Midterm Coursework

Introduction to Quantitative Research Methods (PUBL0055)

Instructions

- The coursework will be posted on Moodle on 2 November 2018 at 6pm, and is due on 7 November 2018 at 2pm. Please follow all designated SPP submission guidelines for online submission as detailed on the PUBL0055 Moodle page. Standard late submission penalties apply.
- This is an assessed piece of coursework (worth 25% of your final module mark) for the PUBL0055 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 PUBL0055 page on Moodle.
- Coursework should be submitted via the 'PUBL0055 - Term 1 Assessments' 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 three 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.
- The word count for this assessment is 1000 words. This does not include the appendix, or any words (or numbers) contained within tables.
- Please submit your type-written (numbered) answers in a single document. Create an appendix section at the end which contains all the R code needed to reproduce your results (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.
- 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*.

Part 1: 25 points

Questions in part 1 do not require using R, but you can use R if you choose to do so. You are also allowed to use a calculator for any of the questions in part 1.

Background

You're working as a consultant for a candidate for the United States Senate in the upcoming midterm elections. The candidate is running as an Independent against a Democratic incumbent and a Republican challenger. With just a few days before the election, the candidate has asked you to analyse the latest polling data. The dataset consists of support for your candidate among likely voters who are either registered Democrats or registered Republicans.

Dataset

For each group, you have a poll of 1000 voters in each of the 33 counties in the state. The dataset shows how many respondents from each group support your candidate.

Registered party	**Number of voters who support the candidate (out of 1000)**
Democrat	170, 448, 369, 182, 394, 206, 258, 433, 503, 426, 409, 421, 355, 516, 226, 535, 489, 337, 464, 508, 325, 521, 533, 533, 237, 476, 312, 493, 411, 464, 192, 689, 398
Republican	377, 373, 284, 249, 402, 336, 377, 263, 352, 341, 290, 365, 416, 299, 334, 353, 363, 346, 285, 371, 387, 333, 431, 349, 307, 294, 374, 254, 366, 416, 274, 343, 415

1a. Descriptive statistics

- Calculate the mean and standard deviation of registered Democrats who support your candidate.
- Calculate the mean and standard deviation of registered Republicans who support your candidate.
- Present your answer in a formatted table as shown below:

	Mean	Std Deviation
Registered Democrats		
Registered Republicans		

1b. Difference in means and confidence interval

Use the data above to complete the following calculations. For each question you should show how you arrived at your answer. (i.e. Calculating the answer in R and reporting it here is not sufficient to get full marks)

- Calculate the difference in means between registered Democrats and registered Republicans.
- Calculate the standard error of the difference in means.
- Calculate the 95% and 99% confidence intervals of the difference in means (you can use the normal distribution for this question).

1c. Summary

- Write a brief summary of the findings that include both statistical and substantive interpretation of the difference in means and the confidence intervals. Your summary must clearly state whether, on the basis of the polling responses, you have evidence that the candidate has more support among registered Democrats or registered Republicans.

Part 2: 35 points

All questions in part 2 require the use of R. Do not include R code in your answers. Instead, create an appendix section at the end of your document containing all R code necessary to reproduce your responses.

Background

Part 2 is based on a study of premature mortality in Great Britain between 2012 and 2014. The dataset from this study includes information on premature mortality for 378 local authorities in Great Britain from 2012 to 2014. Premature mortality is measured as the number of individuals that die before the age of 70 in a cohort of 100,000. In addition to total premature mortality, the dataset also includes a breakdown by gender and socioeconomic indicators such as income, education and employment for each local authority.

Dataset

You can access this data in two ways:

1. You can download the `Premature Mortality in Great Britain 2012-2014` data file from Moodle, copy it to your working directory, and load it into R as we have been doing in class.

   – or –

2. You can run the following line of code in R and this will load the data directly from the course website:

```
pmdata <- read.csv("https://uclspp.github.io/datasets/data/pmgb2012_2014.csv")
```

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

Codebook

The codebook describes the variables in the dataset.

Variable	**Description**
code	Unique idenifier for each local authority
country	1 = England, 2 = Scotland, 3 = Wales
pop_density	1 = low, 2 = medium, 3 = high
pmdeaths_total	Number of premature deaths out of 100,000
pmdeaths_female	Number of premature deaths among women, out of 100,000
pmdeaths_male	Number of premature deaths among men, out of 100,000
mean_income	Mean income in the local authority
edu_level3	Qualification: proportion of the population with A level
edu_level4	Qualification: proportion of the population with degree-level education or equivalent

2a. Descriptive Statistics

- Using the appropriate measures, report and interpret the central tendency and dispersion for the following variables:
  - *edu_level3*
  - *edu_level4*
  - *pop_density*

2b. Visualization

- Produce a scatter plot of premature mortality (*pmdeaths_total*) on the y-axis and degree-level education (*edu_level4*) on the x-axis
- Provide an explanation of the substantive meaning of the graphs. What do they tell us about the association between premature mortality and levels of education in Great Britain?
- Produce a box plot that compares premature mortality in England, Scotland, Wales.
- What does the plot tell us about how premature mortality varies across the three countries?

2c. Difference in Means

- Calculate the mean difference between premature mortality among men and women in Great Britain.
- Conduct t-test to establish whether the difference between the premature mortality of men and women is statistically significant at the 95% confidence level.
- Interpret the results of the t-test both statistically and substantively
- Interpret the confidence interval of the difference in means

2d. Linear Regression

- Estimate a linear regression model to analyse the relationship between mean income and premature mortality in each local authority. The dependent variable is *pmdeaths_total* and the independent variable is *mean_income*
- Present a table with the output of the regression model
- Interpret the main coefficient of interest (*mean_income*)
- Interpret the estimated intercept term of the regression model
- Interpret the R^2 term of the regression model

Part 3: 40 points

This question requires you to interpret and communicate the findings of two linear regression models. The data is from an article that studies the relationship between salaries of legislators and representation of the working-classes in state legislatures in the US.

Background

If politicians in the United States were paid better, would more working-class people become politicians? It is often argued that if politicians are paid too little, then it is economically too difficult for lower-income citizens to hold positions of office. This could mean that low-paying political jobs lead to the under-representation of working-class people in politics. On the other hand, if politicians are paid more, then holding political office might become more attractive to wealthy people, and this might also lead to the under-representation of working-class people. To investigate these two contrasting hypotheses, we will examine data on the salaries paid in different state legislatures in the US and the percentage of legislators who come from working-class backgrounds.

Dataset

The dataset includes salaries of state legislators from all 50 states in the US. It also includes variables measuring information unique to each state such as the length of the legislative session and the number of staffers in each legislature. The occupational backgrounds of legislators are also included, as well demographic data on the makeup of the population in each state. A detailed description of the dataset is provided in the table below.

Variable	**Description**
pct_worker	Percentage of legislators from working-class backgrounds
salary	Average salary of legislators in $100,000s
session_len	Length of legislative session (in days)
staff_size	Average number of full-time permanent staffers in the legislature
term_limits	Binary indicator (0 or 1) of term limits for state legislators
income	Average per-capita income (in $1000s)
income_inequality	Percentage of income to top 1% of earners
pct_union	Percentage of workers belonging to a labour union
pct_black	Percentage of state residents who are Black
pct_urban	Percentage of state residents living in urban areas
poverty_rate	Percent of state residents living below the poverty line

3a. Multiple Linear Regression

This question requires you to interpret and communicate the findings of two linear regression models from Table 1.

Model 1 presents results from a simple linear regression, where the independent variable is *salary.* Model 2 presents results from a multiple linear regression which includes a number of explanatory variables. The dependent variable for both models is the percentage of legislators from working-class backgrounds.

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.

	Model 1	Model 2
(Intercept)	155.49	−199.48
	(41.00)	(103.85)
salary	−0.56	−0.61
	(0.10)	(0.13)
term_limits		0.26
		(0.84)
income		−0.03
		(0.05)
income_inequality		−0.26
		(0.11)
poverty_rate		−0.05
		(0.07)
pct_union		0.12
		(0.04)
pct_black		−0.06
		(0.02)
pct_urban		−0.03
		(0.02)
R^2	0.19	0.35
Adj. R^2	0.18	0.31
Num. obs.	200	200

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

Table 1: Legislative salaries and working-class representation