Final Coursework

Quantitative Data Analysis (POLS0083)

Instructions

Submission Formalities

- The final assessment is posted on Moodle on 3rd January 2023 at 9am, and is due on 11th January 2023 at 2pm. Please follow all designated SPP submission guidelines for online submission as detailed on the POLS0083 Moodle page. Standard late submission penalties apply.
- The coursework should be submitted via the 'POLS0083 - Assessment 2 - 2,000 Word Final Coursework (70%)' 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 submission.
- This is an assessed piece of coursework (worth 70% of your final module mark) for the POLS0083 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 teaching team questions about the coursework.
- Along with the coursework questions, the necessary data sets for the coursework can be found on the POLS0083 page on Moodle.

Coursework Formalities

- The word count for this assessment is 2,000 words. This does *not* include the code, your output, or any words (or numbers) contained within tables or figures. The word count must be clearly indicated on the first page of the assessment submission. Standard word limit penalties apply.
- The coursework consists of two separate sections, each with several questions and subquestions. The marks allocated for each section are indicated in the text. You must complete each question to achieve full marks.
  - Together, the questions are worth 90 marks.
  - 10 marks are reserved for presentation (see below).
- Please submit your type-written (numbered) answers in a single document (preferrably a pdf file, but word files or equivalent are also allowed). Make sure that you include the code, the output (plots, tables etc), and the answers in the document. Double check after uploading your document that all graphs, table, code, and written answers are visible.
- Unless otherwise stated, answers should be written in complete sentences. Be sure to answer all parts of the questions posed and provide a substantive interpretation of the results.
- You can integrate the code with the answers (make sure that it is completely visible), as shown for example in the seminar worksheet solutions. Alternatively, you can create an appendix section at the end which contains all the R code needed to reproduce your results.

  - In either case, your code has to work when we run it. *You do not need to include the code that failed to run*, but just the well-annotated, cleaned-up version.
  - *If you do not provide the code to a question which requires it, any written answer to that question will be disregarded.*
  - Do not screenshot or copy and paste *any* brute R output (e.g. `lm(y~x)`) into your answers. If applicable, create formatted tables that are easy to read.

- Round all numbers to two or three digits after the decimal point.
- 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.
- You may assume that references to *methods* you have used (e.g. difference in means, linear regression, etc) are understood by the reader and do not need definitions, but you do need to be able to explain what they do and how they apply to answering the question.

Presentation (10 marks)

Points will be *deducted* for bad presentation, which includes (but is not limited to):

- Failure to write the answers in full sentences
- Failure to clearly indicate which question is answered where and which code pertains to which question
- Including screenshots from `R` output
- Including long print outs of data sets and objects (e.g., using View(), show())
- Reporting unrounded numbers
- Including unnecessary code or output
- Presenting figures with no or unclear axis labels (or labels that are the unedited variable name in the dataset)
- Presenting tables that are hard to read/not well formatted
- Presenting unnumbered and/or untitled tables and figures
- Referring to the variables in-text by their unedited name from the dataset

Section 1:

Corruption Incidence and Local Health Councils in Brazil

The relationship between well-functioning local institutions and corruption is an important issue across the globe.

To explore the relationship between local governance in the health sector and incidence of corruption, we will conduct some analyses loosely based on the research conducted in Brazil by Avelino, Barberia, and Bilderman (2014).

To measure how established a health council is, Avelino et al. recorded how long a local health council had been established at the time of the audit. Health councils are responsible for overseeing the local provision of health services in the municipality, as well as approving the municipal health budget and monitoring expenditures.

The authors collected data from a set of Brazilian municipalities that had been randomly selected to be audited by the federal government. Detailed memos are produced for each source of federal funding in the selected municipalities, including for health grants. Auditors include “evidence reports” to these memos, depending on the number of irregularities identified.

To measure corruption, the authors looked through the evidence reports for federal grants, and counted the number of such reports that mention irregularities. The municipal-level percentage of these evidence reports that mentioned irregularities was treated as the corruption index score, ranging between 0 and 100.

The author’s primary hypothesis is that the more well-established a health council is in a municipality, the more likely corrupt practices will be uncovered. This is because the authors believe that “local governments acquire expertise to manage the health system over time and each additional year represents a marginal gain in local capacity”, including in the control of corruption. Thus, we would expect more established health councils (that is, those that are older) to have a lower incidence of corruption.

The variables that will be included in the analysis are as follows:

Name	Description
`municipality`	Unique code number for municipality
`corruption`	Numeric corruption index (0-100), percent of evidence reports mentioning irregularities
`council.age`	Number of years the health council has been established when the municipality was audited
`margin`	Margin between the elected mayor and the runner-up candidate in the previous election, in percentage points
`reelected`	Dummy variable that identifies whether the mayor was in their second term at the time of the audit. If so, the value is “1”, otherwise “0”
`transfers`	Federal government grants as a share of total health expenditures in the municipality (percentage)
`poverty`	Percentage of individuals below the poverty line

The data is stored in `brazil.csv`. Once you have downloaded this file and placed it in the relevant folder, it can be loaded into R as follows:

```
brazil <- read.csv("data/brazil.csv")
```

Question 1 (6 marks)

We are first interested in exploring the data set and conducting some descriptive analyses.

a. For how many of the municipalities do the authors have no data on the age of the health council? **(2**

For Q1:
- a. OK
- b. MISSING interpretation and check names of axis
- c. OK

marks)

b. Plot and interpret a boxplot of the health council age (`council.age`). **(2 marks)**

c. Interpret the median and mean of the variable `corruption` **(2 marks)**

For Q2:
- a. OK
- b. OK but compare to sofi's
- c. MISSING

Question 2 (8 marks)

We then proceed with a simple linear regression analysis.

a. Fit and present a simple linear regression with the corruption index as the outcome and age of council as the explanatory variable. **(1 mark)**

the independent variable (X)
lm (dependent variable - independent variable, data =)

b. Discuss the statistical and substantive significance for the intercept and the estimated regression coefficient for `council.age`. Is the intercept meaningful in this model? **(4 marks)**

c. Under which assumptions can we interpret the regression coefficient as the average effect of council age on corruption? **(3 marks)**

FOR Q3:
- a. OK
- b. OK but check
- c. MISSING discussion and comparison
- d. OK

Question 3 (10 marks)

As the authors did in the original study, we now add a number of other municipal-level explanatory variables to our regression model: margin of victory for the Mayor in the last election; whether the Mayor is re-elected; and the poverty level.

a. Fit a multiple linear regression model, adding `margin`, `reelected`, and `poverty` to the model in the previous question. Present this model alongside the simple linear regression model. **(1 mark)**

b. How has the estimated coefficient for `council.age` changed? What does that tell us about the variables we have added to the model? **(3 marks)**

c. Discuss and compare the model fit for the multiple and the simple linear regression models. **(3 marks)**

d. What is the predicted corruption index score for a municipality health council that is 10 years old, that has a re-elected Mayor, where the Mayor won the last election by 12 percentage points, and where the poverty level is 50? **(3 marks)**

Question 4 (14 marks)

Although it was not explored in the original paper, we are interested in whether the relationship between the age of the health council and incidence of corruption differs between municipalities with and without a reelected Mayor.

a. Fit a multiple linear regression model, adding an interaction between `reelected` and `council.age` to the multivariate model from the previous question. Present this model alongside the model without the interaction. **(1 mark)**

b. Interpret the estimated coefficient for `margin`. You do not need to discuss statistical significance. **(2 marks)**

c. Calculate and interpret the 95% confidence interval for the estimated coefficient of poverty. **(3 marks)**

d. Interpret the relationship between `council.age` and `corruption`. **(3 marks)**

e. Using the model you estimated in *4.a*, calculate the fitted values for health councils with ages between 0 and 20 years, separately for municipalities with and without reelected mayors. Set the electoral margin to 10 percent and poverty score to 50 percent. Present the fitted values visually and describe what the graph shows. **(5 marks)**

FOR Q4:
- a. OK
- b. MISSING interpretation
- c. MISSING interpretation
- d. MISSING interpretation
- e. OK ma check again

Section 2:

Asset trading and attitudes to peace

What are factors that determine the extent to which people support peace processes? Research suggests that (ethnic) violence tends to harden ethnic identities and increase out-group discrimination (see, e.g., Shayo & Zussman 2017). While one approach to decrease exclusionary attitudes has been to encourage individuals to put themselves in the shoes of members of the out-group (see, e.g., Adida, Lo & Platas 2018), Jha & Shayo (2019) argue that exposure to financial markets might be another factor that can help promote more inclusionary, pro-peace attitudes in protracted conflicts. The basic idea is simple: conflict tends to be financially costly. Financial markets "demonstrate the shared risks from conflict and the returns from peace" (p.1561-1562). Therefore, individuals who invest in financial assets that are negatively affected by conflict (e.g., stocks of companies located in conflict areas) might have a better understanding of the financial risks of conflict and, in addition, be (financially) negatively affected by conflict. Accordingly, they hypothesize that individuals exposed to financial markets will become more pro-peace. To test their theory, they conduct a field experiment in Israel, in which they randomly assigned a sample of Israeli voters to a financial asset treatment group or a control group. Individuals in the treatment group received vouchers to invest in specific stocks or indices from Israel and the Palestinian Authority. Participants were surveyed before and after the experiment.

This section is loosely based on the experiment that Jha & Shayo conducted and uses a (modified) version of the data they collected. You can download the data set as `trading.csv` from the POLS0083 Moodle page. We will examine whether investing in Israeli or Palestinian stocks affects pro-peace attitudes among Israelis.

The data set contains the following variables:

Variable name	Description
`assettreat`	Treatment assignment, `1` if respondent was assigned to a treatment group (Israeli or Palestinian stock), `0` if respondent was assigned to the control group
`asset_comp`	Treatment uptake, `1` if respondent was assigned to a treatment group and actually completed the instruction session and accepted their assigned assets, `0` otherwise
`isrstock`	Israeli stock treatment, `1` if respondent was assigned to trading Israeli stocks, `0` otherwise
`palstock`	Palestinian stock treatment, `1` if respondent was assigned to trading Palestinian stocks, `0` otherwise
`tradestock6all`	Pre-treatment financial market exposure, `1` if respondent bought/sold shares in the 6 months before the experiment, `0` otherwise
`age`	Respondent age (in years) before experiment began
`faminc`	Monthly family income in NIS (New Israeli Shekel) before experiment began
`religion`	Respondent religion before experiment began
`female`	Respondent sex, `1` if respondent is female, `0` otherwise
`BA_or_higher`	Respondent education, `1` if respondent had a BA degree or higher before experiment began, `0` otherwise
`left_2013`	Pre-treatment (2013) voting behaviour, `1` if respondent voted for a left-wing (pro-peace) party in 2013, `0` otherwise
`left_2015`	Post-treatment (2015) voting behaviour, `1` if respondent voted for a left-wing (pro-peace) party in 2013, `0` otherwise
`p_index_2013`	Pre-treatment (2013) peace attitudes index, higher values indicate higher support for peace. Index is based on respondent's answer to four questions, you can check out Table B14 in the paper's B Supplementary Appendix for a list of questions
`p_index_2015`	Post-treatment (2015) peace attitudes index, higher values indicate higher support for peace

You can load the data set by using the following command:

```
trading <- read.csv("data/trading.csv")
```

Question 1 (9 marks)

Data preparation and description:

a. How many individuals received Israeli stocks, how many received Palestinian stocks, and how many were assigned to the treatment group, but did not receive Israeli or Palestinian stocks? **(2 marks)**

b. Drop the individuals who were assigned to the treatment group, but received neither Israeli nor Palestinian stocks from the data set. **For the remainder of Section 2, we will work with this subset. (1 mark)**

c. Among those who were treated, what is the proportion of those who took up the treatment (i.e., participated in the training and actually traded afterwards)? Is there a difference in uptake between respondents who were assigned Israeli and Palestinian stocks? Have a think about why (or why not) this might be the case. **(3 marks)**

d. Let's also explore our main outcome, support for peace. Produce a graph to inspect the central tendency and the spread of the peace index in 2013 (pre-treatment) and 2015 (post-treatment). What can you tell us about the spread? What would you consider a "substantively meaningful" change in a respondent's attitude towards peace? **(3 marks)**

Question 2 (9 marks)

Now we are interested in whether being exposed to financial markets, i.e., trading stocks, affects attitudes to peace.

a. Estimate the impact of being assigned to receive the treatment (`assettreat`) on attitudes toward peace using the difference in means. Make sure to only use the post-treatment measure (`p_index_2015`). Does receiving stocks to trade increase support for peace? Present your output, provide a brief explanation and comment on the substantive significance of the effect. **(3 marks)**

b. Calculate the standard error of the estimate ("by hand" in R) and use it to compute 99% confidence intervals. Interpret your results. **(3 marks)**

c. Explain the concept of a "sampling distribution". What is the shape of the sampling distribution in this example? Why is it relevant here? **(3 marks)**

Question 3 (9 marks)

The data set stems from a field experiment. Participants were randomly assigned to receive stocks or not. One worry might be that the randomisation did not work properly.

a. Why might this be a worry? Why is a randomised field experiment useful when it comes to estimating the causal effect of stock trading on support for peace? **(3 marks)**

b. If indeed randomisation did not work properly, which potential confounder (focus on factors that we have data on in this example) are you particularly worried about? Explain how this factor could bias our estimate of the effect of trading. **(3 marks)**

c. Test whether randomisation worked properly by examining the balance of two potentially confounding variables between treatment and control groups. What do you conclude? **(3 marks)**

Question 4 (14 marks)

So far, we have focused on estimating the average treatment effect of being assigned to receive stocks. However, some participants in the treatment group were assigned Israeli stocks and others Palestinian stocks. We will now examine whether there are heterogeneous treatment effects, depending on whether participants were assigned to invest in Israeli or Palestinian stocks.

a. If the authors' hypothesis about the effect of being exposed to financial markets is true, should we expect that the treatment effect is similar for both respondents assigned to receive Israeli stocks and those assigned to receive Palestinian stocks? Explain your answer. **(2 marks)**

b. Let's examine whether being exposed to assets of the opposing group in conflict could have particularly strong effects on attitudes towards peace. Formulate a null and an alternative hypothesis about the difference in the effect between the two treatment groups. **(3 marks)**

c. Test your hypothesis. Present and interpret your results. What's your answer? **(3 marks)**

d. One important concept in quantitative data analysis is statistical significance. Briefly explain what statistical significance is, and what influences whether it is high or low in the case of your analysis in *4.c.* **(3 marks)**

e. In your analysis in *4.c* are you worried more about making a Type I or Type II error? Briefly explain which you are worried about more and why. **(3 marks)**

Question 5 (7 marks)

We could also use this data set to employ a difference-in-differences design.

a. Explain in your own words under which assumptions we can use a difference-in-differences design to identify the causal effect of a treatment. Are these assumptions met in this case where we are looking to identify the causal effect of being exposed to financial markets on attitudes to peace? **(3 marks)**

b. Compute the difference-in-differences point estimate of the effect of being assigned stocks on attitudes to peace. Interpret your finding and comment on the substantive and statistical significance of your results. **(4 marks)**

Question 6 (4 marks)

Considering your findings (your answers to questions 1-4) briefly evaluate the internal and external validity of your results. What do we mean by internal and external validity? Comment on whether and why your findings have high/low external and internal validity.