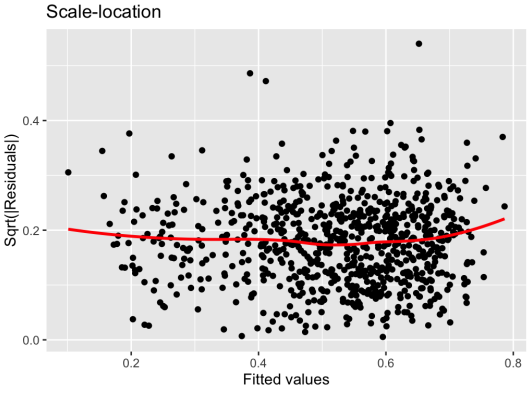


Fig11

Contribution

Both equally contributed