

Project Instructions

For this project, you will be required to use data from one of the following articles:

Angrist, J. and V. Lavy (2009). [The effects of high stakes high school achievement awards: Evidence from a randomized trial](#). *American Economic Review* 99(4), 1384 – 1414. [Data]

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015). [The miracle of microfinance? Evidence from a randomized evaluation](#). *American Economic Journal: Applied Economics* 7(1), 22 – 53. [Data]

Banerji, R. J. Berry, and M. Shotland (2017). [The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in India](#). *American Economic Journal: Applied Economics* 9(4), 303 – 337. [Data]

Bertrand, M., and S. Mullainathan (2004). [Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination](#). *American Economic Review* 94(4), 991 – 1013. [Data]

Chong, A. I. Cohen, E. Field, E. Nakasone, and M. Torero (2016). [Iron deficiency and school-
ing attainment in Peru](#). *American Economic Journal: Applied Economics* 8(4), 222 – 255. [Data]

Gneezy, U., J. List, J. Livingston, X. Qin, S. Sadoff, and Y. Xu (2019). [Measuring success in education: The role of effort on the test itself](#). *American Economic Review: Insights* 1(3), 291 – 308. [Data]

Muralidharan, K. and V. Sundararaman (2011). [Teacher performance pay: Experimental evidence from India](#). *Journal of Political Economy* 119(1), 39 – 77. [Data]

You must let me know (via email) which of these articles you are interested in no later than 8am on January 25; if I do not hear from you by then, I will make the choice for you.

As noted on the syllabus, you are not being asked to replicate the article that you are getting your data from. Instead, once you let me know which of these articles you are interested in, I will suggest a slight variation on it for you to do (e.g., if the original article analyzed performance for all students, I might suggest that you focus only on the performance of boys).

Ultimately, your aim in this project is to answer a *causal* (not “casual”) question such as the following: Does being placed into a small class cause students to perform better academically? To do so, you will be required to use a regression model of the following form:

$$\text{Outcome}_i = \alpha + \beta \text{Treatment}_i + \mathbf{X}_i \gamma + U_i,$$

where Outcome_i is the outcome (e.g., a test score) for the i th individual, Treatment_i is equal to 1 if the i th individual receives the treatment (e.g., being placed into a small class) and 0 otherwise, \mathbf{X}_i is a *vector* of control variables (e.g., age, gender, etc.) for the i th individual, and U_i is an idiosyncratic error term.

The main parameter of interest is β , which is known as the “average treatment effect” or ATE (no one really cares what α or γ are). Thus, the null hypothesis you will want to test is $H_0 : \beta = 0$. If any of this is unclear to you, please make sure to spend some time watching my videos that review the background material you are expected to be familiar with from your previous courses in statistics/econometrics.

Submission Instructions

The project will be completed through 4 “instalments” each worth 20% of your final grade (the other 20% is the test). You should think of these instalments not as 4 separate pieces of work, but rather 4 versions of the same piece of work, each one being “better” than the one that came before it. That is, each instalment should not only add new features, but also improve the *existing* features (this means fixing any technical errors you had previously, making your writing more clear, etc.).

Instalment submissions must be made via a private Google Drive folder that I will share with you once you let me know which article you are interested in. Specifically, each instalment will require you to upload files named `paper-x.pdf` and `code-x.txt`, where `x` is the instalment number (e.g., your first instalment will require files named `paper-1.pdf` and `code-1.txt`). The file named `paper-x.pdf` is to be a PDF file containing the latest version of the write-up for your project. The file named `code-x.txt` is to be a plain text file containing the latest version of your R code. For the first instalment, you will also need to upload your data file (do not modify this file in any way, i.e., do not rename it or convert it to a different format). All of these files must be contained entirely within the Google Drive folder I share with you; please do not create any subfolders within this folder! If you have done everything correctly, you will have uploaded exactly 9 files in this folder by the end of the course (2 for each instalment plus 1 containing your data).

Please note that failure to precisely follow the above submission guidelines will result in a mark of zero. For example, if you were to upload your write-up as a Word file rather than a PDF file, or your R code as a rich text file rather than a plain text file, you would get a zero.¹

Your R Code

The single most important thing to keep in mind about this project is that I need to be able to replicate all of your results. To run your code, I will set my working directory to the Google Drive folder I have shared with you and enter the following command:

```
source("code-x.txt")
```

(where `x` is the instalment number). Your code needs to be written so that the above produces every single number that appears in your write-up. If this doesn't work for any reason, you will get a mark of zero. Thus, I recommend that you work in exactly the same fashion yourself rather than “interactively” (i.e., typing commands directly into the R console). Indeed, before submitting any instalment, you should re-start the R console and run the above command to make sure you get the results you are expecting.

For the purpose of this project, you are not permitted to use any R packages except for `haven` (for reading data saved in Stata format) and `sandwich` (for computing HC standard errors).

Below are some general guidelines for your R code. If you fail to follow of any of these guidelines, you will get a mark of zero.

- Do use the following as your very first line to ensure R's memory is cleared: `rm(list=ls())`
- Do not include any line beginning with `>` (i.e., lines that you copied from the R console).
- Do not include any calls to the `setwd()` function.
- Do not include any “path” references when reading in your data. That is, you should have something like `read.table("data.txt")` rather than `read.table("/Users/JaneDoe/ECN723/data.txt")`, and just manually set the working directory in R to the location where you've saved your data file (remember: when I run your code, I will set my working directory to the Google Drive folder containing I have shared with you, i.e., the folder containing your data file).
- Do not include any calls to the `install.packages()` function or the `remove.packages()` function. However, do make sure to include a call to the `library()` function for any package(s) you use.
- Do not include any calls to functions that open a graphical interface such as the `View()` function (you can use this yourself if you would like, but it will just create an error for me).
- Do not create separate data frames for your treated and non-treated groups. You should have a single data frame containing all of your observations, and within this data frame, there should be a treatment variable equal to 1 for observations in the treated group and 0 for observations in the non-treated group.
- Use the `attach()` function exactly once (and make sure to do so only after you have “cleaned” your data).

¹Please make sure you understand the difference between “plain text” and “rich text”.

Your Write-up

You will need to create a short write-up describing precisely what you have done/found. It must be no more than 10 pages in length, but all else equal, shorter is better (clearly explain everything you are doing in detail, but keep it concise).

Your write-up should be written so that it would be easy for another student in this course to read it and understand exactly what you have done/found. That is, your “target audience” consists of readers who know something about economics and econometrics, but don’t necessarily know anything about the *specific* topic you are writing about (do not assume that your readers have read the article that you obtained your data from). This means that you can skip explaining straightforward things like how to calculate a T -statistic and put all of your energy into explaining the design of your experiment, what all of your variables measure, and what your results tell you about your causal question of interest.

Your write-up must be split into 3 sections:

1. Introduction

This section should very clearly explain what your causal question of interest is, and how your experiment is designed. Make sure to explain exactly what your “treatment” is. This section should be 1 to 2 pages in length.

2. Data and Model

This section should provide a very clear explanation of the model you are estimating and how all the different variables in it are defined. *Be very specific.* For example, if your outcome variable is “TestScore” you need to explain exactly what this is measuring, i.e., what kind of test it is, when the test took place, what the score is out of, etc. Information about your outcome, treatment, and control variables should be summarized in a table (call it Table 1; see `ecn723-project-sample.pdf` for an example).

You should also include a table here providing the sample mean (and its standard error) of all of these variables for the entire sample and also for each group (treated and non-treated) separately; this table should also clearly list the total number of observations as well as the number of observations in each group (call it Table 2; see `ecn723-project-sample.pdf` for an example).

In addition to your “full” regression model that includes all of your variables, you will also be required to estimate a “basic” version of it that does not include your control variables, i.e., a model of the following form:

$$\text{Outcome}_i = \alpha + \beta \text{Treatment}_i + U_i.$$

Rather than writing out equations for both models, however, just write out the equation for your full model and then explain in words that your basic model is identical but excludes the control variables.

You do not need to go into any of the details about your econometric methods, but you should clearly state what methods you are using. For example, you might tell us that you are estimating the parameters in your model using OLS and that you are providing us with HC standard error for them. Finally, make sure to clearly describe exactly what hypothesis you will be testing and how this relates back to your causal question of interest.

Overall, this section should be 3 to 4 pages in length.

3. Results

This section should clearly describe your results. You should have a table here showing your average treatment effect estimates (and their standard errors) from your basic and full models (call it Table 3; see `ecn723-project-sample.pdf` for an example). Remember that no one cares about the estimates of α or γ ; all that we care about is your estimate of β (the ATE). Most importantly, you need to formally test the hypothesis you described in Section 2 (do this using the results from both your basic model and your full model, but base your overall conclusion on the full model as it should provide a more accurate estimate of the average treatment effect). This section should be about 2 pages in length.

Your write-up does not need a “Conclusions” section or any appendices (remember that I have your R code, so there is no need to include it in your write-up). You only need the 3 sections (and 3 tables) described above; no more, no less.

In addition to this outline, you must adhere to the following formatting guidelines:

- Use 1 inch margins on all sides, and number each page inside the bottom margin (centered).
- Use “justified” alignment for all paragraphs (i.e., text stretched out from the left margin to the right margin).
- Double-space everything (except footnotes and notes for tables, which should be single-spaced). However, do not include an extra space between sections (i.e., there should be exactly one line between the last word of a section and the title of the next section, not two or three).
- Use a 12 pt font size for everything (except footnotes and notes for tables, which should be 10 pt).
- Do not include a title page. The first line of text should be your main title (centered and in bold), the second line of text should be your name (centered), and the third line of text should be title of the first section (left-justified and in bold), and so on.
- Do not indent the first line of the first paragraph of a section, but do indent the first line of each subsequent paragraph.
- Use bold for your main title, the number/title of each section, and the title of each table, but nowhere else.
- Use footnotes rather than endnotes.
- Do not paste any R code or output into your write-up.
- Tables should only contain horizontal lines, and these horizontal lines should only be at the top of the table, after the header row, and at the bottom of the table.
- Above each table, you must write “**Table X: Blah blah blah**” (without the quotation marks) where “X” is the table number and “Blah blah blah” is the description.
- Always refer to tables by writing “Table X” (without the quotation marks) where X is the table number (notice that Table is capitalized). For example, you might write “... are shown in Table 1”.
- You do not need a “References” section since you are only going to cite one paper (the paper you found your data from). Instead, include a full reference to this paper in a footnote the first time you mention it, and always refer to it as “Lastname1 and Lastname2 (year)” (if there are two authors) or “Lastname1 et al. (year)” (if there are 3 or more authors). For example, you might write something like “Angrist and Lavy (2009) estimate...” or “Banerjee et al. (2015) examine...”. Do not ever write first names, article titles, or journal names in the main body of text.

All of these formatting rules are demonstrated in the file named `ecn723-project-sample.pdf`. Please read it very closely. If you don’t follow these formatting guidelines, I will just stop reading and give you a zero.

Timeline

Again, you must let me know (via email) the paper that you would like to get your data from no later than January 25 at 8am; if you fail to do so, I will make the choice for you. The important point is that you need to start working on your first instalment absolutely no later than January 25; if you think you can do it on the weekend of February 13/14, you are setting yourself up for disaster (you will probably find the first instalment to be the most difficult one since it will require you to get your data imported into R and “cleaned up”; indeed, each instalment should be progressively easier for you).

For each instalment, there are a set of *minimum* tasks that you need to achieve:

Instalment	Due (8am)	Minimum Tasks
1	February 15	<p>-Create the basic layout of your write-up and ensure you have it formatted properly. If you do not follow the formatting guidelines on this or any other instalment, you will get a mark of zero.</p> <p>-Write the entirety of Section 1.</p> <p>-In Section 2, give a detailed description of what your outcome/treatment/control variables are and complete Table 1.</p> <p>-After getting rid of any observations with NA values (for any of your variables), compute the number of observations you have in the entire sample and in both the treated and non-treated groups (you will know you are on the right path if the total number in these two groups is equal to the number in the entire sample; you will get a mark of zero if this is not the case). Fill these values into the bottom row of Table 2.</p> <p>-Make sure that your data file is uploaded into the Google Drive folder I share with you.</p>
2	March 8	<p>-Compute your summary statistics and complete Table 2. To check that you are on the right path, make sure that (a) the sample mean of your treatment variable for the entire sample is equal to the number of observations in the treated group divided by the number of observations in the entire sample, and (b) for every variable, the sample mean for the entire sample lies somewhere between the sample mean for the treated group and the sample mean for the non-treated group. If either of these conditions is not satisfied, you will get mark of zero.</p> <p>-Specify your regression model and describe the hypothesis you will be testing in order to complete Section 2 (this should come after Table 2).</p> <p>-Read over Section 1 again and spend some time to improve your writing (please don't think it is already "perfect"; your writing can always be improved). <u>Do not neglect this step!</u></p>
3	March 29	<p>-Use OLS to estimate your basic model and fill in the first column of Table 3. To check that you are on the right path, use the numbers in Table 2 to compute the two-sample T-statistic for comparing the mean of the outcome variable between the treated and non-treated groups (you can just do this by hand to check for yourself; do <u>not</u> use the <code>t.test()</code> function in R as it makes some silly assumptions); the numerator and denominator of this test statistic should be equal to the estimated coefficient on your treatment variable and its standard error, respectively (you don't need to report the value of this test statistic in your write-up; just compute it in R to check that you are on the right path). If this condition is not satisfied, you will get mark of zero.</p> <p>-Use the results in the first column of Table 3 to test the null hypothesis that the coefficient on your treatment variable is equal to zero (if you don't do this correctly, you will get a mark of zero). Discuss your findings in Section 3 right below Table 3.</p> <p>-Read over Sections 1 and 2 again and spend some time to improve your writing. Again, <u>do not neglect this step!</u></p>
4	April 19	<p>-Use OLS to estimate your full model and fill in the second column of Table 3. Use these results to again test the null hypothesis that the coefficient on your treatment variable is equal to zero (if you don't do this correctly, you will get a mark of zero). Discuss your findings in Section 3, making sure to compare it to what you found using your basic model. In case your findings differ, you should base your conclusion on the full model as it should provide a more accurate estimate of the average treatment effect. Make sure that Section 3 is very clearly written as you will not have an opportunity to revise it.</p> <p>-Read over Sections 1 and 2 again and spend some time to improve your writing.</p>

Remember that `paper-2.txt` and `code-2.txt` should be improved/expanded versions of `paper-1.txt` and `code-1.txt`, respectively, and so on. Nothing is “written in stone”; you can add/remove/modify any part of your code or write-up for any new instalment. For example, even though you will have written your introduction for the first instalment, you still need to put some effort into *improving* that section in each subsequent instalment.

Feedback

Inside the private Google Drive folder I share with you, there will be a file named `feedback.txt` that I will use to give you feedback on each instalment (this will be updated within one week of every new submission; I will make sure to indicate which instalment I am referring to so that there is no confusion). Please make sure to incorporate all of the feedback I leave into your next instalment. The absolute worst thing you can possibly do in this course is to ignore my feedback. If I start reading a new instalment and see that you have ignored the feedback I gave on your previous instalment, I will just stop reading and give you a zero.