Introduction to the Context and Data

Context, Original Dataset and Adjustment

The given dataset ‘ReferendumReuslts.csv’ contains data on 1070 electoral wards on Brexit giving the numbers of ‘Leave’ and ‘Remain’ votes cast in each ward in 2017, each linked with 45 corresponding variables. It contains 267 ‘Leave’ with missing values, which we removed, resulting in a modified dataset with 803 records. The leave vote data was converted into a ratio to facilitate the analysis of the factors that affected the percentage of Leave votes. The mean of ‘Leave’ across all wards is 0.5482, with a range of 0.1216 to 0.7897, indicating that slightly more residents tended to vote in favour of leaving. The dataset includes a variety of variables related to age, ethnicity, region, social grade, education, employment, and deprivation. According to Martin Rosenbaum’s (2017) report on the referendum of Brexit, we may mainly focus on age, education, region, and ethnicity aspects.

Age-related variables in the dataset include distinct age groups (ranging from 'Age_0to4' to 'Age_90plus'), 'MeanAge', and 'AdultMeanAge', which respectively indicate the percentage of permanent residents in each age group, the mean age of all permanent residents, and the mean age of adult permanent residents.

Ethnicity-related variables in the dataset include 'White', 'Black', 'Asian', 'Indian', and 'Pakistani', which indicate the proportion of permanent residents in each ethnic group.

Region-related variables 'RegionName' and ‘AreaType’ indicate the name of the region where the ward is located and the type of administrative area in which the ward is situated.

For Education, the dataset includes variables such as ‘NoQuals’, ‘L1Quals’, ‘L4Quals_plus’, and ‘Students’ representing the percentage of permanent residents with no

academic/professional qualifications, only 'Level 1' qualifications, educated to the degree level or above, and those who are students.

Approach

The analysis should decrease the candidate variables and recognize significant features of the data that could affect the modelling.

Regions:

In comparison, 'RegionName' provides a more specific geographical location than 'AreaType', which could potentially contribute more to the model. The more detailed the geographic information, the more accurate the model can be, as location can have a significant impact on voting patterns.

Based on the similarity of the pairwise distance of means of leave voting in different areas, we used a cluster tree to group regions with similar leave voting traits into four groups, naming the new variable ‘RegionGroup’: Group 1 is East Midlands, Group 2 is East of England and South East, Group 3 is London, and Group 4 is the remaining regions. We then renamed the variable ‘RegionGroup’.

Age:

We excluded the 18 variables related to age for the 0-17 age groups from the dataset since the minimum voting age in the UK is 18. The omission led to the removal of ‘MeanAge’, which included all age groups of residents.

Furthermore, we observed high correlations among most of the distinct age groups. To address this issue, we decided to use 'AdultMeanAge', which represents the age-related influence factor. This approach ensures that the analysis considers the age group eligible to vote.

Postals:

According to Martin Rosenbaum (2017), individuals who voted by mail have a slightly higher tendency to support remaining in the European Union compared to those who voted in person at a polling station. The boxplot analysis further suggests a slight difference between the two groups. Therefore, it may be beneficial to conduct further statistical research to determine the significance and contribution of 'Postals' to the final model.

Residents & Household:

As ‘Residents’ and ‘Households’ do not show a clear linear relationship with ‘Leave’, it might not be necessary to include both variables in a linear model.

Density:

'Density' has a negative linear relationship with ‘Leave’. Then it may be useful to conduct further statistical research to determine its statistical significance and contribution to the final model.

Ethnicity:

We start by analysing the correlation between each ethnicity. The table shows there is a high negative correlation between "White" and "Asian" (-0.90) and "White" and "Indian" (-0.66).

Additionally, there is a high positive correlation between "Asian" and "Indian" (0.73) and "Asian" and "Pakistani" (0.75). Then we considered using PCA to simplify the data without losing important traits and reduce the effect of multicollinearity since the high correlation indicates the high strength of the linear relationship that might exist between two variables.

Based on the output of the PCA, the first two principal components explain a large proportion of the variance in the data, accounting for 0.6354 and 0.2013 of the total variance, respectively. This means PC1 and PC2 contain a significant amount of information from the original variables.

Therefore, we created new variables ‘EthnicityPC1’ and ‘EthnicityPC2’ containing PC1 and PC2 values for later linear model building.

House owing & renting:

The correlation analysis showed a strong positive correlation of 0.89 between the variables "Owned" and "Owned Outright", suggesting a high degree of collinearity. To address this, we performed principal component analysis (PCA) and found that PC1 explains 94.32% of the total variance. Using PC1, we created a new variable called ‘Owning’.

However, the correlation between ‘PrivateRent’ and ‘SocialRent’ is extremely small (0.08), therefore we decided to keep both variables in the initial model to assess their statistical significance.

Education:

The scatter plot of ‘Students’ (bottom right) in [Figure 1] does not show clear linearity. Furthermore, the correlations between ‘Students’ and the other variables are small. As a result, it may be a good idea to exclude ‘Students’ when creating a model.

However, despite the large correlations among the other three variables, we decided to keep 'NoQuals', 'L1Quals', and 'L4Quals_plus' in the initial model due to the significance of each variable's meaning.

Deprivation:

As the scatter plots [Figure2] of 'Deprived' and 'MultiDepriv' have the similar trend, we investigated that their correlation is dramatically high (0.9746). We decided to omit one of the variables since the new variable created through PCA may be meaningless and hard to understand.

After examining the model to evaluate the performance of these variables, we would decide on which one to omit.

SocialGrade:

The scatter plots of the three social grade-related variables suggest a similar trend in the relationship with ‘Leave’. Additionally, there is an inclusion relationship among these variables. Specifically, households in social grades C2, D, and E are included in the households in social grades C1, C2, D, and E. Similarly, the households in social grades D and E are also included in the household in social grades C1, C2, D, and E. To address issues of inclusion, we have decided to use PCA to reduce multicollinearity, creating a new variable named ‘SocialGrade' to represent the values of PC1.

Occupation:

Upon examining the scatterplot [Figure3], we observed that there is no apparent linear relationship between ‘Unemp’ or ‘UnempRate_EA’ and ‘Leave’. Considering the inclusion relationship among ‘Unemp’, ‘UnempRate_EA’, and ‘RoutineOccupOrLTU’, we decided to omit the first two variables in our model-building process.

To further support this decision, we analysed the correlation between each type of employment. The table shows a high correlation between all of them, with coefficients greater than 0.7. Notably, the correlations between ‘RoutineOccupOrLTU’ and the other three variables are particularly high, with coefficients all greater than 0.8.

Finally, we categorised the occupations into two groups: ‘RoutineOccupOrLTU’ and ‘HigherOccup’.

Model building and checking

Step 1: select important covariates for the generalised linear models.

Due to the unclear linear relationship and potential confounding effects of related covariates, we made the decision to retain the following variables in their original form for Model1: 'AdultMeanAge', 'Postals', 'Density', 'SocialRent', 'PrivateRent', 'NoQuals', 'L1Quals', 'L4Quals_plus', 'HigherOccup', and 'RoutineOccupOrLTU'.

By analysing the data and using PCA, we decided to replace ethnicity-related and social grade-related variables with PCA values:'EthnicityPC1' and 'EthnicityPC2' and 'SocialGrade’. And replaced the variables 'Owned' and 'OwnedOutright' with 'Owning'.

However, based on the variance of the Pearson Residuals (74.09673) of Model 1, indicates that the binomial model reveals an obvious overdispersion. Then we used quasibinomial instead for Model 2.

The p-values for ‘EthnicityPC1’ and ‘EthnicityPC2’ in Model 2 are large. However, according to Martin Rosenbaum (2017) and his report on the referendum, Ethnicity plays a crucial role in some areas, and a combination of ethnicity with other variables accounts for the significant variation in votes. Therefore, we decided to retain two PCs and search for further interactions.

The summary table for Model2 shows that there is no statistically significant distinction between the 'Deprived' and 'Multidepriv' variables. However, due to their high correlation and to simplify the model, the 'Deprived' variable was excluded while taking the inclusion relationship into account.

After removing the 'Deprived' variable in our Model 3, we observed a large p-value for 'PrivateRent', indicating less statistical significance with 'Leave'. Consequently, we considered excluding 'PrivateRent' from future models. However, based on the plot, we found that 'PrivateRent' may still have an impact on 'Leave', as the proportion of leave decreases visibly as the percentage of private renting increases. Therefore, we decided to keep 'PrivateRent' in our analysis and looked for additional interactions.

Step2: Add suitable interaction terms

According to Martin Rosenbaum (2017), the relationship between 'Leave' and 'Ethnicity' may be more complex than other covariates and dependent on various factors such as the ward's different regions. To investigate this, we included interaction terms 'EthnicityPC1' * 'RegionGroup' and 'EthnicityPC2' * 'RegionGroup' in our analysis. The new model summary showed a small p-value for both 'EthnicityPC1' and 'EthnicityPC2', indicating that including these two covariates with interactions in Model 4 is justified.

In addition, Martin (2017) also mentioned that deprived, predominantly white housing estates located towards the urban periphery tended to vote for Leave in the referendum across multiple urban areas. Hence, we considered the interaction ‘RegionGroup’ with ‘MultiDepriv’.

Considering the high p-value for ‘HigherOccup’, we found that it became statistically significant when adding an interaction with ‘L4Quals_plus ‘, which indicates there could be some relationship between people who are highly educated and who obtained a higher occupation.

Using the F test comparing models when adding each interaction, the small p-values associated with the deviance in each ANOVA table support all the interactions to be added to Model 4. Also, the small p-values in the summary table of model4 are evidence to keep ‘HigherOccup’ and those interactions.

- ‘Ethnicity’ and ‘RegionGroup’

  Figure4 illustrates the variation in ethnic composition across different regions. The clustering of colours in certain areas of the plot indicates that there are distinct groups of regions with similar ethnic compositions. This finding suggests that the rate of increase in leaving votes proportion differs across different region groups.

- ‘RegionGroup’ and ‘MultiDepriv’

  In Figure5, different colours in the plot indicate varying strengths of the relationship across different regions. This provides evidence that the rates of increasing the leaving votes proportion when ‘MultiDepriv’ increases are different in different region groups, supporting the claim.

- ‘HigherOccup’ and ‘L4Quals_plus’

Figure6 indicates that wards with a higher percentage of residents who have both higher occupation and higher education tend to have a lower proportion of leaving votes.

Comparing the fit of all models:

Model 4's performance appears to be satisfactory when compared with all the other models with a lower dispersion parameter and lower residual deviance.

There are no noticeable changes in model 2 from model 1 since we only change the response type. But in Model 2 the error terms were adjusted accordingly.

Considering overfitting and the meaning of each variable, we reduced variables in Model 3 compared with Model 2. The overall fit of Model 3 is slightly worse regarding the residual deviance (60819>58222) and dispersion parameter (77.36>74.10).

When comparing Model3 and Model4, we use ANOVA. The Deviance column shows the difference in deviance between two models, which is positive (11231). The Pr(>F) column indicates the small p-value associated with the F-statistics. The results suggest that Model 4 is a significantly better fit for the data than Model 3.

We decided to use Model 4, as it improved the fitting and reduced the dispersion situation by adding interactions between variables based on Model 3. Both the dispersion parameter (63.94) and the residual deviance (49587) are smaller compared with previous models.

By looking at the fitted plots [Figure7], it is evident that Model4 performs better than Model2 and Model3. Specifically, Model2 and Model3 show less fitting in the [0.6,0.8] range.

Conclusion:

In the final model we found that the factors that significantly affect the proportion of Leave are age, postal, social grade, house renting, and education.

‘AdultMeanAge’: Wards with higher adult mean age would have a higher proportion of leave votes.

‘Postals’: Postal voters slightly favour remaining in the EU compared to in-person voters.

‘HouseRenting’: The final model indicates that the individuals who own their accommodation are more likely to vote in favour of leaving the EU, compared to those who rent their homes.

‘Education’: The analysis suggests that the individuals with lower levels of education tend to have a higher inclination towards leaving. There may be evidence suggesting that residents with higher education tend to support remaining in the EU, which is opposite to the trend observed in lower educated individuals.

As interpreted in Step 2 of the "Model Building" part, the interactions between variables have combined impact on the proportion of votes for leaving.

When multicollinearity is present, it can be challenging to isolate the individual effects of each variable on the outcome of the model. This is because all the variables in the model are interconnected, and the impact of one variable on the outcome is dependent on the other variables in the model. As a result, it becomes difficult to discern the unique influence of each variable on the overall model outcome.

Limitation:

Data:

While the dataset may have included the most relevant factors, it is still possible that important variables were not included, and the dataset was oversimplified. For example,

factors such as the proportion of unemployed individuals in a ward, the covariates related to age groups, or other relevant demographic factors may have been left out of the analysis.

While building our model, we mistakenly treated ‘RegionGroup' as a numerical variable. Compromising the model's interpretability and potentially decreasing its accuracy. However, this approach enabled us to consider more complex interactions between 'RegionGroup' and other response variables, resulting in an improved model fit. Despite recognizing that treating ‘RegionGroup' as a categorical variable would have preserved interpretability, time constraints prevented further exploration of this option. Nevertheless, we believe that our approach captured important relationships in the data and yielded valuable insights.

Model:

When focusing on the fitted plot [Figure7], we find that there are several outliers, which may cause unreliable results since the outliers can significantly affect the distribution of the data and distort the model's predictions.

And also, the outliers might be considered noise in the data which may reduce the accuracy of the model. Since these values account for less than 1% of the observations, they are not a significant issue.

Although the error term is approximately normally distributed, the QQ-plot [Figure8] shows slightly heavier tails than what would be expected under the assumption of normality for the final model.

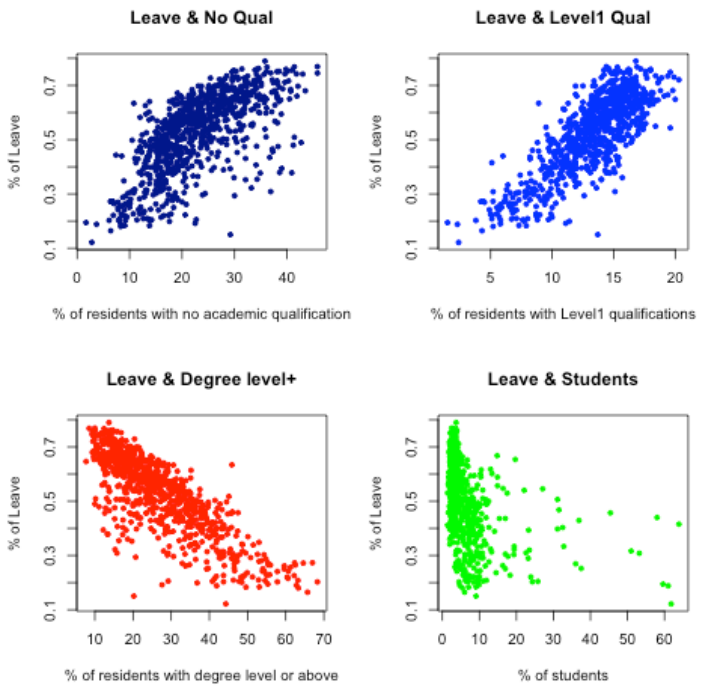

Figure 1. Scatter plots between Education and Leave

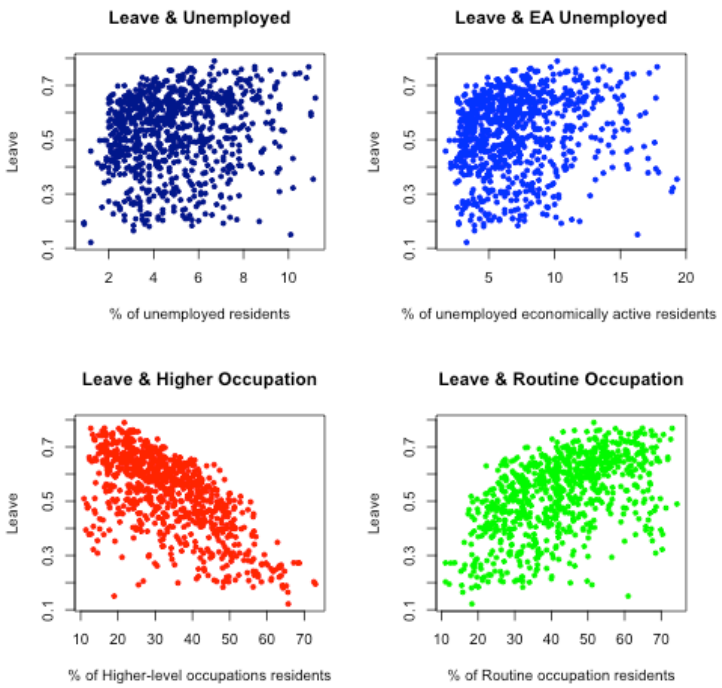

Figure 3. Scatter plots between Occupation and Leave

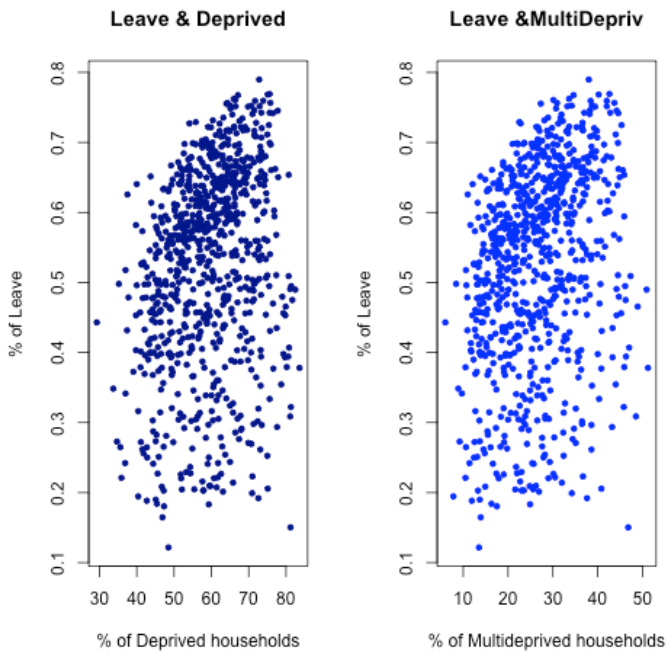

Figure 2. scatter plots between 'Deprived'/ 'MultiDepriv' and 'Leave'

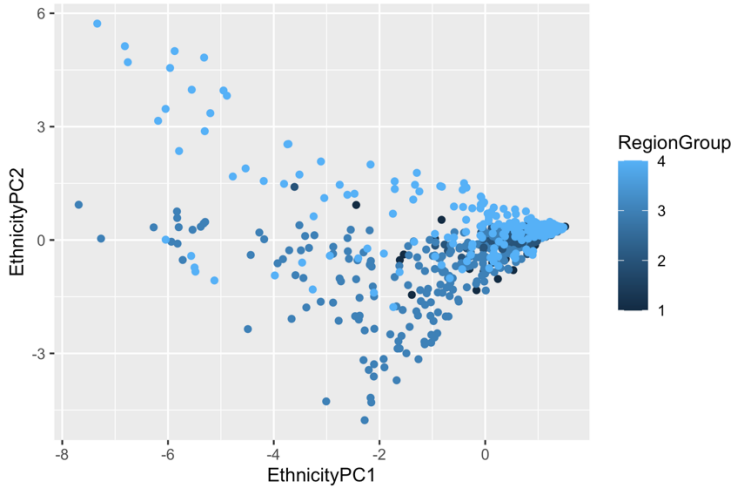

Figure 4. Scatter plots between Ethnicity PCs among RegionGroup

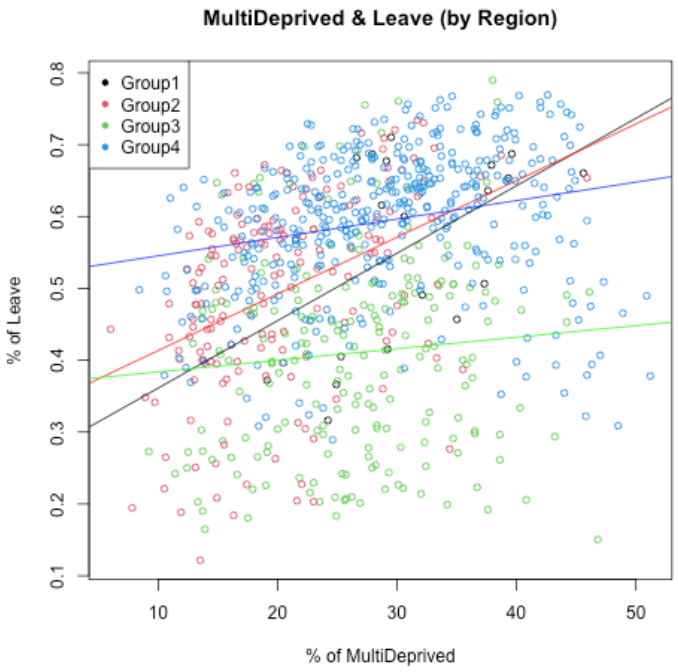

Figure 5. Multideprivation against Leave among region groups

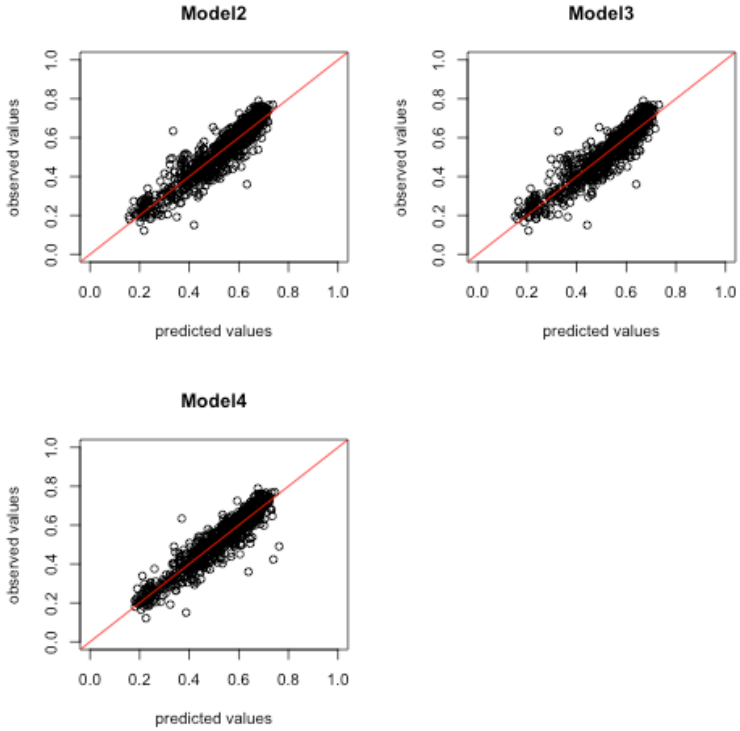

Figure 7. fitted plots for Model2 (top left), Model3(top right), Model4(bottom left)

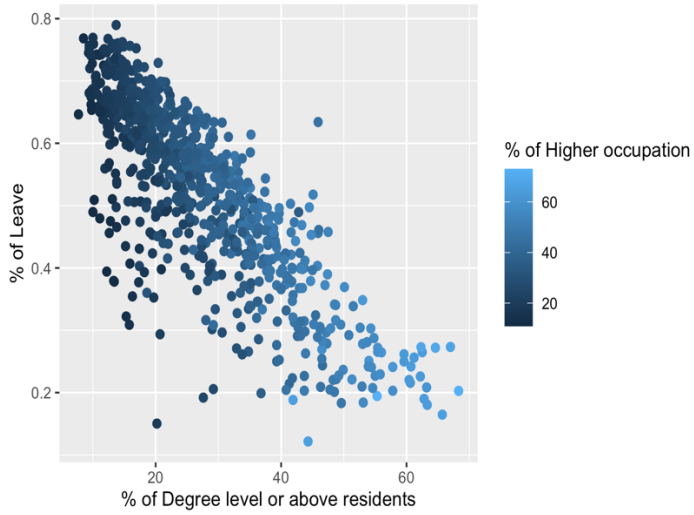

Figure6. scatter plots of 'L4Quals_plus' against %leave votes among 'HigherOccup'

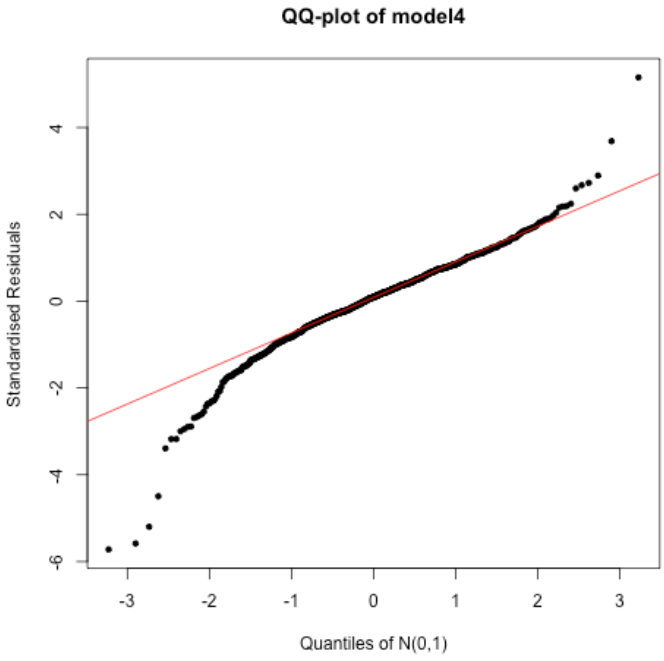

Figure 8. QQ-plot for Model4

Each member contributed equally.