# Final Take-Home Paper

## PUBL0055 Introduction to Quantitative Methods

2023-24

## Instructions

### Submission Formalities

- The final assessment is a 7-day Take Home Paper and is posted on Moodle on 3rd January 2024 at 2pm, and is due on 10th January 2023 at 2pm. Please follow all designated SPP submission guidelines for online submission as detailed on the PUBL0055 Moodle page.
  - The standard extension for SoRAs with exam arrangements and ECs are 14h only (see sections 6.3.7 and 6.3.9 of the Regulations)
  - Late submission penalties for take-home papers apply (see section 6.3.11 of the Regulations)
  - If you experience technical issues, you must inform the PS team on polsci-pg@ucl.ac.uk ASAP with evidence.
- The coursework should be submitted via the 'PUBL0055 - Assessment 2 - Take Home Paper [7 Days] (70%)' links 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!
  - Remember that you can only submit once. You can check your similarity score before submission via the Turnitin Similarity Checker.
- This is an assessed piece of coursework (worth 70% 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 teaching team questions about the coursework.
- Along with the coursework questions, the necessary datasets for the coursework can be found on the PUBL0055 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. Standard word limit penalties apply.
  - When answering, **do not include the question text in your submission**, as these would go towards your word count.
  - The word count must be clearly indicated on the title page of the assessment submission. **If not included, the turnitin word count (which counts R code) will be used to decide on word penalties.**
- The coursework consists of two separate sections, each with several questions. The marks allocated for each section are indicated in the text.
  - Together, the questions are worth 90 marks.
  - 10 marks are reserved for clarity of presentation (see below).

- Please submit your type-written (numbered) answers in a single document, **preferably a pdf file, as the formatting will not be affected when we look at it on TurnitIn for marking**, 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 one document.

- 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.

  - If a question only asks you to create a table, a figure, or only to subset data, without explicitly asking you to report a number or interpret it, this means the question does not require an answer sentence.

- You can integrate the code into your document (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 must run correctly when we run it. **You do not need to include the code that failed to run**, but just the final, 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 raw 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 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 be able to explain what they do and how they apply to answering the question.

- Unless explicitly stated which one you should use, you can use *any method you see fit* to get the answer to a question. For example, if asked to calculate the difference in means, with no further specification, you can choose to do this by hand or with bivariate regression.

## Presentation (10 marks)

Marks will be *deducted* for bad presentation, which includes:

- 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

- 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

# Section 1 (40 marks)

**The Effects of Educational Television**

Is educational television an effective teaching aid? "The Electric Company" was a television programme that ran on US TV from 1971 to 1977. The programme used sketch comedy to provide an entertaining way of helping elementary school children develop their grammar and reading skills. It was widely credited by many teachers in US schools as having important effects on [...] cond-, third-, and fourth-grade children. In this section, you will analyse data from an experime[...] mly assigning classes of children to watch "The Electric Company". You will investigate what rea[...] made classes as part of this experiment.

The unit of analysis in this [...] ren, and there are 192 classes in the data. Each class was either treated (to watch the progra[...] tch the program). The outcome of interest is the average score on a reading test administered a[...] called `post.score`. In addition to the treatment and outcome, the data also contains informati[...] the class and the score on the same reading test as administered before the treatment took place.

| Name | Description |
|------|-------------|
| `class_id` | A numerical identifier for each class |
| `grade` | The school grade of the class (1st through 4th) |
| `treatment` | 1 if the class was treated, 0 otherwise (randomized) |
| `pre.score` | Class reading score *before* treatment, at the beginning of the school year |
| `post.score` | Class reading score at the end of the school year |

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

```
electric <- read.csv("data/electric-company.csv")
```

## Question 1 (8 marks)

a. Calculate the number of classes in each grade. Is the number of classes in each grade the same? For which grade are there the most observations? **(2 mark)**

b. Make a box plot which depicts student scores at the beginning of the year as a function of the grade they are in. Which grade had the highest median reading score at the beginning of the year? Which grade had the most dispersed scores at the beginning of the year? **(3 marks)**

c. Create a scatter plot which compares student scores at the beginning of the year to student scores at the end of the year. Briefly comment on what you see. **(3 marks)**

## Question 2 (12 marks)

a. Calculate and report the average effect of the treatment on the class reading score at the end of the school year. **(1 mark)**

b. Explain whether we can interpret your answer to a. as the causal effect of television on student scores. **(2 marks)**

c. Calculate and report the standard error of the difference in means (you should do this 'by hand'). Then, calculate and report the t-value for the difference in means. Considering the t-value, can we reject the null hypothesis of no effect of the treatment at the 95% and 99% confidence levels, respectively? **(3 marks)**

d. Construct and interpret the 95% confidence interval for the difference in means estimate. **(2 marks)**

e. Briefly explain the concept of a "sampling distribution". To illustrate this, create a function that samples 200 observations from the full data set and then calculates the difference in means between the treatment and control group. Replicate this 5000 times and then plot the resulting sampling distribution in a histogram. Briefly comment on what you see. **(4 marks)**

**Question 3 (10 marks)**

a. Estimate three linear regression models. The first should predict `post.score` with only the `treatment` variable. The second regression model should be the same as the first, but should also control for student `grade`. The third model should be the same as the second, but should also control for `pre.score`. Report the results of all three models in one regression table. **(2 marks)**

b. Compare these models [...] of the variation in `post.score` they "explain". What does this tell us about the relati[...] grade a student is in and reading ability, and 2) students' prior performance on the tes[...]nce on the test? **(4 marks)**

c. Are the estimates of t[...]t different across the three models? Why do you think that is? Support your argumen[...]sults from relevant tests of pre-treatment covariate balance. **(4 marks)**

**Question 4 (10 marks)**

a. Run a linear regression model with `post.score` as dependent variable and `treatment`, `grade` as well as their interaction as independent variables. Report the results in a table. **(1 mark)**

b. What is the meaning of the interaction coefficients? **(3 marks)**

c. Based on the model from a., calculate and report the estimated treatment coefficients for *each* grade. How does the effect of the treatment differ as grade increases? (**4 marks**)

d. Using the model from a., calculate and report the fitted values for treated and untreated 2nd graders. Interpret these values in substantive terms. (**2 marks**)

## Section 2 (50 marks)

### Removal of Lenin Statues, Electoral Turnout & Pro-Russian Backlash

What are the electoral consequences of changes to symbols of nationhood and a nation's past? In a recent article, Rozenas & Vlasenko investigate the effect of the demolition of statues of Lenin during the 2013-14 Euromaidan protests, the so-called "Leninopad" (meaning: "Leninfall" or "Lenin's free fall"), on overall electoral turnout and the pro-Russian voter share.

Since gaining independence [...] ere had been some, albeit limited, efforts to 'de-Sovietise' public spaces and symbols of natio[...]. In 2013, then-President Viktor Yanukovich relatively suddenly decided to build closer ties wi[...] the European Union. This sparked nationwide protests against the government starting in Nove[...] maidan"[1], which also led to an unprecedented wave of demolitions of Soviet-era Lenin monume[...]

In this section, you will use [...] he data Rozenas & Vlasenko used in their paper. The dataset is entitled `leninopad.rda` and you can find it on the PUBL0055 Moodle page. The data includes the following variables:

| Variable name | Description |
|---|---|
| precinct_id | Unique numerical indicator for each electoral precinct. |
| election | Year of election. Includes elections in 2004, 2006, 2007, 2010, 2012, 2014. Note that the elections in 2004, 2010 and 2014 are all presidential elections, whereas those held in 2006, 2007 and 2012 were all parliamentary elections. |
| post_leninopad | Dummy variable indicating treatment *period*, i.e. whether the election happened *after* the outbreak of Euromaidan in 2013 and the start of the demolition of Lenin statues (=1), or *before* (=0). |
| number_statues | Total number of Lenin statues present in the electoral precinct. Is equal to `NA` for all precincts where no Lenin statue was present before the start of the period of analysis. |
| ever_removed | Dummy variable indicating treatment *group* status, i.e. whether the electoral precinct is one of those who had at least one statue removed (=1), or not (=0). |
| statue_removed | Dummy variable indicating *treatment*, i.e. whether at least one statue was removed in the electoral precinct **and** the election occurred *after* Leninopad (=1), or not (=0). |
| overall_turnout | Share of registered voters that voted, measured in percent (%). |
| pro_russian_turnout | Share of registered voters that voted for pro-Russian candidates, measured in percent (%). |
| precinct_size | Official classification of precincts by electorate size. Factor variable with 3 levels: small, intermediate, large. |
| dist_to_kiev | Distance to Kiev (where the first Lenin statue 'fell') in km. |
| km_road | Length of the roads that are within the precinct. Can be understood as a proxy for urbanisation. |

You should load the data by using the following command:

```
load("leninopad.rda")
```

---

[1]The Euromaidan protests eventually led to the Maidan Revolution and Yanukovich secretly fleeing Kiev to end up in Moscow in February 2014.

## Question 1 (10 marks)

Let's begin by preparing the data for analysis and describing it.

a. How many *unique* electoral precincts are there in the dataset? **(1 mark)**

b. As we are interested in the effect of *removing* a Lenin statue on electoral outcomes, we need to remove all those precincts which did not have any Lenin statues before the start of Leninopad. Create a subset of the data removing all those observations [where] a Lenin statue was present before the start of the period of analysis, i.e. where the number of [statues] [is larger] [than 0]. [Ho]w many unique electoral precincts are there in the subsetted data? **(2 marks)**

*You will be using this sub[set]* [of the data] *for the remainder of this section, so you can call this new data* leninopad *as well.*

c. What is the proportio[n of treated] precincts, i.e. those who had at least one statue removed during Leninopad? **(2 marks)**

d. Let's have a look at the distribution of the number of Lenin statues still present in the last election *before* Leninopad. What are the shares of electoral precincts with 1, 2, 3 and 4 statues, respectively, still standing during the 2012 election? Report the values as *percentages* rather than proportions. **(2 marks)**

e. Create two histograms, one for the distance from Kiev and one for the kilometers of roads present in the electoral precinct. Do either of these variables need to be log-transformed? Why or why not? Create a new, log-transformed, version of the variable(s) for which you concluded that it made sense to do so. **(3 marks)**

## Question 2 (15 marks)

One way we could investigate the relationship between statue removal and electoral outcomes is to compare a cross-section of the treated and untreated observations after the treatment occurred. Create a new dataset by subsetting the data from 1.b to only the observations from 2014 using the following code:

```
leninopad14 <- leninopad[leninopad$election==2014,]
```

a. With this subset, calculate and report in substantive terms the difference in mean turnout between those electoral precincts where a statue was removed and those where no statue was removed. **(2 marks)**

b. Is this difference in means you just calculated a credible estimate of the causal effect of removal of statues on turnout? If yes, why; if not, why not? **(3 marks)**

c. With this subset, fit a multiple regression model with turnout as dependent variable, and, as independent variables, treatment status, size of the electoral precinct, as well as the distance to Kiev and kilometers of roads. *If you decided to log one (or both) of the latter two variables in question 1.e., you should use the logged version(s) of the variables you created.* Present the regression results in a table. **(2 marks)**

d. Interpret the intercept in substantive terms. Is the intercept meaningful in this instance? **(3 marks)**

e. Interpret the coefficient for treatment status in substantive terms. How does it compare to your results from b.? **(3 marks)**

f. Under which circumstances can the estimated average treatment effect from a selection on observables design be interpreted as *causal*? Are these conditions met in the multivariate regression from c.? **(2 marks)**

## Question 3 (12 marks)

As we have access to data about the outcome variable *before* Leninopad started for both the treated and control precincts, we can adopt a difference-in-differences approach.

a. Subset the dataset you created in 1.b to only include the observations from the election years just before (2012) and just after (2014) treatment occurred. **(1 mark)**

b. Calculate and report the difference-in-differences estimate for overall turnout. Do the calculation 'by hand', i.e. by calculating the relevant differences between the relevant group means. Interpret it in substantive terms, making sure to pay attention to the size of the effect. **(3 marks)**

c. Calculate the difference-in-differences using linear regression with an interaction between treatment *group* and treatment *period*. Be mindful to choose the correct variables to do this. Report and compare the difference-in-differences estimate you find to the one from b. **(3 marks)**

d. Is the difference-in-differences estimate statistically significant? To answer this, conduct a hypothesis test with $\alpha$-level $= 0.05$ on the coefficient from the interaction you got from c. by taking the following steps. You can find all the values you need by running summary() on the model you created in c. **(5 marks)**

   1. State the null hyp...
   2. Report the t-valu... relevant critical value.
   3. Report the p-valu... ...ntive terms and compare it to the chosen $\alpha$-level.
   4. State a conclusion... ...esis.

## Question 4 (8 marks)

We also have information on... ...voters who voted for pro-Russian candidates, which allows us to investigate the mechanism behind the effect of statue removals in more detail. For this question, you should use the same data from only 2012 and 2014 you used in the previous question.

a. Calculate the difference-in-differences estimate with linear regression like you did in 3.c., but this time with the *share of pro-Russian voters* as dependent variable. Present this regression together with the regression from 3.c. in the same table. **(2 marks)**

b. Interpret the difference-in-differences estimate in the pro-Russian vote model in substantive terms. **(3 marks)**

c. Compare the estimated treatment effect of Lenin statue removal on overall turnout to the one on pro-Russian turnout. What can we conclude about the driving factors behind the effect of Lenin statue removal on overall turnout from this? **(3 marks)**

## Question 5 (5 marks)

a. On what assumption does a difference-in-differences analysis rely on to make credible conclusions about *causal* effects? What does it imply in the case of the removal of Lenin statues and overall electoral turnout in Ukraine? **(2 marks)**

b. Assess the plausibility of the assumption by plotting the trends in electoral turnout in the treatment and control group, respectively, over all of the elections starting from 2004. For this question, you need to again use the data you created in question 1.b. Describe what you see. Is the assumption reasonable in this case? **(3 marks)**