ICA 2 STAT0023

1 Exploratory Data Analysis

The dataset *ReferendumResults.csv* contains information on the voting data for 1070 electoral wards during the 2016 EU-referendum in the United Kingdom. 267 wards have missing data and were saved separately in *testing*, resulting in a modified dataset with 803 wards. Our aim is to construct a statistical model to predict the proportion of leave votes within a ward using a selection of covariates from the dataset. There are 49 variables provided: proportion of Leave votes each ward *proportion_leave* is calculated using $NVotes$ and *Leave*, while the other 47 can be allocated into 8 groups by definition: Age Group, Education, Occupation, Housing Status, Social Status, Geographic Characteristics, Ethnicity and Miscellaneous. The proportion of leave votes within wards ranges from 0.12 to 0.79 with a mean of 0.52. The dataset generally exhibits strong correlations among variables within their respective groups, as well as across different groups.

1.1 Age

The Age group encompasses $MeanAge$, $AdultMeanAge$ and proportion of permanent residents between 0 and 90 years old per ward. $AdultMeanAge$ appears to be the most representative in this group because it has a strong linear relationship with *proportion_leave* at r=0.48, compared to $MeanAge$ at r=0.37. The relationship between $AdultMeanAge$ and *proportion_leave* can be observed from Figure 1. $AdultMeanAge$ is also a natural choice as under-aged residents cannot vote.

The age bracket variables also tend to have a strong negative correlation with each other. For example, Age_20to24 and Age_65to74 with r = -0.52. This may be because the age groups are complementary: higher proportion of the elderly per ward implies a smaller proportion of young people.

1.2 Education and Occupation

The Education group consists of $NoQuals$, $L1Quals$, and $L4Quals\_plus$. $L4Quals\_plus$ expressed the strongest relationship with 'proportion_leave', showing a clear negative linear trend (r=-0.81). Which implies that wards with a lower proportion of educated residents tend to have a higher proportion of leave votes.

$Students$, $Unemp$, $UempRate\_EA$, $HigherOccup$ and $RoutineOccupOrLTU$ are variables in the Occupation group that describe the ward population categorised by occupations. Scatter-plot of $HigherOccup$ against $proportion\_leave$ (Figure 2) shows that wards with a higher proportion of residents in higher occupations tend to vote less leave votes (r=-0.65).

However, as covariates in both Education and Occupancy measure similar indicators of socioeconomic status, they are highly correlated with each other. For example, wards with more highly-educated residents have fewer less-educated residents due to complementary effects. Similarly, the same wards also tend to have more people in higher occupations compared to routined occupations, as shown in the correlation matrix for $NoQuals$, $L4Quals\_plus$, $HigherOccup$ and $RoutineOccupOrLTU$.

1.3 Housing Status and Social Status

The Housing Status group contains $Owned$, $OwnedOutright$, $SocialRent$ and $PrivateRent$, which are the percentage of permanent residents in a ward with the respective ownership and renting status. $PrivateRent$ shows the strongest linear relationship with $proportion\_leave$ (r=-0.59), but it also strongly correlates with $Owned$ and $OwnedOutright$. $SocialRent$ has a small association with $PrivarteRent$ and a weak relationship with $proportion\_leave$ with correlation coefficients close to zero.

A similar pattern can be seen in the Social Status Group, which comprises $Deprived$, $MultiDepriv$, $C1C2DE$, $C2DE$ and DE that explains the proportions of resident/household categorised as deprived or in a social grade. These covariates highly correlated within their group (r>0.85), which can be attributed to the inclusive relationships, namely $C1C2DE$ includes $C2DE$ and DE. They also strongly correlate with variables within the Education and Occupation group, which is contextually reasonable because residents with higher qualifications or having higher occupation tend to have a higher social grade.

When *Deprived* interacts with *L4Quals_plus* in Figure 3, it shows that the wards with a lower proportion of educated residents are more affected by *Deprived*, explained by the light-coloured points, which stand for higher deprivation rates, being shifted down.

1.4 Geographic Characteristic and Ethnicity

Geographic Characteristics includes *AreaType*, *RegionName* and *Density*. *Density* has a strong correlation with *proportion_leave* (r=-0.55). *AreaType* and *RegionName* are categorical variables with 4 and 9 levels. To reduce dimensionality for modelling, *RegionName* can be transformed into 4 levels using hierarchical clustering, which group wards with similar characteristics into a subgroup and the number of groups was defined at a reasonable threshold of 4. The 4 new levels will therefore be *EM*, *EE_SW*, *LON* and *Other*, and included in a new variable called *RegionGroup*. Figure 4 indicates that both variables have a similar relationship with *proportion_leave*, illustrating that the wards in London have a notably lower proportion of leaving votes, compared to wards located elsewhere. *RegionGroup* lists a clearer difference among its levels, suggesting a stronger relationship with *proportion_leave*.

Ethnicity group includes *White*, *Asian*, *Black*, *Indian* and *Pakistani*. *White* seems to be the most representative of the group, due to its complementary relationship with the other ethnic-minority levels. When *White* interacts with *L4Quals_plus*, Figure 5 shows that high White proportions shifts up the proportion of Leave Votes among wards with low proportion of *L4Quals_plus* more than wards with high proportion of *L4Quals_plus*.

1.5 Miscellaneous

Postals, *Residents* and *Households* are all grouped in Miscellaneous. The data for *Postal* are relatively imbalanced, with 48 wards as non-postal and 755 as postal, making it less impactful for further analysis. *Residents* and *Households* also show unclear relationships with *proportion_leave*.

2 Modelling

Model 1: Selection of covariates and interaction terms

As the aim of our model is to predict the proportion of leave votes within a ward, a natural choice is to use a generalised linear model with a logistic link function. GLM with logistic link can generate predictions bounded in [0,1] and allows for heteroscedasticity, which might be the case in socio-economic data. By weighting the observations by NVotes, more importance is given to wards with larger vote counts, as they are likely to have a larger impact on the overall Leave-vote proportion.

In Model 1, covariates were selected based on their contextual relevance and their observed relationship with *proportion_leave* during exploratory data analysis. The covariates were chosen from a set of factors that have been previously identified as potential influencers of Leave Votes, including Education, Age, Ethnicity, Geographic, Housing and Social Status, with Education being considered a particularly strong predictor (Martin 2017). We then selected the most representative variables from each group, and removed variables that either had an unclear linear relationship with *proportion_leave*, or had strong collinearity with covariates in each group, such as $SocialRent$ and $C1C2DE$. Additionally, our EDA revealed that the impact of $L4Quals\_plus$ on the response varied depending on the proportion of White ethnicity or Deprivation rate. As a result, we included interactions between $L4Quals\_plus$ and each of these covariates in Model 1. The summary output is shown in the code.

The diagnostic plots show that there is no systematic pattern in the standardised residual versus fitted values plot. Upon further examining the standardised residual versus fitted values plot in Figure 6, we observed that the standardised residuals surpass the range of [-2,2], indicating evidence of overdispersion in the data. Further analysis revealed that the Variance of the Pearson Residuals is 59.26, exceeding 1 and supporting our previous graphical observation. overdispersion means the data has higher variation than the binomial distribution, and Model 1 fails to capture this excess variation. This implies that the calculations for the standard error and hypothesis test might have been incorrect.

The Cook's Distance plot in Figure 6 also shows that there are 467 influential observations beyond the $8/(n-2k)$ threshold. This could have been resulted from the diagnosed overdispersion in the data, or the highly correlated nature of the covariates.

Model 2: Fixing overdispersion

To address the extra variability in the data, a quasi-binomial model could be more suitable. In Model 2, a "dummy" dispersion parameter is incorporated, which results in inflated standard errors and p-values. Although the value of the estimated coefficients did not change, their revised p-values provide more evidence for some of them to be removed from the model.

Investigations on the Cook's Distance values shows that the number of influential observations dropped to 25, which might result from the change in distribution chosen.

Model 3: Evaluating covariates under quasi-binomial model

To obtain significant predictors for our model, we iteratively removed insignificant covariates based on t-tests from the summary output in Model 2. The covariates *Owned*, *OwnedOutright*, *HigherOccup*, *PrivateRent* were deleted one by one, with each removal supported by an F-test that showed no evidence against the null hypothesis that nested models with one less predictor explains the same information as the full model. Contextual reasons were also accounted for in this removal, as these covariates are not comparably meaningful relative to covariates in other groups. Therefore we decided to omit these covariates.

The residual deviance has increased from 46552 with 781 degrees of freedom in Model 2 to 46773 with 784 degrees of freedom in Model 3. Nonetheless, the proportion of deviance explained in both Model 2 (88.1%) and Model 3 (88.0%) is quite similar. This implies that the removal of four covariates from Model 2 only led to a minimal decrease in residual variance, thereby providing a justifiable reason for opting for the lower dimensionality of Model 3.

Again, no systematic patterns are seen in the diagnostic plots. Cook's Distance values shows that the number of influential observations slightly increased to 27, which might be due to the change in number of covariates.

3 Prediction

Figure 7-8-9 indicate that all models have a similar level of fit on the training data. This could be due to the fact that all models capture the same underlying effects in the dataset. The fitted values for

most wards appear to be similar to observed values, as evidenced by the large number of points falling on the $y = x$ line. However, good performance on the training data does not necessarily guarantee good performance on the testing data. Over-fitting is a potential issue that needs to be considered, and thus Model 3 is preferred due to its lower dimensionality. This reduces the risk of overfitting and improves the generalisability of the model. As investigated from the Modelling section, it can be seen that the standard errors were inflated in Model 2, and remain roughly the same in Model 3.

4 Limitation

Due to the socio-economic nature of the dataset, the presence of high collinearity among covariates make the estimated coefficients are notably sensitive to covariate selection. The approach in this report prioritises the interpretations of each covariates, whereby the most contextually relevant covariates such as in the Education group or Age group are selected, then we progress onto evaluating the model. The disadvantage of doing this is that it is difficult to argue that this is the "best possible" model that can be built for predicting the Leave-Vote proportion.

An alternative approach to find the "best possible" combination could involve combinatorial optimisation, namely by going through all possible combinations of covariates and selecting one with the best metric, namely the smallest Residual Deviance. This method, however, is computationally intensive and potentially leaves out important covariates in each group.

The objective of any predictive model is to achieve the highest possible accuracy while keeping uncertainty at a minimum. However, it is important to note that there exists a trade-off between accuracy, which can be measured by bias, and uncertainty, which is measured by variance. During the model-building process, we took care to strike a balance between these two factors. We ensured that the variance was not too high by including only the most important covariates, and checked that the residual deviance did not increase significantly when any covariate was removed.

However, assessing the generalisability of the model is equally important, and this can be achieved through cross-validation techniques. For example, we can split the dataset into two parts - a training set (90%) and a testing set (10%). If the model performs consistently well on both datasets, it is an indication that the model is not overfitting to the training set, and further improvements can be made

to reduce bias and variance. Metrics such as the proper scoring rule can also be more appropriately used here to compare performances between different models, which was limited in the current approach where R^2 and AIC are not available and proper scoring rule is not appropriate to be used on the training data.

5 Conclusion

Overall, Model 3 supports our findings in the exploratory data analysis suggesting that geographic demography, age, ethnicity, education, employment status, density, social grade and the combined effect among them has a significant role in predicting the proportion of leave votes within each electoral ward. The number of predictors were reduced significantly to 14. As we are including a categorical covariate, the reference group is set as the region East Midlands.

It should be noted that the interpretation of the magnitude and direction of each coefficient in the models is not straightforward due to three reasons. Firstly, the highly correlated nature of the covariates means that the individual effects of each covariate are intertwined with one another, making it difficult to draw meaningful conclusions from each coefficient. For instance, the initial expectation was that regions in London (as indicated by $RegionGroupLON$) would have a higher proportion of Leave-Votes compared to other areas, based on the boxplots generated in the EDA. However, the odds of voting to leave increase by a factor of $exp(\beta_{\text{RegionGroupLON}}) = 1.15$ when comparing $RegionGroupLON$ to the reference group $RegionGroupEM$, which is higher than other areas at 1.11 and 1.03, both compared to the reference group. Secondly, interpretation of the contribution of the covariate to the response is subject to the usual size of the values it takes. For example, although $Unemp$ has a larger coefficient than $L4Quals\_plus$, it should be noted that $Unemp$ has an average value of 4.90 while $L4Quals\_plus$ is 27.17 based on the data. Finally, some covariates are measured in different scales, namely $AdultMeanAge$ with years, and the other covariates are with proportions.

References

Martin, R. (2017), 'Local voting figures shed new light on EU referendum'.

URL: *https://www.bbc.co.uk/news/uk-politics-38762034*

Contribution

In this project both team members contributed equally.

Appendix: Graphs

Figure 1:

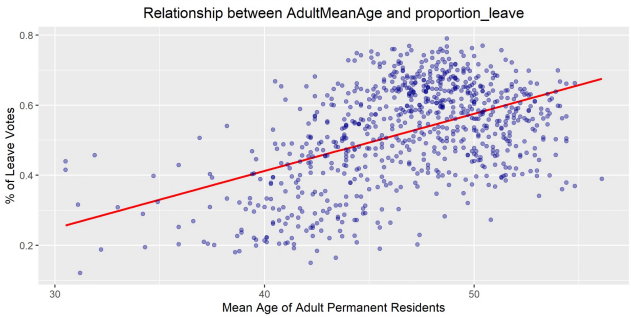

Figure 2:

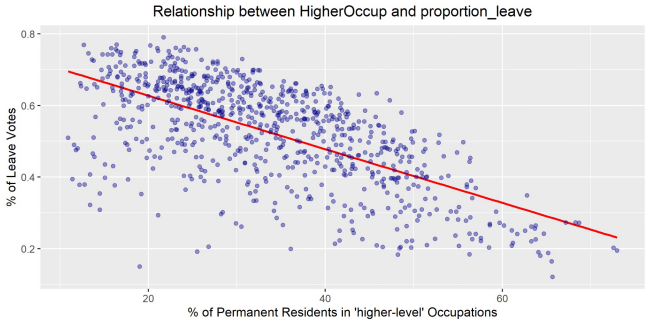

Figure 3:

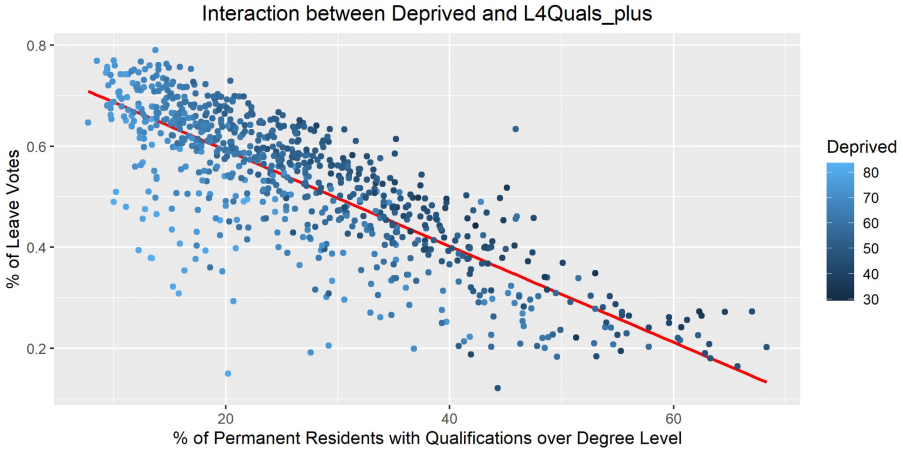

Figure 4:

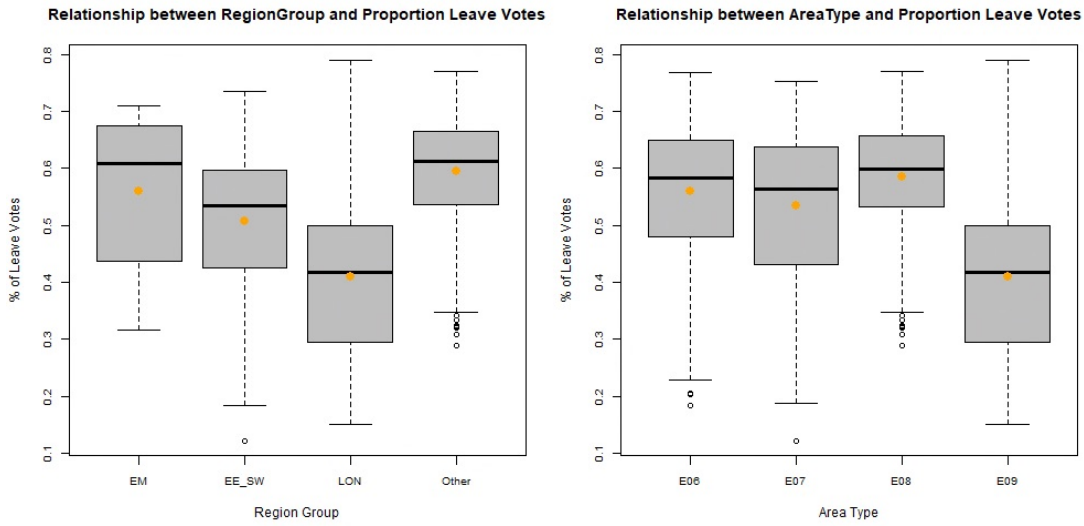

Figure 5:

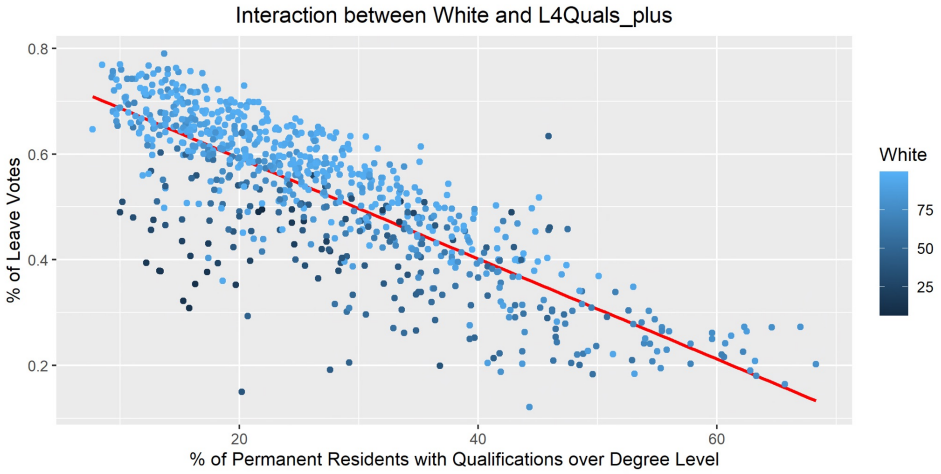

Figure 6:

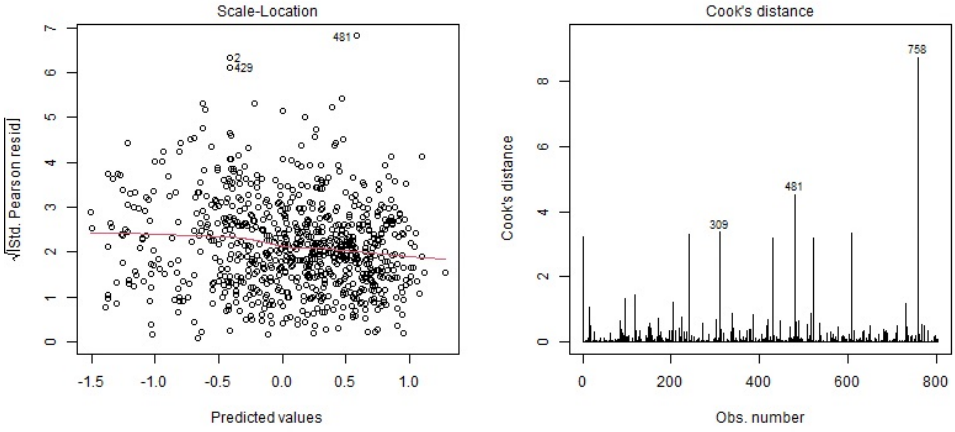

Figure 7-8-9: