Final Coursework

Introduction to Quantitative Methods (PUBL0055)

Instructions

- The coursework will be posted on Moodle on 18th December 2020 at 6pm, and is due on 11th January 2021 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 75% 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 how to complete the essay. If you believe that there may be an error in the questions posed, please email both b.lauderdale@ucl.ac.uk and j.blumenau@ucl.ac.uk to request clarification/correction.
- Along with the essay questions, the provided datasets for the essay can be found in the PUBL0055 page on Moodle.
- The final essay should be submitted via the 'Turnitin Submission: PUBL0055 Essay 2' 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, each based on a different data set. There are five total questions across those three sections. Each question will receive equal weight in your final exam mark.
- Unless otherwise stated, 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 3000 words. This does not include the code, or any words (or numbers) contained within tables or figures.
- Please submit your type-written (numbered) answers in a single document (as a .pdf, .doc, or .docx file). You should create an appendix section at the end which contains all the R code needed to reproduce your results.
- You may assume the 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 explain how they apply to answering the question.
- Round all numbers to two digits after the decimal point.
- Do not screenshot or copy and paste brute R output (e.g. `lm(y ~ x)`) into your answers. Create a formatted table that is easy to read. The one exception to this rule is that you may use `screenreg()` to format a table of regression coefficients.
- 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.

Section 1 - Call to Military Service and Political Attitudes

How does being called to military service affect attitudes and political behaviour? This question has received a lot of attention in political science, and particularly so in the United States, where, until 1973, many men were enrolled into the armed forces on a mandatory basis. In a recent paper, Green, Davenport, and Kolby (2019) examine whether the draft had long-term consequences for the political behaviour and attitudes of those men eligible for the draft. The Vietnam draft randomly selected men who turned 19 prior to 1969 to serve in the US army. Although not all drafted men eventually served, we can investigate whether *being drafted* has long-term effects on attitudes and political behaviour. In particular, in this section, we will focus on the effects of being drafted on *trust* in other people.

The replication data from this study is stored in the `vietnam_draft.Rdata` file, and contains the following variables:

Name	Description
`Drafted`	Whether a respondent was drafted (1) or not (0)
`Wave`	The wave in which the respondent participated in the survey (1-4)
`State`	Respondent's state of residence at the time of the survey (5 unique states)
`date`	Respondent's date of birth (mm/dd/yyyy)
`birthyear`	Respondent's year of birth (1950, 1951 or 1952)
`dem`	Whether the respondent is a registered Democrat (1) or not (0)
`rep`	Whether the respondent is a registered Republican (1) or not (0)
`rural`	Whether the respondent lived in a rural (1) or urban (0) zip code at the time of the survey
`CollegeDegree`	Whether the respondent graduated from college (1) or not (0)
`RaceEthnicity`	`White` if the respondent is non-hispanic white, `Black` if African American, `Hispanic` if respondent is Hispanic
`Vote12Obama`	`TRUE` if the respondent voted for Obama in the 2012 Presidential election, `FALSE` otherwise
`AfghanistanMistake`	`TRUE` if the respondent thought that the invasion of Afghanistan was a mistake, `FALSE` otherwise
`IraqMistake`	`TRUE` if the respondent thought that the invasion of Iraq was a mistake, `FALSE` otherwise
`AttendedPolEvent`	`TRUE` if the respondent attended a political party event in the past two years, `FALSE` otherwise
`trust`	Respondent's general level of trust ranging from "you can never trust other people" (1) to "you can always trust people" (5)
`ideology`	Respondent's political ideology ranging from very conservative (1) to very liberal (5)

After downloading the data and storing it in the relevant folder, you can load this data into R using the following command:

```
load("data/vietnam_draft.Rdata")
```

Once you have successfully run this command, it will be available as a R data frame called `draft`.

Question 1

A. Calculate the difference in mean `trust` between drafted and non-drafted respondents. Provide a brief interpretation.

B. Explain the concept of a "sampling distribution". What is the shape of the sampling distribution for the difference in mean trust between drafted and non-drafted respondents? How is the concept of the sampling distribution relevant and useful for analysing these survey data?

C. Calculate the standard error and the 95% confidence interval for the difference in means using the formulae that we covered (i.e. do not just use the t.test function). Explain what the standard error and confidence interval are in general and what they tell us in this instance.

D. Now estimate the same effect by using linear regression. Present the result in a table, interpret the regression coefficient and compare these estimates to those obtained previously. Can the coefficient of `Drafted` be interpreted causally in this model? Explain why or why not.

Question 2

A. Fit a linear regression model that includes covariates `RaceEthnicity` and `birthyear` in addition to the treatment variable. Make a table that includes this model and the model from Question 1 part D above. What is the estimated average effect of being called to the military? Are these results different from what you obtained in the previous question? If so, why; if not, why not? How should we make the decision about whether to include control variables in this regression?

B. In the multiple regression model that you fit in part A, fully interpret the magnitude and statistical significance of the coefficient for the level `Black` of the `RaceEthnicity` variable.

C. Select two further outcome variables from `Vote12Obama`, `AfghanistanMistake`, `IraqMistake`, `AttendedPolEvent`, and `ideology` that might tell you something about the long-term effects of being drafted on attitudes or political behaviour. Estimate the average treatment effect of being drafted on these outcomes. (For the purpose of this question, you can run linear regression models with a binary dependent variable.) Produce a graph or a table to present your results. Write a short report summarising your findings.

Question 3

Use a multiple linear regression model to investigate whether the effects of being drafted on levels of `trust` are *different* for the different race/ethnicity groups defined in the variable `RaceEthnicity` (`White`, `Black` and `Hispanic`). Write a short essay describing what you find, including an assessment of the strength of the evidence for any differences (or lack thereof) that you find.

In writing this essay, you might want to:

- Provide relevant descriptive statistics for the key variables
- Explain how you set up the model(s) to make this assessment
- Present your regression estimates in a table along with the other models for `trust` that you fit in Questions 1D and 2A.
- Construct fitted values for relevant cases to illustrate the relationships described by the model.

Please note that you need not present all of these, you should aim to clearly describe the data and what they can tell us about differences in the long-run effects of being drafted for individuals in the three race/ethnicity categories in the data set.

Section 2 - The Effect of Municipality Size

How does the size of a political system affect the cost of running it? On the one hand, smaller units might be able to deliver targeted services that keep costs low. On the other hand, larger political units may have more cost-effective service delivery, because they can benefit from economies of scale.

The effect of the size of municipalities — as one example of political systems — on the cost of public service delivery is difficult to assess, because smaller and larger municipalities will differ on many dimensions, not just with regard to their size. One way to deal with potential confounding is to use data from a municipality reform that took place in a small country in 2007 in which many municipalities were quite suddenly merged. We will use the 2007 municipality reform to explore whether mergers affect per capita public service delivery cost. Since the reform took place in 2007, you will consider the post-treatment period to start in 2008.

The CSV file `cost_data.csv` contains the variables presented in the table below. In this data, each row is a municipality-year observation. Municipalities that were merged in 2007 are recorded as one municipality not just after the merger, but also before. For the municipalities that were merged in 2007 (treated), the pre-treatment outcomes are based on the average of the outcomes of the municipalities that were later merged. Note that before the 2007 reform, the number of control and future treated municipalities was similar.

The outcome of interest is public service delivery costs per capita called `Y`. Whether a municipality is merged (treated) or not (control) is indicated as 1 or 0, respectively, in the variable called `treatment`. The data also contains region and municipality identifiers:

Name	Description
`year`	Year (2005-2011: 2005-2007 pre-treatment, 2008-2011 post-treatment)
`Y`	The outcome variable. Public service delivery costs per capita.
`treatment`	Whether the municipality is the result of a 2007 merger (1) or not (0).
`region`	Region ID (4 unique regions)
`municipality`	Municipality ID (87 unique municipalities)

Once you have downloaded the data file and placed it in the relevant folder, it can be loaded into R as follows:

```
costs <- read.csv("data/cost_data.csv")
```

Question 4

A. Calculate, compare and interpret the yearly outcomes by treatment group. What does this tell you?

B. Using the data from the years 2007 (the last year of the pre-treatment period) and 2008 (the first year of the post-treatment period), compute the difference-in-differences estimate of the effect of municipal merger on public service delivery costs.

C. Use an appropriate linear regression to calculate the same difference in differences estimate that you calculated in part B. Show that you get the same numerical value, and use the results of the regression to test the null hypothesis that the difference in differences is equal to zero. Interpret your results in terms of the research question posed at the beginning of this section.

Section 3 - Direct Democracy and Naturalisation

In Switzerland there is a tradition of using direct democracy (referenda) for a wider range of decisions than in most democractic countries. Citizens are called to vote on referendums four to five times a year, and, until 2003, many municipalities decided on immigrants' citizenship applications in secret ballot votes. Even though applications would only get to the vote stage if they fulfill all naturalisation requirements, voters rejected many immigrants' applications at the ballot box. (The requirements include the following: being integrated, respecting the legal order, posing no threat to the security of Switzerland, self-sufficiency, and, depending on the region, mastering the local language.)

In this section, you will investigate why some applicants were rejected while others were not. The data includes the vote results from 2429 applications, and a number of variables with information about the applicants that stems from official leaflets that were sent to voters before the secret ballot vote:

Name	Description
`born_CH`	Whether applicant is born in Switzerland (1) or not (0)
`since`	Number of years since arrival in Switzerland
`male`	Whether applicant is male (1) or female (0)
`kids`	Whether applicant has children (1) or not (0)
`unemployed`	Whether applicant is unemployed (1) or not (0)
`refugee`	Whether applicant has refugee status (1) or not (0)
`married`	Whether applicant is married (1) or not (0)
`percent_novotes_rounded`	Percentage 'no' votes in application referendum
`origin_region`	Applicant's region of origin
`application_decade`	Decade in which application was voted on
`age`	Applicant age (ranging from "0-20 years" to "60+ years")
`education`	Applicant education level (low, middle, high)
`local_language`	Applicant local language skills (ranging from "insufficient" to "perfect")
`integration_level`	Applicant integration level (ranging from "familiar with Swiss customs and traditions" to "indistinguishable from Swiss")
`skill_level`	Applicant skill level (low, middle, high)

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

```
pass <- read.csv("data/passport_data.csv")
```

Question 5

Your task in this section is to investigate the relationship between the share of No votes that an applicant received and the applicant's characteristics. In particular, we are interested in whether applicants are discriminated against based on their region of origin. To explore this question, you should implement two linear regression models with `percent_novotes_rounded` as the dependent variable.

In the first model, the only explanatory variable should be the `origin_region` variable. For the second model, you should build a model which – in addition to the `origin_region` variable – includes exactly **three** additional explanatory variables that you think might be useful to include from the supplied dataset. You should explain why you think these particular variables are important to include, given that our main interest is in the relationship between region of origin and application no votes. Please note that, for the second model, you should *not* estimate several different models and present the results, but rather you should argue theoretically why you chose certain variables.

You should write up the results of these models as if they were to be published in a political science journal article with a focus on communicating the substantive meaning of your results. In your discussion of these

models, you should focus on communicating the substantive implications of the regression that you implement, paying particular attention to the relationship between an applicant's country of origin and the outcome of his or her citizenship application. You may wish to focus on the following:

- Provide descriptive statistics and/or plots to provide the reader with an overview of the dependent variable and the important explanatory variable(s) that you intend to use.
- Provide a well-formatted table of regression output which includes the key information about the models you have estimated.
- Discuss both the statistical and substantive significance of the relationships that you illustrate.
- Discuss model fit, using appropriate statistics.
- Discuss whether or not or under which conditions we should consider the estimates you present to be causal effects of region of origin.
- Discuss weaknesses of your analysis, and potential alternative analysis designs that you might use (given different data) to evaluate this research question.