Midterm coursework
Introduction to Quantitative Research Methods (PUBLG100A/B)

- The coursework will be posted on Moodle on 3 November 2017 at 2pm, and is due on 8 November 2017 at 2pm. Please follow all designated SPP submission guidelines for online submission as detailed on the PUBLG100A/B Moodle page. Standard late submission penalties apply.
- This is an assessed piece of coursework (worth 25% of your final module mark) for the PUBLG100A/B module; collaboration and/or discussion of the coursework 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 seriously.
- As this is an assessed piece of work, you may not email/ask the course tutors or teaching fellows questions about the coursework.
- Along with the coursework itself, the datasets for the coursework can be found in the PUBLG100A/B page on Moodle.
- Coursework should be submitted via the ‘PUBLG100A(B) Essay 1 Turnitin Submission’ link on the course Moodle page. You will need to click the ‘Submit Paper’ link at the bottom of the page. When presented with the ‘Submit Paper’ box, the ‘Submission Title’ should be your candidate number, and you should upload your document into the box provided.
  - Please remember to state only your candidate number on your coursework (your candidate number is made up of four letters and one number e.g. ABCD5). Your name and/or student number must not appear on your coursework.
- The coursework consists of four sections; you must complete each part of each section to achieve full marks.
- Where appropriate, answers should be written in complete sentences. Be sure to answer all parts of the questions posed and interpret the results.
- PLEASE SUBMIT YOUR TYPE-WRITTEN (NUMBERED) ANSWERS IN ONE DOCUMENT. CREATE AT THE END AN APPENDIX SECTION CONTAINING ALL R CODE NEEDED TO REPRODUCE YOUR RESPONSES (you do not need to include the code that failed to run, but just the cleaned-up version. Your code has to work when we run it). FAILURE TO INCLUDE THE R CODE MEANS THAT THE COURSEWORK WILL BE MARKED INCOMPLETE.
  - An example of the formatting we require for the coursework is given in the document ‘PUBLG100 Midterm Coursework Example’, which is also on the Moodle page. Note that the length of this example document is not indicative of the expected length of the coursework.
- You may assume the methods you have used (e.g. t-test, linear regression, etc) are understood by the reader and do not need definitions, but you do need to explain the intuition of these methods.
- Round all numbers to two digits after the decimal point.
- Do not copy and paste *any* brute R output (e.g. `summary(lm(y ~x))`) into your answers. Create a minimally formatted table, e.g. with the `screenreg` command as seen in class. If that does not work, re-create by hand such a table.
- Assign every table and figure a title and a number and refer to the number in the text when discussing a specific figure or table.
- All variable names in the coursework are written in *italics*.

Datasets

"UK General Election 2017 data.Rdata"

This dataset contains the results of the 2017 UK general election, by party, for each of the 632 UK mainland constituencies (excluding Northern Ireland).

Variables:

- Key – Constituency identifier key
- $Country$ – Country of constituency (England = E, Scotland = S, or Wales = W)
- $Region$ – Geographic region of constituency
- $Con17$ – Conservative Party percentage of the constituency vote in 2017
- $Lab17$ – Labour Party percentage of the constituency vote in 2017
- $LD17$ – Liberal Democrat Party percentage of the constituency vote in 2017
- $UKIP17$ – United Kingdom Independence Party percentage of the constituency vote in 2017
- $SNP17$ – Scottish National Party percentage of the constituency vote in 2017
- $Other17$ – Other parties percentage of the constituency vote in 2017
- $PctHomeOwners$ – Percentage of constituency population who own their own home
- $PctWhiteBritish$ – Percentage of constituency population who self-identify as white British
- $PctUnemployed$ – Percentage of constituency population who are long-term unemployed
- $PctHighEducation$ – Percentage of constituency population who have high levels of education (A-level and above)
- $PopDensity$ – Population density of the constituency
- $Winner2010$ – Identity of the party that won the constituency in 2010
- $Winner2015$ – Identity of the party that won the constituency in 2015

You can access this data in two ways.

1. You can download the "UK General Election 2017 data.Rdata" data file from Moodle, and load it into RStudio as we have been doing in class.
2. You can run the following line of code in RStudio and this will load the data directly from github.

```
results <- read.csv("https://uclspp.github.io/datasets/data/UK_General_Election_2017_data.csv")
```

These two ways of loading the data will produce identical results.

1 20 points

You are a researcher who is interested in exploring levels of support amongst business leaders for Britain's exit from the European Union. You conduct a small survey and randomly select 36 CEOs to rate their level of support for an EU exit agreement on a scale from 0 to 100 (where 0 indicates complete opposition, 50 indicates neither support nor opposition, and 100 indicates complete support).

You get the following scores: 98, 74, 65, 78, 17, 65, 82, 72, 68, 74, 59, 95, 51, 94, 39, 34, 68, 17, 90, 86, 65, 57, 33, 78, 29, 80, 65, 29, 82, 46, 55, 50, 32, 43, 90, 3.

1. Calculate the mean and standard deviation of the scores in your sample.

2. Calculate the 95% and 99% confidence intervals around the mean, under the assumption that the survey responses follow a normal distribution. Show your work, step by step.

3. Provide both a statistical and substantive interpretation for the mean and the 95% and 99% confidence intervals. You should conclude whether, on the basis of this sample, you have evidence for or against the statement that "Business leaders back Brexit."

2 20 points

You are a consultant for a major government department. The government has been running a trial jobs training programme, in which 500 unemployed individuals were randomly assigned to two equally sized groups. Individuals in the treatment group ($n_t = 250$) received an 8-week job training programme, while individuals in the control group ($n_c = 250$) received no training.

Following this trial, the government surveyed all individuals in the sample, and recorded their income over the subsequent 12 months. The mean income level of those in the training group was $\bar{Y_t} = £18,593$, with a sample standard deviation of $s_t = £2779$. The mean income level for the control group was $\bar{Y_c} = £18,123$, with a sample standard deviation of $s_c = £3068$.

The minister in charge of your department wants to use the difference in means as the basis for a press release which claims that the training scheme successfully increases earnings. You should write a paragraph explaining why you think it would be either appropriate or inappropriate for the minister to use this figure. You should support your answer with appropriate evidence from a statistical test that you have learned on the course, and you should explain the intuition behind the statistical approach you use.

3 30 points

This question uses data on the results of the 2017 British General Election. The data can be found in the accompanying "UK General Election 2017 data.Rdata" file, and the variables in the data are described on page two of this document.

1. Using the appropriate measures, report and interpret the central tendency and dispersion for the following variables: a) Liberal Democrat party vote share (*LD17*) and b) the winning party of the constituency in the 2015 election (*Winner2015*).

2. Produce scatter plots of the Conservative vote share in 2017 in each constituency (*Con17*) against a) the percentage of homeowners in each constituency (*PctHomeOwners*), and b) the percentage of unemployed people in each constituency (*PctUnemployed*). Provide an explanation of the substantive meaning of these graphs. What do they tell us about patterns of Conservative party support in the UK?

3. Produce a box plot which depicts the distribution of the Labour Party's constituency-level vote share in each UK region (*Region*).

4. Calculate the mean difference in the Labour Party's share of the vote (*Lab17*) in England and in Scotland. Do the same for England versus Wales, and for Wales versus Scotland (hint: you will need to make use of the variable *Country* for this analysis). Conduct t-tests to establish whether these country differences are statistically significant at the 95% confidence level. Explain your results.

   Hint: You can perform a two-sample t-test for the difference in means between two continuous vectors (of arbitrary lengths) in *R* in the following way:

   ```
   t.test(first.sample.vector, second.sample.vector, mu = 0, conf = .95)
   ```

   where `first.sample.vector` is the vector of values for your first subgroup, and `second.sample.vector` is the vector of values for your second subgroup.

5. Estimate a linear regression model where the dependent variable is the Liberal Democrat share of the vote in a constituency (*LD17*), and the independent variable is the percentage of the population of a constituency with high levels of education (*PctHighEducation*).

   - Present a simple table with the output of this regression
   - Interpret the main coefficient of interest (*PctHighEducation*) both statistically and substantively
   - Interpret the estimated intercept term of the regression
   - Interpret the R-squared term of the regression

4 30 points

This question requires you to interpret and communicate the findings of two linear regression models. The data is from an article that studies discrimination against immigrants in Switzerland.

In Switzerland, decisions about citizenship applications of foreign residents are made by direct popular vote (referendum) in the municipalities in which foreign residents wish to settle. In a typical citizenship referendum, local voters are provided with official voting leaflets that explain the citizenship request of an individual, with a detailed description of each immigrant applicant. These leaflets include details on several *personal characteristics* of each applicant. Voters then cast a secret ballot on each individual request, and applicants with a majority of ‘yes’ votes are granted Swiss citizenship.

As we are interested in describing the factors that are related to discrimination, our dependent variable is the proportion of “no” votes received by each applicant. For the purposes of this question, this variable should be considered as continuous. Our observations are therefore at the level of the referendum for each individual citizenship applicant. There are 2429 referendum observations in the data.

We are interested in the effects of the *personal characteristics* of each applicant on the proportion of “no” votes that the applicant receives. We will focus on the following variables:

- *Male* – a binary variable equal to 1 if the applicant is male, and 0 if the applicant is female
- *Married* – a binary variable equal to 1 if the applicant is married, and 0 if the applicant is not married
- *Years since arrival* – a continuous variable for the number of years that the applicant has been living in Switzerland prior to the citizenship application
- *Unemployment* – a binary variable equal to 1 if the applicant is currently unemployed, and 0 if the applicant is currently employed
- *Turkey* – a binary variable equal to 1 if the applicant originally came from Turkey, and 0 if the applicant originally came from another country
- *Years of Education* – a continuous variable for the number of years of education that the applicant has completed
- *Age* – a continuous variable for the age (in years) of the applicant

Use the model results presented in table 1 on page 7. Model 1 presents results from a simple linear regression, where the independent variable is *Unemployment.* Model 2 presents results from a multiple linear regression which includes a number of explanatory variables. The dependent variable for both models is the proportion of “no” votes received by each citizenship applicant.

Your task is to interpret the models and write up the results as if you were writing the discussion for publication in a major journal/book. Interpret the two models statistically and substantively, and in comparison to one another. You should focus on determining which variables have coefficients that are significantly different from zero, and what the effect sizes mean in substantive terms. Simply listing the significant effects will be insufficient to receive full marks. You should also comment on how the estimates differ between the two models, and on the fit statistics of the two models.

Table 1: Multiple linear regression

	Proportion voting 'No' in referendum	
Independent variables	Model 1	Model 2
Unemployment	19.70	5.60
	(5.32)	(2.65)
Male		0.74
		(0.61)
Married		0.36
		(0.80)
Years_since_arrival		−1.75
		(0.39)
Turkey		13.26
		(1.23)
Years of Education		-1.20
		(0.93)
Age		0.01
		(0.02)
Constant	35.9	32.7
	(7.29)	(3.29)
Observations	2,429	2,429
R^2	0.47	0.67

Note: Figures in parentheses are the standard errors of the regression coefficients