Assessed Practical: Brexit

Assignment due: 12pm Friday March 29

On June 23rd, 2016, The UK had a national referendum to decide whether the country should leave the EU ('Brexit'). The result, a win for the Leave campaign, surprised many political commentators, who had expected that people would vote to Remain. Immediately people began to look for patterns that coud explain the Leave vote: cities had generally voted to Remain, while small towns had voted to Leave. England and Wales voted to Leave, while Northern Ireland and especially Scotland voted to Remain.

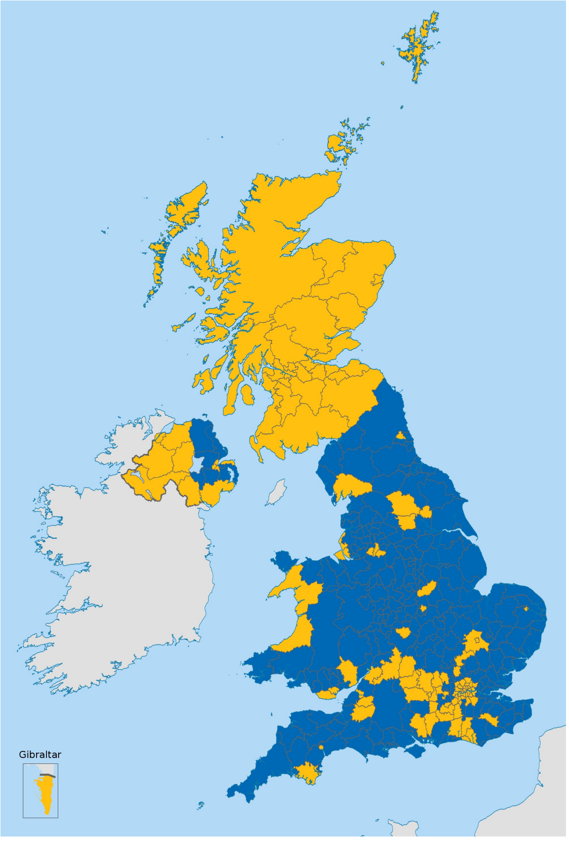

Figure 1: EU referendum vote by electoral ward. Yellow indicates Remain, blue indicates Leave

In the next few days, the Guardian newspaper presented some apparent demographic trends in the vote, based on the ages, incomes, education and class of different electoral wards (https://www.theguardian.com/politics/ng-interactive/2016/jun/23/eu-referendum-live-results-and-analysis). The Guardian's analysis stopped at showing these results graphically, and commenting on the apparent patterns. We will go one better by doing some real statistical analysis of the data.

I have scraped the data from the Guardian's plots into a data file (brexit.csv) which you can download from MINERVA

There are 6 attributes in the data. The 5 possible input variables are:

- abc1: proportion of individuals who are in the ABC1 social classes (middle to upper class)
- medianIncome: the median income of all residents
- medianAge: median age of residents

- withHigherEd: proportion of residents with any university-level education
- notBornUK: the proportion of residents who were born outside the UK

These are normalised so that the lowest value is zero and the highest value is one.

The output variable is called voteBrexit, and gives a TRUE/FALSE answer to the question 'did this electoral ward vote for Brexit?' (i.e. did more than 50% of people vote to Leave?).

Tasks (week 6):

1. Fit a logistic regression models using all of the available inputs. Identify the direction of each effect from the fitted coefficients. Compare these with the plots shown on the Guardian website. Do they agree?
2. Present the value of each coefficient estimate with a 95% confidence interval. Which input would you say has the strongest effect?
3. Using AIC, perform a model selection to determine which factors are useful to predict the result of the vote. Use a 'greedy' input selection procedure, as follows: (i) select the best model with 1 input; (ii) fixing that input, select the best two-input model (i.e. try all the other 4 inputs with the one you selected first); (iii) select the best three-input model containing the first two inputs you chose, etc. At each stage evaluate the quality of fit using AIC and stop if this gets worse.

Tasks (week 7):

1. Use the **rpart** package to create a decision tree classification model. Visualise your model and intepret the fitted model.
2. Compare your decision tree model and your logistic regression model. Do they attribute high importance to the same factors? How do you intepret each model to explain the referendum vote?
3. Which model would you use if you were explaining the results for a newspaper article, and why?