Problem 8: UK Election Results, 2015-2019

Assigned: 23 November
Due: 17 December
Maximum Mark: 25 Points
Maximum Submission: 10 pages

Note: this is a lengthy problem. Use the Rmarkdown chunk options to abbreviate output to ensure you do not exceed the page limit!

The following three files on Canvas (obtained from the House of Commons Library) contain detailed results from the last three UK general elections: `GE2015-results.csv`, `GE2017-results.csv`, and `GE2019-results.csv`.
Columns relevant to this problem are:

ons_id:	Office of National Statistics constituency identifier code
country_name:	Nation of the UK (England, Scotland, Wales, or NI)
region_name:	Region of the UK
valid_votes:	Total number of valid votes cast
con:	# of votes for the Conservative party candidate (elections only)
lab:	# of votes for the Labour party candidate (elections only)

A separate file (`demographics.csv`) contains demographic data for UK constituencies:

ons_id:	Constituency identifier code
income:	median income in £
age.0.15:	% of population between age 0-15
age.65.over:	% of population age >65
foreignborn:	% of population not born in UK
employment:	% employment rate
outofwork:	% out of work
white:	% white
commute.car:	% commuting by car
commute.bike:	% commuting by bike
health.good:	% rating their health as "good"
health.bad:	% rating their health as "poor"

In this problem you will construct a prediction model to investigate trends in the Conservative vote share across different consistencies and different years. The response variable will be the Conservative vote proportion (out of the number of valid votes cast), while the explanatory variables will be constituency demographic information, plus the year of the vote, nation, and/or region.

1. (a) Load all relevant CSV files into R data frame variables. Make sure that any relevant categorical data is stored as a factor, not as strings.
(b) Add the year of the election (2015, 2017, or 2019) as an additional, *categorical* column to each results data frame.
(c) Merge the three separate year-specific results data frames into a single data frame containing all results. (This data frame should have 1950 rows and 33 columns.)
(d) Merge the combined results data frame together with the demographics data (paired by constituency code) to produce a single data frame.
(e) Calculate the Conservative vote share, then apply a logit transform to this value. Assign the transformed variable to a new column of the data frame: this will be your response variable. If the transform produced any infinity or bad values (associated with zero votes), remove the offending rows from the data frame.

(f) Further simplify this dataframe to remove extraneous columns (columns that do not contain either your chosen response variable or an explanatory variable). *Note*: It might be easiest to create a new dataframe (using the `data.frame` command) containing *just* the columns you want to retain.

(g) Remove rows associated with Northern Ireland constituencies from the data set (British mainland political parties have a minor presence in Northern Ireland).

(h) Finally, use the summary() command to summarize your data frame variable. (Make sure the output appears in your report.) The data frame should have 1893 rows and between 9 and 14 columns: verify this using nrow() and ncol().

All these steps should be done in R. Do not modify the data files "by hand". If your code cannot be run independently on the original data files you will not receive credit.

2. Produce a pairs plot showing the relationships graphically between all continuous variables (do not include categorical variables). Make sure the plot is legible by choosing the point and/or image size appropriately.

3. For each of the continuous explanatory variables, determine whether the relationship with the response variable is positively correlated (+), not significantly correlated (o), or negatively correlated (-). Summarize your results in a table.

4. For categorical explanatory variables, determine whether the *variable* is associated with a statistically significant difference in the response variable (Y) or not (N). If there is an association, state which group predicts the highest value of the response variable and which group predicts the lowest value of the response variable. Summarize your results in a table.

5. Carry out a series of multiple linear regression analyses on the data:

(a) A null model.

(b) A regression using all explanatory variables with no interactions.

(c) A regression using all explanatory variables, *plus* one-way interactions between year and each of the other variables.

(d) A regression using all explanatory variables interacting (one-way) with all other explanatory variables.

(e) A stepwise-simplified version of model (c).

(f) A stepwise-simplified version of model (d).

(Note: You do not have to show any output for this question, just the R code.)

6. For each of the models above, calculate the following:

(a) Degrees of error freedom

(b) Degrees of model freedom

(c) Multiple R-squared

(d) Adjusted R-squared

(e) Residual standard error

(f) Akaike information criterion

(h) PRESS

Summarize the results in a table.

7. Which model is "best"? *Briefly* explain any reservations or caveats.

8. For the "best" model you have chosen, produce a four-panel check plot. Note if any issues of concern are present.

9. Identify one categorical variable, one continuous variable, and one interacting variable pair, all with highly significant ($p < 10^{-12}$) contributions to the variance reduction in the "best" model. For each, write one or two sentences explaining what the model implies about the relationship between the relevant explanatory variable(s) and response variable.