University of St Andrews | FOUNDED 1413

# ID5059 Practical2 Individual Report

Student ID: 180009169
Student Subject: Msc Applied Statistics and Datamining

## 1. Executive Summary

This report aims to provide analysis and models of relationships between Party, Remain percentage and Brexit motions in the first round of Brexit Indicative Votes. Methods applied in this report include Principal Components Analysis (PCA), Logistic Regression and K-Nearest Neighbour, while introduce other classification models briefly. The report finds Labour Alternatives motion might be useful to distinguish Conservative, Labour, and Scottish National Party. Also, Conservative and Democratic Unionist Party are likely to have similar political views about Brexit voting, while Labour, Liberal Democrat and Scottish National Party are on the other side. Moreover, MPs with high remain percentages are likely to against Confirmatory public motion. As for models, both logistic model and k-nearest neighbour have high accuracy of predicting parties. The analysis conducted in this report also have some imitations. For instance, it does not have cross validation and lack visualisation (i.e. ROC curve) when assessing the accuracy of models.

## 2. Introduction

### 2a) Background

The first round of Brexit Indicative Votes held on the 27th March 2019. According to the definition from Institute for Government (2019), “indicative votes are votes by MPs on a series of non-binding resolutions. They are a means of testing the will of the House of Commons on different options relating to one issue.” Although no motion got more than fifty percent aye, these votes show the tendency and help to narrow down options in the future rounds.

### 2b) Data Description

This report is based on RStudio (RStudio Team, 2015) and Brexit indicative votes data which are from the guardian webpage (Holder, Voce, & Clarke, 2019). The original dataset has 13 columns and 623 rows. The columns include member name, party, constituency, remain percentage, stance and eight Brexit options,

while each row is a MP. In addition, the values for eight Brexit options are assigned as 1, 0 and -1. Positive one means “for”, negative one means “against”, and zero means “abstain”.

The original dataset might not clean enough mainly for three issues. The first one is that some MPs obtained throughout, these data would just be noise, have no contribution to and even distort our models. The other thing is some parties have too few members. It can make predictions for these parties be difficult because the predictions can be arbitrary when data size is tiny. The last issue is that there are three missing values in the remain percentage column. To fix the first two problems, it might be reasonable to omit these data points in the beginning. However, for missing percentages, it may be better to impute them with the mean rather than just removing them since our dataset is already quite small. After cleaning data, we have 597 rows and 13 columns now. In addition, although we have 13 variables, member name and constituency variables seem to be meaningless to our predictions. Therefore, we just ignore them.

## 3. Exploratory Data Analysis

Before fitting models, we should get familiar with our data. Therefore, we did some preliminary data analysis. In this section, some basic description and principal components analysis are performed.

### 3a) Statistical Description

First, we simply plot bar charts (Figure1) to give us an overview of the for/against percentage for each Brexit option and trends of different parties. In Figure1, each bar chart is plotted for one Brexit option labelled by one letter. B stands for “No Deal”, D stands for “Common Market”, J stands for “Customs Union”, H stands for “EEA/Efta without customs union”, K stands for “Labour Alternatives”, L stands for “Revoke Article 50”, M stands for “Confirmatory public vote” and O is “Contingent Plan”. The x-axis values are against(-1), abstain(0) and for(1), while the y-axis values are the counts of them.

Intuitively, the votes are fairly separated among parties in all alternative Brexit options except Labour Alternatives (option K). Most Conservative MPs against it and nearly all Labour MPs voted for, while all Scottish National MPs abstained. This means Labour Alternatives (option K) may be the important variable to predict parties.

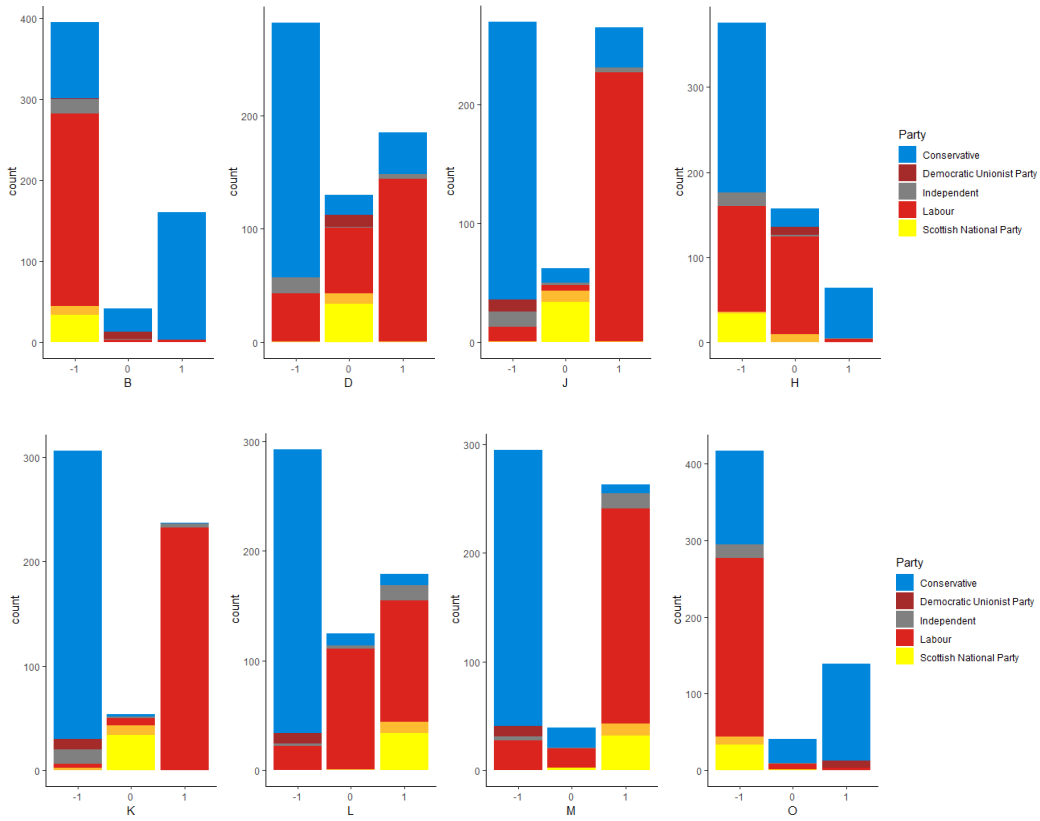



**Figure1. Overview of votes for eight alternative Brexit options** The blue bar is Conservative Party, tends to against option D, J, K, L and M. The red bar is Labour Party, tends to against option B and O, while for option J and M. The yellow, dark red and grey bar denote the votes of Scottish National Party, Democratic Unionist Party and Independent Party respectively.

After ploting barcharts, we plotted pairwise correlation coefficients between the remain percentage and eight options (Figure2). The correlation coefficient shows how linear dependent two variables are. The range of correlation is between negative one and positive one. The signs indicate positive or negative relationships. The absolute value shows how close the relationship. For examples, negative one means the two variables are perfectly negative linear, positive one means the two variables are perfectly positive linear, while zero shows they have no linear relationship.

Figure2 shows that, except for remain percentage and option H, most variables have linear relationships with others more or less. The results for remain percntage might be concerned since it indicates that other variables may not be able to explain remain percentage well. However, correlation coefficients only show linear relationships. Therefore, it is still possible for non-linear relationships.

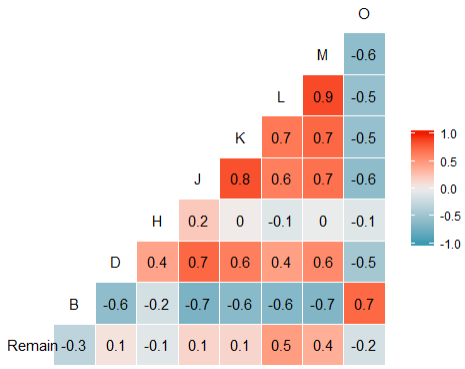



**Figure2. Pairwise Correlation Coefficients between Remain Percentage and Brexit Options** Red squares mean positive correlation coefficients. Bule squares mean negative correlation coefficients. Darker colours means higher absolute correlations corfficients.

### 3b) Principal Components Analysis

It is hard to identify patterns in our dataset when it has many variables. For instance, we have to look through all the plots, bars, colours in Figure1 to find whether any structure there. Moreover, it is easy for our eyes to miss some unnoticeable patterns. Therefore, we performed Principal Components Analysis (PCA) in this report. PCA is a technique for multivariate data analysis. The main aim of it is to reduce data dimensions while retaining as much information as possible. PCA combines original variables into new variables which called Principal Components and order them by importance. Then PCA projects original data points in new axes so we could see if there is any cluster. The projection matrix is called loadings and the new transformed observations are called scores.

Remain percentages are scaled before doing PCA since it has different range with other numeric variables. The first task of PCA is to choose the number of principal components. According to the cumulative proportion (Appendix A), the first four components could retain nearly 90% information, which is relatively high. Moreover, we plotted Scree graph (Appendix B) and calculated Kaiser's criterion (Appendix C). The Scree plot suggests us choose the first five components, but Kaiser's criterion suggests only the first two. The Scree plot usually keep too much components, while Kaiser's criterion tends to retain too few. After trading off, we made the cut off points as four. The Table1 displays the loadings, which can be considered as weights of original variables. The signs indicate the direction original variables contribute to principal components. Blanks in the table are not missing values but values which are too small to present.

**Table1. Loadings for the first four Principal Components**

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Remain | 0.156 | 0.481 | 0.741 | 0.120 |
| B | -0.380 | | -0.106 | 0.433 |
| D | 0.337 | -0.389 | | 0.274 |
| H | | -0.668 | 0.552 | |
| J | 0.398 | -0.171 | -0.150 | 0.188 |
| K | 0.383 | | -0.316 | 0.282 |
| L | 0.373 | 0.317 | | 0.131 |
| M | 0.395 | 0.198 | | 0.179 |
| O | -0.336 | | | 0.747 |

According to the left plot in Figure3, there are clear clusters among parties and nearly vertical boundary in the middle of plot. The distance between parties indicates the similarity of them. Parties which gathered are more likely to vote similarly. Therefore, Conservative and Democratic Unionist Party might have similar political views about Brexit voting, but Labour, Liberal Democrat and Scottish National Party are on the other side, while Independent Party is near the boundary. Combined the plot with the loadings table to analyse, all options except B (No Deal) and O (Contingent Plan) have positive weights for PC1. Therefore, Conservative and Democratic Unionist Party have small values in PC1 might because many MPs from the two parties vote for motion B and O. On the contrary, most MPs of other parties against the two motions.

The right plot in Figure3 are coloured by remain percentages. We classified percentages with two levels, “low” if the percentage is under the mean and “high” if the number is above the mean. As the plot shows, points with high remain percentages have higher PC2 values than points with low percentages in general. According to loadings, motion M (Confirmatory public vote) dominates PC2 since it has much larger absolute weight than other variables. Furthermore, because the weight is negative, the high remain percentage might indicate voting against to option M.

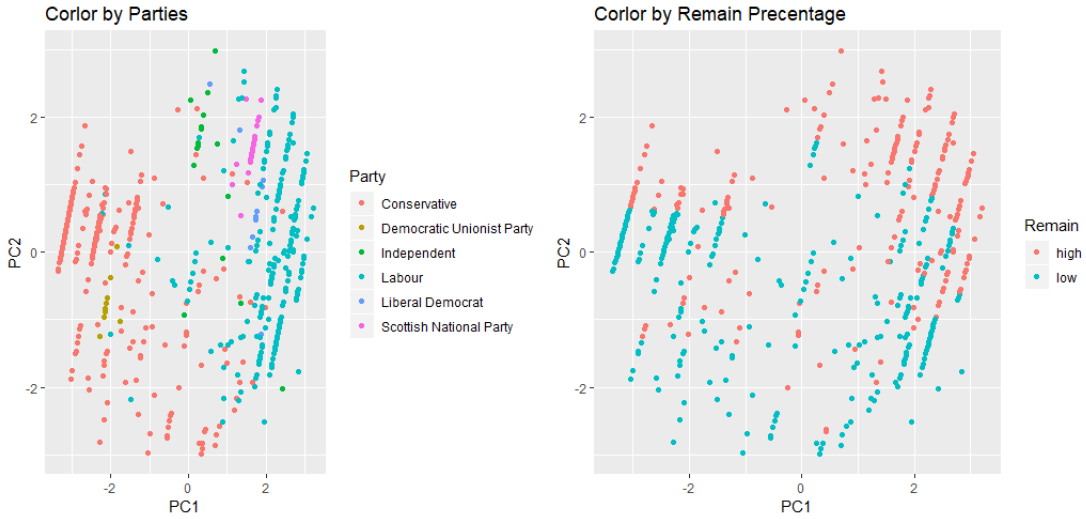



**Figure3. Scores plotted on PC1 and PC2 with different parties and remain percentages** The points in left graph are coloured by parties. The points in right graph are coloured by remain percentages that classified with "low" and "high". There are clusters along PC1 in the left plot and clusters along PC2 in the right plot.

Similarly, we plotted coloured scores on PC3 and PC4 as Figure4. It is unlikely to observe clear structure in the left plot. The remain percentage contribute most in the PC3. Therefore, showing no structure in the left plot could be the evidence that the remain percentage has weak relationship with party. Although the right graph gives us pattern along PC3, it is meaningless because it just indicates points in the high remain percentage group tend to have larger values for the remain percentage, which is absolute but useless.

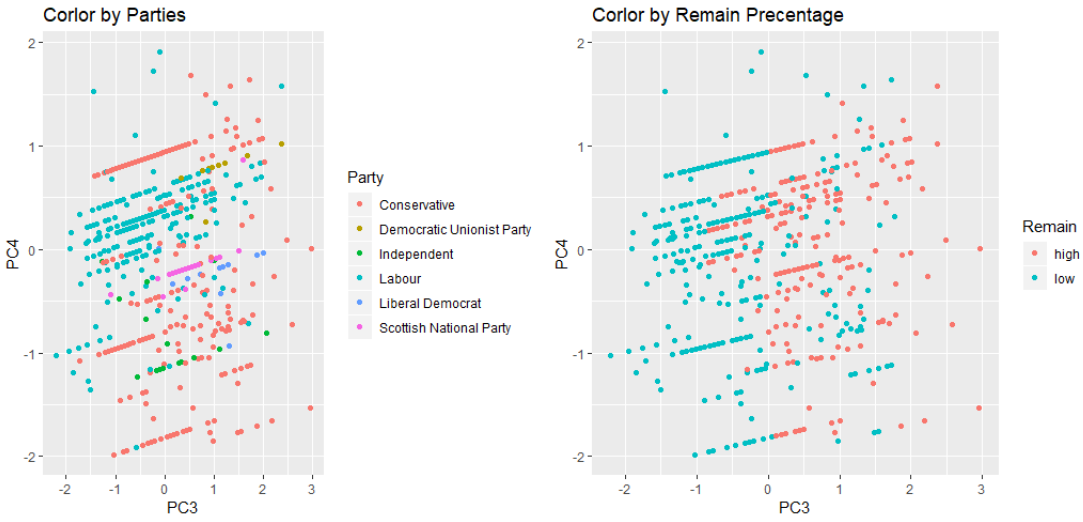



**Figure4. Scores plotted on PC3 and PC4 with different parties and remain percentages** The points in left graph are coloured by parties. The points in right graph are coloured by remain percentages that classified with "low" and "high". It seems no cluster in the left plot. There are some clusters along PC3 in the right plot.

## 4. Methods

Seven classification methods are conducted to predict Party, which are multinomial logistic regression, weighted k-nearest neighbour, decision tree, random forest, neural net, naïve bayes and support vector machine. For the second question, linear regression is applied. This section would stress logistic regression and k-nearest neighbour, while gives some basic understanding of others.

### 4a) Logistic Regression

Logistic Regression is an algorithm which models the probability of the response variable rather than models the response directly (James, et al., 2013). It introduces a link function: $p(X) = \frac{e^{\eta}}{1+e^{\eta}}$ , where η is a linear equation: $\eta = \beta_0 + \beta_1 X$. Therefore, maximum likelihood can be applied to fit the model above. Introducing this link function to logistic model guarantees the estimated probability always larger than zero and under one, no matter which value $\eta$ is. Otherwise, the estimated probability might be negative or beyond one, both situations make no sense in the real world. Therefore, this property is important and the key point for the logistic model. Another advantage of logistic regression is that we can calculate p-value

for each covariate to judge whether it is significant. However, logistic models can be difficult to interpret and visualise.

Normally, logistic regression is designed for binary classification. However, in this report, our response variable Party has more than two classes, so we performed multinomial logistic regression. It has similar algorithm as normal logistics regression but extended for multiple classes.

### 4b) K-nearest Neighbour

The algorithm of K-nearest Neighbour is easy to understand. It first identifies K points in the training data that are closest to the point in the test data that we want to predict, then finds the most common value among the K points, finally assigns this value to the predicted point. The number K can be decided by users. In the report, we used an improved version for this method by weighting every training point with its distance to the predicted point, which makes the method more reasonable. The main advantage of this method is it do not require any statistical assumption such as normality or independency. However, k-nearest neighbour is not able to tell us the significance of explanatory variables. It is also impossible to get fitted values for training data.

### 4c) Other methods

Decision Tree is a widely used model that have a series of nodes, branches and leaves. Each node denotes a test, each branch denotes an outcome of the test, and each leaf denotes a class. Random Forest is a method that contains lots of decision trees. Because of it, random forest avoids the overfitting issue that decision tree would face. Naïve Bayes method is based on conditional probability and has strong independency assumption. Support Vector Machine aims to find a hyperplane in N (the number of features) dimensions space to classify data points.

## 5. Model Results

### 5a) Multinomial Logistic Regression

The first logistic regression includes Remain percentage and all eight motions as explanatory variables. The misclassification rate for training data is about 3.77%. It means the model has 3.77% chance to make a wrong prediction. However, when we apply the model to the test data, the rate increases to 7.5%. This is

a sign that the model might be overfitting. After checking the p-values (Appendix D) for each variable, we found all p-values of Remain percentage are much higher than 0.1. It is a strong evidence that Remain percentage is not significant to all parties. Therefore, we removed the variable when we built the second logistic model. The misclassification rate for the second model is similar when applying to training data. However, it decreases to 6.67% when testing with test data, which is an improvement compared with previous model. In addition, the p-values (Appendix E) show that all explanatory variables are significant to at least one party now. Table2 shows the confusion matrix for logistic regression. The figure in the diagonal is the number of correct predictions. It is obvious that most predictions are right.

**Table2. Confusion Matrix for Logistic Regression**

| Reference / Prediction | Conservative | Democratic Unionist Party | Independent | Labour | Liberal Democrat | Scottish National Party |
|---|---|---|---|---|---|---|
| Conservative | 52 | 0 | 1 | 1 | 1 | 0 |
| Democratic Unionist Party | 1 | 5 | 0 | 1 | 0 | 0 |
| Independent | 0 | 0 | 3 | 1 | 0 | 0 |
| Labour | 1 | 0 | 0 | 43 | 0 | 1 |
| Liberal Democrat | 0 | 0 | 0 | 0 | 0 | 0 |
| Scottish National Party | 0 | 0 | 0 | 0 | 0 | 9 |

### 5b) Weighted K-nearest Neighbour

K-nearest Neighbour cannot be used to predict the training data because of its property. Therefore, we only apply it to test data. The max value of K is set as 9, which means RStudio would build models from K=1 to K=9 and give the best model. The results show the model with K=5 is the optimal one with 94.17% accuracy. According to the confusion matrix for k-nearest neighbour (Table3), the prediction rarely get wrong, which proves the great performance of this method.

**Table3. Confusion Matrix for K-nearest Neighbour**

| Reference / Prediction | Conservative | Democratic Unionist Party | Independent | Labour | Liberal Democrat | Scottish National Party |
|---|---|---|---|---|---|---|
| Conservative | 52 | 0 | 1 | 2 | 0 | 0 |
| Democratic Unionist Party | 0 | 5 | 0 | 0 | 0 | 0 |
| Independent | 0 | 0 | 3 | 1 | 0 | 0 |
| Labour | 0 | 0 | 0 | 43 | 0 | 0 |
| Liberal Democrat | 1 | 0 | 0 | 0 | 0 | 0 |
| Scottish National Party | 1 | 0 | 0 | 0 | 1 | 10 |

## 6. Conclusion

To sum up, there are main conclusions. The first is that Conservative and Democratic Unionist Party might have similar political views about Brexit voting, but Labour, Liberal Democrat and Scottish National Party are on the other side. The second is that most aye of motion B and O are from Conservative and Democratic Unionist Party. On the contrary, most MPs of other parties against the two motions. The third one is that the high remain percentage might indicate voting against to option M. The fourth is that remain percentage is not statistically significant for the party prediction. The last finding is that logistic regression and k-nearest neighbour method both can predict parties well. In addition, the accuracy rates of them are both over 90%.

Reference:

Holder, J., Voce, A. and Clarke, S., (2019). How did your MP vote in the indicative votes? Available at: https://www.theguardian.com/uk-news/ng-interactive/2019/mar/27/how-did-your-mp-vote-in-the-indicative-votes (Accessed: 16th April 2019)

Institute for Government (2019). Indicative votes. Available at: https://www.instituteforgovernment.org.uk/explainers/indicative-votes (Accessed: 19th April 2019)

James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

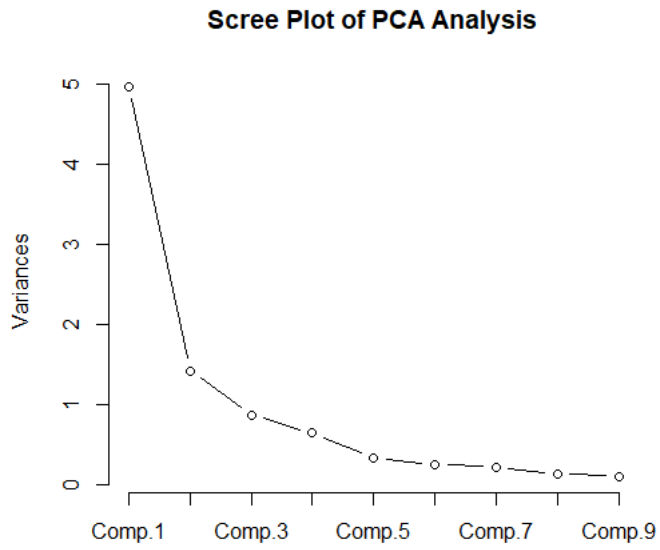
RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL: http://www.rstudio.com/.

## Appendix A. Importance of Components

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|---|---|---|---|---|---|---|
| Standard deviation | 2.2302062 | 1.1945713 | 0.93700303 | 0.80517016 | 0.58628161 | 0.50971197 |
| Proportion of Variance | 0.5526466 | 0.1585556 | 0.09755274 | 0.07203322 | 0.03819179 | 0.02886737 |
| Cumulative Proportion | 0.5526466 | 0.7112022 | 0.80875498 | 0.88078820 | 0.91897999 | 0.94784736 |

| | Comp.7 | Comp.8 | Comp.9 |
|---|---|---|---|
| Standard deviation | 0.47638379 | 0.37158833 | 0.32303929 |
| Proportion of Variance | 0.02521572 | 0.01534199 | 0.01159493 |
| Cumulative Proportion | 0.97306308 | 0.98840507 | 1.00000000 |

## Appendix B. Scree Plot for PCA




## Appendix C. Kaiser's Criterion

| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|
| [1]4.9738195 | 1.4270006 | 0.8779747 | 0.6482990 | 0.3437261 | 0.2598063 | 0.2269415 | 0.1380779 | 0.1043544 |

## Appendix D. P-values for the First Logistic Model

| | (Intercept) | Remain | B | D | H | J | K |
|---|---|---|---|---|---|---|---|
| Democratic Unionist Party | 0.0000000 | 0.2586537 | 0.00000000 | 0.00000000 | 0.0000000 | 3.684355e-07 | 3.205346e-05 |
| Independent | 0.5128667 | 0.1611610 | 0.27845167 | 0.20093314 | 0.3071453 | 2.705547e-01 | 4.127736e-03 |
| Labour | 0.7794854 | 0.9244404 | 0.08801971 | 0.01165011 | 0.1267093 | 5.519363e-01 | 4.821384e-08 |
| Liberal Democrat | 0.0000000 | 0.3611077 | 0.00000000 | 0.93069718 | 0.0724330 | 4.795059e-02 | 2.075843e-02 |

```
Scottish National Party     0.0000000 0.8445756 0.00000000 0.00000000 0.0000000 0.000000e+00
1.336467e-04
                                    L           M           O
Democratic Unionist Party 0.00000000 1.663559e-03 0.000000000
Independent               0.06241858 2.068173e-01 0.759670702
Labour                    0.07265776 4.226939e-05 0.006218096
Liberal Democrat          0.85824797 0.000000e+00 0.000000000
Scottish National Party   0.00000000 1.144141e-01 0.000000000
```

## Appendix E. P-values for the First Logistic Model

```
 (Intercept)          B           D          H           J           K           L
 Democratic Unionist Party 0.002723866 0.82502448 0.169598666 0.34887278 0.97818667 8.409628
 Independent               0.256564227 0.34321238 0.299611891 0.33733894 0.23988953 2.286909
 Labour                    0.713038798 0.09988913 0.009275429 0.05092536 0.71760002 4.191321
 Liberal Democrat          0.000000000 0.00000000 0.856030429 0.13528782 0.08491095 2.157286
 Scottish National Party   0.000000000 0.00000000 0.000000000 0.00000000 0.00000000 2.671109
                                       M           O
 Democratic Unionist Party 6.241258e-01 0.023783682
 Independent               1.085447e-01 0.597573551
 Labour                    1.378218e-06 0.002175917
 Liberal Democrat          0.000000e+00 0.000000000
 Scottish National Party   8.793598e-02 0.000000000
```