Week 1: Introduction to Causal Inference

Tanisha Mohapatra & Fernando Sanchez Monforte

HT 2025

Overview In this week's lab session,[1] we will cover:

- Some basics of R Markdown
- Potential outcomes

We will also explore the data used in a cross-national study investigating whether predominantly Muslim societies are disadvantaged in democratization.

[1] Lab sessions are developed based on materials from past iterations of this course, generously shared by Sascha Riaz, Ria Ivandic, and Nelson A. Ruiz.

1 Introduction

1.1 Causality

- This week's lecture delved into the concepts of causality and the potential outcomes framework. It discussed causality as the relationship between two sets of events: one (the effects) being a direct consequence of the other (the causes).
- Our discussion then shifted to the logic of counterfactuals, framed in the context of "If X had or had not occurred, Y would or would not have ensued." This highlights that constructing counterfactuals is essential for making valid causal claims.
- Another key point we covered was the issue of omitted variable bias, which can occur even when some covariates are controlled for. This bias arises from our inability to control for confounders when we lack data on them.

1.2 Potential Outcomes

- We also took an in-depth look at the Potential Outcomes Framework. Central to our discussion was *The Fundamental Problem of Causal Inference*, which describes the impossibility of observing different treatment outcomes for the same unit simultaneously.
- We further examined various estimands, including the average treatment effect (ATE) and the average treatment effect on the treated (ATT).
- The lecture also addressed selection bias, specifically relating to the pre-treatment differences in outcomes between treated and control groups.
- For this week's practical session, we will revisit some of the core principles of the "potential outcomes" framework and introduce basic applications using R.

Before we proceed further:

1. Create a folder called `lab1`.
2. Download the data and the empty RMarkdown file (both available on Canvas).
3. Save both in this `lab1` folder.
4. Open the Rmd file.
5. Set `lab1` as the working directory using the `setwd()` function or by clicking on "More".

2 R Markdown Refresher

If you are struggling to remember how R works since the last time you used it (which is reasonable: coding is like foreign language acquisition, if you don't use it you lose it), then you might want to check out chapters 1-4 of this R refresher. More resources are linked at the end of this document, in Section 6.

`Markdown` is a tool for converting plain text into formatted text. `R Markdown` utilizes the markdown syntax in order to combine formatted text with code in a single document. Within RStudio, R Markdown is a specific type of file format for making dynamic documents. It allows you to simultaneously (1) save and execute code, and (2) produce high quality documents that include both code and text. For the purposes of our labs, R Markdown allows us to include code chunks and the text that helps explain them in an easy-to-read manner. You will use R Markdown to write your assignments and the final exam, so it's important that you feel comfortable using the software.

2.0.1 Headers

You can create section headers by using a `#`, for example, `# Section`. Under each of these, you can create subsections, like `## Subheader`, or sub-sub sections, like `### Sub-sub header`.

2.0.2 Emphasis

You can add emphasis by making text **bold** or *italic*. To bold text, add two asterisks or underscores before and after a word or phrase. To bold the middle of a word for emphasis, add two asterisks without spaces around the letters. Typing `I just love **bold text**` will produce: I just love **bold text**. To italicize text, add one asterisk or underscore before and after a word or phrase. To italicize the middle of a word for emphasis, add one asterisk without spaces around the letters: `*italicized text*` shows up as *italicized text* in your output file.

To emphasize text with bold and italics at the same time, add three asterisks or underscores before and after a word or phrase. To bold and italicize the middle of a word for emphasis, add three asterisks without spaces around the letters. If you want to say something `***really important***`, it'll show up as ***really important***!

2.0.3 Basic blockquotes

To create a blockquote, add a `>` in front of a paragraph.

> I added the > at the beginning of this paragraph, and got a blockquote :)

2.0.4 Multi-paragraph blockquotes

For multi-paragraph blockquotes, in between each paragraph, add a new line with just a `>`.

> Par 1: The rubber duck, with a sparkle in its eye and a causal inference textbook twice its size, was ready to dive into a world where 'cause' and 'effect' were more than just words in a dictionary.
>
> Par 2: It wasn't the brightest bulb in the chandelier, often confusing correlation with causation, but its enthusiasm could light up the entire department.

2.0.5 Figures

Insert an image

In a new line, type `![Caption for the image](path/to/image.png)`. You can also use `knitr::include_graphics()` within a code chunk, for example:

```
knitr::include_graphics("figures/rubber_duck_doing_stats.png")
```

Figure 1: A rubber duck doing statistics for me

Insert a figure from data

You can also include figures directly from analysis done within code chunks:

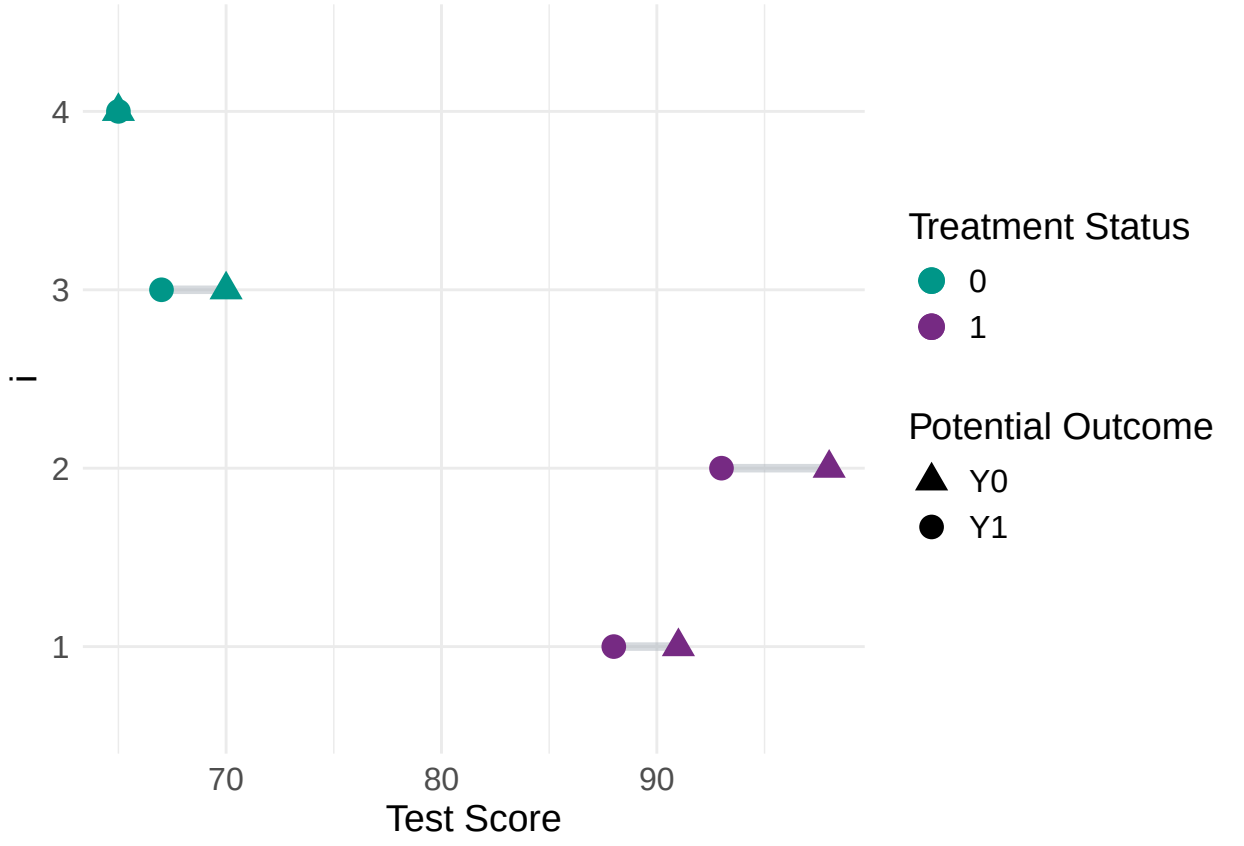
```
# Data
i <- c(1:4)
Y_i <- c(88, 93, 70, 65)
D_i <- c(1, 1, 0, 0)
Y1_i <- c(88, 93, 67, 65)
Y0_i <- c(91, 98, 70, 65)

# Create a data frame
data <- data.frame(i, Y_i, D_i, Y1_i, Y0_i)

# Create the dumbbell plot
plot <- ggplot(data) +
  geom_segment(aes(y = as.factor(i), x = Y0_i, xend = Y1_i, yend = as.factor(i)),
               color = "#aeb6bf", linewidth = 1.5, alpha = 0.5) +
  geom_point(aes(x = Y0_i, y = as.factor(i), color = as.factor(D_i), shape = "Y0"), size = 4) +
```

```
  geom_point(aes(x = Y1_i, y = as.factor(i), color = as.factor(D_i), shape = "Y1"), size = 4) +
  scale_color_manual(values = c("#009688", "#762a83"), name = "Treatment Status") +
  scale_shape_manual(name = "Potential Outcome", values = c(17, 16), labels = c("Y0", "Y1")) +
  labs(x = "Test Score", y = "i") +
  theme_minimal() +
  theme(
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )

plot
```



3 Potential Outcomes Recap

The following questions are designed to help you get familiar with the potential outcomes framework for causal inference that we discussed in the lecture.

Task 1.1
Explain the notation Y_{0i}.

Task 1.2
Explain the notation Y_{1i}.

Task 1.3
What is the difference between Y_{0i} and Y_i?

Task 1.4
Can we observe both Y_{0i} and Y_{1i} for any individual unit at the same time?

Task 1.5
If D_i is a binary variable that gives the treatment status for subject i (1 if treated, 0 if control), what is the meaning of $E[Y_{0i}|D_i = 1]$?

Task 1.6
The table below contains the potential outcomes (Y_{1i} and Y_{0i}) and the treatment indicator (D_i) from a hypothetical experiment with 6 units.

$Unit$	Y_{1i}	Y_{0i}	D_i
1	2	2	1
2	3	-1	1
3	-1	9	1
4	17	8	0
5	12	9	0
6	9	1	0

Complete the following calculations by hand:

1. List the observed outcomes (Y_i) for the experiment based on the table above.
2. Calculate the "true" average treatment effect (ATE) based on the potential outcomes.
3. Calculate the "true" average treatment effect on the treated (ATT) based on the potential outcomes.
4. Calculate the "estimated" average treatment effect based on the naive difference in group means for treatment and control conditions from the observed outcomes. Explain the difference between this estimate and the "true" average treatment effect.

4 Islam and Authoritarianism

In a famous paper titled "Islam and Authoritarianism," Steven Fish asks whether Muslim societies are less democratic.[2] He runs a series of cross-sectional regressions of countries' *Freedom House scores* (an indicator of the level of a country's democracy) on characteristics of the countries, including whether they are predominantly Muslim.

The paper's dataset is in the spreadsheet `fishdata.csv`, which you can download from Canvas. You should load the data using the `read.csv()` function. This data contains the following variables (among others):

- `FHREVERS` - Freedom House scores, a measure of democracy where higher values indicate that a country is more democratic and lower values indicate greater authoritarianism
- `MUSLIM` - 1 if a country is predominantly Muslim, 0 otherwise
- `GDP90LGN` - the country's GDP in 1990
- `GRW7598P` - the country's average annual economic growth from 1975-98, in percent
- `BRITCOL` - 1 if the country was a British colony, 0 otherwise
- `OPEC` - 1 if the country is a member of the OPEC group of oil-exporting countries, 0 otherwise

Task 2.0
View the first 6 rows of this data. What is the unit of observation?

Task 2.1
Taking subsets and summarizing variables:

1. How many countries are predominantly Muslim?
2. What percentage of countries are predominantly Muslim?
3. How many countries have a GDP of above 3.0 in 1990?
4. How many countries are both Muslim and a former British colony?
5. How many countries have either an average economic growth from 1975-98 of above 0.6
6. Create a new dataset consisting only of countries that are both Muslim and a member of OPEC.

(Hint: Use square brackets to denote subsets of a variable or dataset. You'll also need the `length()` function.)

Task 2.2
What is the difference in mean Freedom House score between Muslim and Non-Muslim countries? Calculate it both by hand and using a regression, verifying that your answers are identical.

Task 2.3
Is the difference in means calculated above likely to be biased? If so, in which direction and why?

Task 2.4
Conduct a t-test for the difference in means calculated above using the `t.test()` function. Is the difference statistically significant?

Task 2.5
Conduct the t-test again, this time coding it by hand. Confirm that your answer is identical to the answer you calculated in the question above.

[2] M. Steven Fish (2002). "Islam and Authoritarianism." World Politics, 55 (1): 4-37.

Task 2.6
Calculate:

1. The percentage of Muslim countries that are former British colonies
2. The percentage of non-Muslim countries that are former British colonies
3. The correlation between being a former British colony and Freedom House score, controlling for being Muslim [3]

Use these results to explain the impact that controlling for BRITCOL has on the estimated effect of MUSLIM.

Estimate a linear regression with `FHREVERS` as the dependent variable and `MUSLIM` as the independent variable. How do the results from your regression relate to the difference-in-means that you calculated in Task 2.2?

Task 2.7
Repeat task 2.6 for `OPEC`, `GRW7598P` and `GDP90LGN`. For `GRW7598P` and `GDP90LGN`, simply calculate the correlation between each one and `MUSLIM` and don't estimate the percentages like in steps (i) and (ii).

Task 2.8
Now estimate a regression of `FHREVRERS` on `MUSLIM`, `BRITCOL`, `OPEC` and `GRW7598P`. Again, do the results make sense?

[3] Note that steps 1 and 2 are akin to measuring the correlation between being a Muslim country and being a former British colony. We do it this way because a correlation coefficient is defined only for two continous variables, and these are both binary.

5 R Markdown, LaTeX, and PDFs

For your problem-sets and exam, you will submit two files: (1) a PDF report/document with your analysis and write-up, and (2) the associated R/Rmd script that contains your code. To produce PDF reports that combine your write-up (plain text) and analysis (in code chunks), you can use R Markdown files. You can also incorporate latex options, e.g., for typesetting, math mode, and formatting. Instructions to create your first PDF file from R Markdown (Rmd) are summarized below:[4]

For MacOS:

- Before proceeding, ensure your system is running MacOS 11 or higher.
- Download and install MacTex.pkg from https://www.tug.org/mactex/mactex-download.html. This file is quite large so download can take some time.
- Download and install R if you haven't already done so, from https://cran.rstudio.com/. Then, also download and install RStudio from https://www.rstudio.com/products/rstudio/download/.
- Once these installations are complete, launch RStudio. Click File, then New File, R Markdown. If RStudio prompts you to install any required packages that are missing, select Yes.
- You'll get a dialog box like the one shown in Figure 2. Enter the document title, author, date, and ensure you select the PDF output option. This sets up a basic YAML header in your Rmd file. You can modify this as needed.
- Before producing your PDF document, you need to save your Rmd file. Select the directory where you want your PDF to be stored. If prompted, set character encoding to your system's default or to UTF-8.
- Knit your Rmd file to get your PDF document.

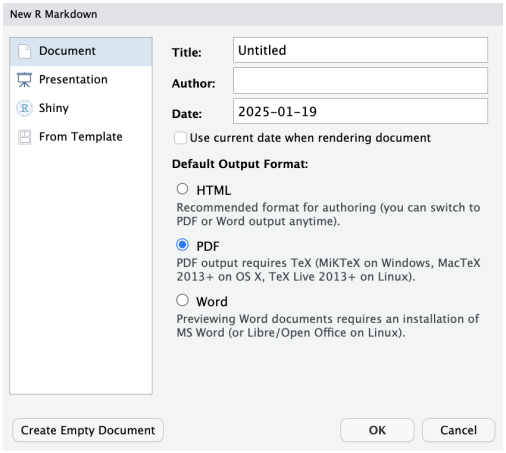


Figure 2: New Rmd file dialog box (MacOS)

[4] These instructions were compiled from helpful tutorials published by Søren L Kristiansen: linked here (Mac version), and here (Windows version).

For Windows:

- Before proceeding, if you are using Windows 10, figure out whether your operating system is running 32-bit or 64-bit.
- Download and install MikTeX from http://miktex.org/download. Find the version which is compatible with your system (from the step above). Download and install the 'Installer' version.
- Download and install R if you haven't already done so, from https://cran.rstudio.com/. Then, also download and install RStudio from https://www.rstudio.com/products/rstudio/download/.
- Once these installations are complete, launch RStudio. Click File, then New File, R Markdown. If RStudio prompts you to install any required packages that are missing, select Yes.
- You'll get a dialog box where you can enter the document title, author, date. Ensure you select the PDF output option. This sets up a basic YAML header in your Rmd file. You can modify this as needed.
- Before producing your document, you need to save your Rmd file. Select the directory where you want your PDF to be stored. If prompted, set character encoding to your system's default or to UTF-8.
- Click on Knit PDF. The first time you try to generate a PDF like this, RStudio will find that one or more LaTeX packages are missing, and prompt an installation. Click the Install button, and repeat this as many times as prompted (there could be several missing packages). Your PDF file will generate once all requirements have been installed by RStudio.

You can now easily include math notation in your PDF document. For example, if I type `$\beta_{0}$` within a sentence, the PDF file formats it as β_0! Now, say I want an equation in a new line. I need to use two `$` signs, so if I type `$$Y_{i} = \beta_0 + beta_1 X + \epsilon_i$$`, it'll look like this:

$$Y_i = \beta_0 + beta_1 X + \epsilon_i$$

You can also use latex environments like `\begin{itemize}`, and latex commands such as `\textbf{boldtext}` for **boldtext**. For more details on how to use latex styling, packages, etc. in your Rmd PDFs, refer to this section from the R markdown cookbook linked in Section 6.

6 Resources and help

The more you use R (or any other language), the better you will get at it. This includes doing things the wrong way and learning what produced your error!

1. For 'How do I ...?' or for troubleshooting errors in your code:
    - CRAN website: searching with appropriate keywords will direct you to relevant packages and functions, with explanations of syntax and usage.
    - Google, Stackoverflow: always look up your errors first. The same applies for 'How do I ...?' – learning how to search for solutions is a key skill. Someone has likely run into your issue before, or has asked the same question before.
    - Claude, ChatGPT, etc.? Only use *cautiously* for help with debugging *your* code: be mindful that these models **do** give wrong answers at times. Usual disclaimers apply regarding ethics/plagiarism.
2. For help with producing PDF documents:
    - RMarkdown for Scientists by Nicholas Tierney covers the essentials of using Markdown.
    - R Markdown: The Definitive Guide by Yihui Xie, J. J. Allaire, Garrett Grolemund delves into more details and other advanced features.
    - This Markdown guide contains handy summaries and examples of markdown syntax

6.1 Some general resources

1. Tutorials to refresh R, unless you are very familiar: (all three cover similar material, just pick one you like best and work through it)
    - The R Guide by WJ Owen
    - Intro to R by Venables and Smith
    - Simple R by John Verzani
2. Detailed guides:
    - R for Data Science by Wickham and Grolemund (focused on 'tidyverse')
    - R Cookbook by Long and Teetor
    - Undergraduate guide to R by Trevor Martin
3. Functionality, structures, workflow etc.:
    - Intro to statistical programming methods with R by Beckman et al.
    - Intro to programming with R by Stauffer et al.

Cheat sheet for:
Base `R` is linked here
Data tidying with `library(tidyr)` is linked here.
Data wrangling with `library(dplyr)` is linked here.
Visualizing with `library(ggplot2)` is linked here.
Producing reports with `R Markdown` is linked here.