UNIVERSITY COLLEGE LONDON

DEPARTMENT OF POLITICAL SCIENCE/SCHOOL OF PUBLIC POLICY

POLS0010 Data Analysis Term II

ESSAY QUESTIONS 2019

Guidelines for Completing and Submitting POLS0010 Term II Essay

- Read the guidelines below to avoid losing unnecessary marks.
- The assessment is due on Tuesday 30th April 2019, 2.00pm. **It has two parts (I and II), both of which need to be submitted together. Part I of the essay is worth 60 marks. Part II is worth 40 marks.**
- Please follow all designated Department of Political Science submission guidelines. These may be different to those of your home department. You must submit one copy of your essay via Turnitin.
- The datasets for the essay can be found in the 'POLS0010 Term 2 Assessment 2019' folder in the 'Assessment Folder' section of Moodle
- **The word limit for both Parts I and II is 3,000 words**, excluding your R script appendix (see below). You can divide the word limit as you like between the two parts.
- This is an assessed piece of coursework for the POLS0010 module; collaboration and/or discussion with anyone is strictly prohibited. The rules for plagiarism apply and any cases of suspected plagiarism of published work or the work of classmates will be taken very seriously.
- Please read the guidelines in the 'Term 2: Essay Information' folder in the 'Assessment' section of Moodle if you require further guidance on logistics or plagiarism
- You may open up the datasets and work on the essay questions anytime up until the submission date. There is no limit on the number of times you may open the data files. Be sure to save your data files and R script file.
- You should include a copy of your R script as an appendix to your essay. **FAILURE TO INCLUDE THE R SCRIPT WILL INCUR A 10 POINT PENALTY.** Note that your R script file should be neatly presented and easy to follow, including comments indicating the question being addressed.
- Include in your write-up all charts/plots and tables that you produce
- Answers should be written in complete sentences; no bulleting or outlining.
- You may assume the methods you have used (e.g. logit regressions, etc) are understood by the reader and do not need definitions, but you do need to say which techniques you have used and why.
- As this is an assessed piece of work, you may not email/ask the course tutors for help with the essay questions.

PART I

This part of the final essay contains two questions. You must answer both of them. Question A is worth 30 points and Question B is worth 25 points.

Up to an additional five points will be awarded for clarity of presentation, especially tables and figures. See Lecture 6 for guidelines on presentation.

Both questions require you to write a brief report. It is up to you how you structure the reports, but it is advisable to keep introductory material to a minimum, given the word limit. Your reports should discuss your methods, your results and the conclusions that you draw from them.

QUESTION A: Switzerland's Gun Control Referendum [30 points]

In Switzerland in 2011, a legally binding referendum was held that would have banned people from keeping guns at home, as well as introducing stricter background checks for those wishing to purchase them. The referendum failed, with 56% of voters opposing it. For this question, suppose that the referendum is going to be repeated next year, and the pro-gun-control campaign asks for your advice.

Specifically, the campaign group want to run an advertising campaign targeted at groups who are most likely to support the new referendum, to persuade them to turn out and vote. Your job is to tell them which types of people are most supportive of gun control. To help measure the likely effectiveness of their advertising, they also want to know how much each characteristic matters in explaining support. You'll use a survey of voters taken after the first referendum that asked about support for gun control. You need to:

i) Choose a logit model that predicts support for gun control, carefully justifying your selection of variables for the model. You must use a minimum of three independent variables.

ii) Present the model's findings in ways that clearly explain how much the variables matter in explaining support for gun control

You should present your approach and your findings in the form of a brief report. It should conclude by explaining which types of people you think the campaign should target. The dataset is called "s" and is contained in the file "swiss.Rda". It contains the following variables for each individual in the survey:

Variable name	Variable description
VoteYes	dependent variable: =1 if respondent voted for gun control, 0 otherwise
female	=1 if female, 0 otherwise
age	in years
LeftRight	individual's own assessment of how right-wing they are on a scale from 0-9, where higher values mean more right-wing [treat this as a continuous variable]
trust	respondent's trust in government. =1 if trusts government, 0 otherwise
university	=1 if respondent has a university degree, 0 otherwise
urban	=1 if respondent lives in an urban area, 0 otherwise
suburb	=1 if respondent lives in a suburban area, 0 otherwise

QUESTION B: Estimating Constituency-Level Results from the EU Referendum [25 points]

In the 2016 UK referendum on leaving the EU, the results of the vote were not released for individual electoral constituencies. However, many scholars would like to know why people voted to leave the EU, and how support for leaving differed across constituencies. One previous study has already estimated constituency-level support for 'leave' in an authoritative way. Your tasks in this question are (i) to produce estimates of the percentage of voters that voted 'leave' in every constituency using multilevel modeling and post-stratification that are as close as possible to this existing set of estimates, as measured by the Mean Absolute Error (MAE), and (ii) to use your results to explain why people voted to leave.

You need to:

i) Estimate an appropriate logistic multilevel model explaining voting for leave, using the predictors in the dataset.[1]

ii) Present the multilevel model results and interpret how the variables affect voting to leave the EU (**Note:** you do **not** need to discuss statistical significance).

iii) Produce post-stratified estimates of the percentage of people who voted 'leave' in all 631 constituencies in England, Scotland and Wales

iv) Compare your results to the existing estimates using the Mean Absolute Error

You should present and explain your approach and results in a brief report, explaining why your estimates do or not perform well compared to the existing estimates. **Note:** if you cannot get very close to the existing results, do not worry. Your grade depends on the quality of your analysis, presentation and interpretation, not how close your results are to the existing estimates.

The **survey data** is called "e" and is in the file "eusurvey.Rda". It comes from the 2017 British Election Study and it contains the following variables:

Variable name	Variable description
cname	constituency name
ccode	constituency code
leave	dependent variable: =1 if respondent voted to leave EU, 0 if respondent voted to remain in the EU
votecon	=1 if respondent voted Conservative in the 2015 election, 0 otherwise
voteukip	=1 if respondent voted UKIP in the 2015 election, 0 otherwise [note: UKIP is the United Kingdom Independence Party, which campaigns in favour of the UK leaving the EU]
female	=1 if female, 0 otherwise
age	in years
highed	=1 if respondent is educated to degree level or higher, 0 otherwise
lowed	=1 if respondent has no educational qualifications, 0 otherwise
c_con15	percent vote for Conservative party in the constituency, 2015 election
c_ukip15	percent vote for UKIP in the constituency, 2015 election
c_unemployed	constituency unemployment rate, percent
c_whitebritish	percent of constituency population who are white British
c_deprived	percent of constituency population living in poverty

[1] As in the practical exercise, use the option "nAGQ=0" to avoid estimation errors

Post-stratification data for the 631 constituencies is called “post” and is contained in the file “eupoststrat.Rda”. Each row contains one particular demographic group in one constituency. In addition to the variables in “e”, it also contains these variables:

Variable name	Variable description
c_count	Number of people in the demographic group
c_total	Number of people in the constituency
percent	percent of constituency represented by the demographic group

Finally, the **comparison data** containing the existing estimates by constituency is called “est” and is in the file “existing_estimates.Rda”. In addition to the constituency name and code, it contains the existing estimate of the leave vote share for each constituency (called *estimate*).

PART II

This part of the final essay contains one question. It is worth 40 points. 5 points are reserved for clarity of presentation, especially tables and figures. See Lecture 6 for guidelines on presentation.

The question requires you to write a brief report. It is up to you how you structure the report, but it is advisable to keep introductory material to a minimum, given the word limit. Your report should discuss your methods, your results and the conclusions that you draw from them.

QUESTION C: Describing and Classifying Tweets [40 points]

Many companies monitor social media posts in order to gauge how customers feel about their company and their competitors. For this question, imagine that you have been hired as a consultant by one of the major American airline companies to analyse tweets about airlines. They want to find out how people talk about airlines on Twitter, and then build a predictive tool that can classify tweets in future into 'negative' or 'positive' sentiment toward airlines, to help them respond better to their customers in real time. They have provided you with a dataset of 11,541 tweets about airlines that have been labelled as 'negative' or 'positive' by their staff. The dataset also identifies which airline each tweet is talking about.

Your task is to prepare a brief report that describes the tweets, and recommends a classification method for future tweets. You need to:

1. Use appropriate tools to describe the tweets. In particular, what words are associated with negative or positive sentiment? How does word usage differ across the different airlines?
2. Use your analysis from (1) to build a short dictionary of negative and positive words describing airlines, then use it to classify tweets as 'negative' if they contain more negative than positive language, and 'positive' otherwise [code for creating your own dictionary is provided below]
3. Use an appropriate supervised machine-learning method to classify the tweets into 'negative' and 'positive'
4. Compare the performance of your classifiers from (2) and (3), and use this analysis to decide which one would be the better classifier for the company to use for future tweets

Here is some advice for part (2):

- Your dictionary should contain a minimum of 5 words and a maximum of 15 words in each category
- You are not expected to exhaustively compare the performance of different dictionaries. Instead, simply choose **one** dictionary based on your analysis from (1), and explain how you chose the words.

The dataset for this question is called "tweets" and is contained in the file "tweets.Rda". It contains the following variables:

Variable name	Variable description
text	The text of each tweet
sentiment	Labeled sentiment of each tweet: 1=negative, 0=positive
airline	The airline company featured in the tweet: United, JetBlue, American Airlines, US Airways, Virgin America or Southwest

You should first create a corpus of tweets using the following code:

```
speechCorpus <- corpus(tweets$text, docvars = tweets)
```

Code for creating a dictionary:

You can create a dictionary called "mydict" in R that contains two categories ('negative' and 'positive') using the following code:

```
neg.words <- c()
pos.words <- c()

mydict <- dictionary(list(negative = neg.words,
                          positive = pos.words))
```

You need to insert your chosen lists of negative and positive words in 'neg.words' and 'pos.words'. This dictionary can then be used with the quanteda package in exactly the same way as any of the existing built-in dictionaries.