Problem Set 2

Due Monday September 18, 2023, 10am

To successfully complete this problem set, please follow these steps:

1. Download the contents of the assignment Dropbox link in a folder (directory) on your computer designated for this problem set only. Follow the directory structure, e.g., putting the `data` folder containing the datasets as a subdirectory of the project directory.

2. Insert your answers in the yellow boxes using Microsoft Word, and prepare a single `.R` script for what you produce. Save the word document as a `.PDF`.

3. Please submit the PDF to the designated `PS-XX: pdf` link and your R Script to the `PS-XX: R` link.

(1) Your name:

Nikhilla Bhuvana Sundar

(2) Group members, if any:

(3) Compliance with the Academic Code on problem set[1] (sign with an X below)

X

[1] You may use the same code from classmates, Ed Discussion Board, instructors, and generative AI. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, or use generative AI to provide potential code for you, but you must each type your own answers and your own code.

In this problem set, we will continue with:

> Banerjee, A., Duflo, E., Karlan, D. et al., 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*. (Canvas reading link)

Gradescope will have pre-installed the following packages:

```
tidyverse
scales
gt
modelsummary
```

This problem set is about the analysis of randomized control trials and estimating the average treatment effect with the SATE (sample average treatment effect). Read up to the section "Country-by-country variation" before starting on this problem set or as you work through it.

As always, hints are provided at the end.

Problem 1

Suppose the causal effect of the Graduation program on an outcome Y was τ for all individuals. In other words,

$$Y_{1i} - Y_{0i} = \tau$$

for all individuals i. Show mathematically that, if being assigned to the Graduation Program, D is randomly determined, then the difference between the expected outcome among individuals assigned to treatment $E[Y|D_i = 1]$ and the expected outcome among those assigned to control $E[Y|D_i = 0]$ is τ.

Given that, $Y_{1i} - Y_{0i} = \tau$

$We\ have\ Y_{1i} = \tau + Y_{0i}$ (1)

When $D_i\ is\ randomly\ assigned, the\ difference\ between\ E[Y|D_i = 1]\ and\ E[Y|D_i = 0]\ captures\ the\ causal\ effect\ of\ the\ treatment$

$= E[Y|D_i = 1] - E[Y|D_i = 0]$

$= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$

From (1) we have,

$= E[\tau + Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$

$= \tau + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$ τ can be taken out as expectation is a linear operator

And the above equation leaves us with τ

Problem 2

The dataset `graduation_household.rds` includes the household-level variables measured for this study. We analyzed a subset of this dataset in the previous problem set. `rds` is a format for any generic R object.[2]

One naming convention that the authors use is to denote outcomes measured at different survey waves by the suffixes `_bsl` (baseline), `_end` (endline 1), and `_fup` (follow-up, or endline 2). The variables include:

Variable name	Description
`id`	Household ID
`country`	Country in which Survey took place
`assignment`	Whether the household was assigned to the *Graduation* treatment. A numeric variable that is 1 if assigned to treatment, 0 if not.
`ctotal_pcmonth_bsl`	Total monthly per capita consumption at baseline (2014 USD PPP)
`ctotal_pcmonth_end`	Total monthly per capita consumption at Endline 1 (2014 USD PPP)
`ctotal_pcmonth_fup`	Total monthly per capita consumption at Endline 2 (2014 USD PPP)

2.1

Show **two** histograms of "Total monthly per capita consumption at endline 1" (hereafter referred to as endline 1 consumption) in the same plot, side by side: one for the households assigned to treatment and another for the household assigned to control groups. Use the `facet_wrap()` layer in ggplot to make smaller multiples of the same type of geometry repeated for each group of a different variable. facets are different from geometry / scale / theme layers.

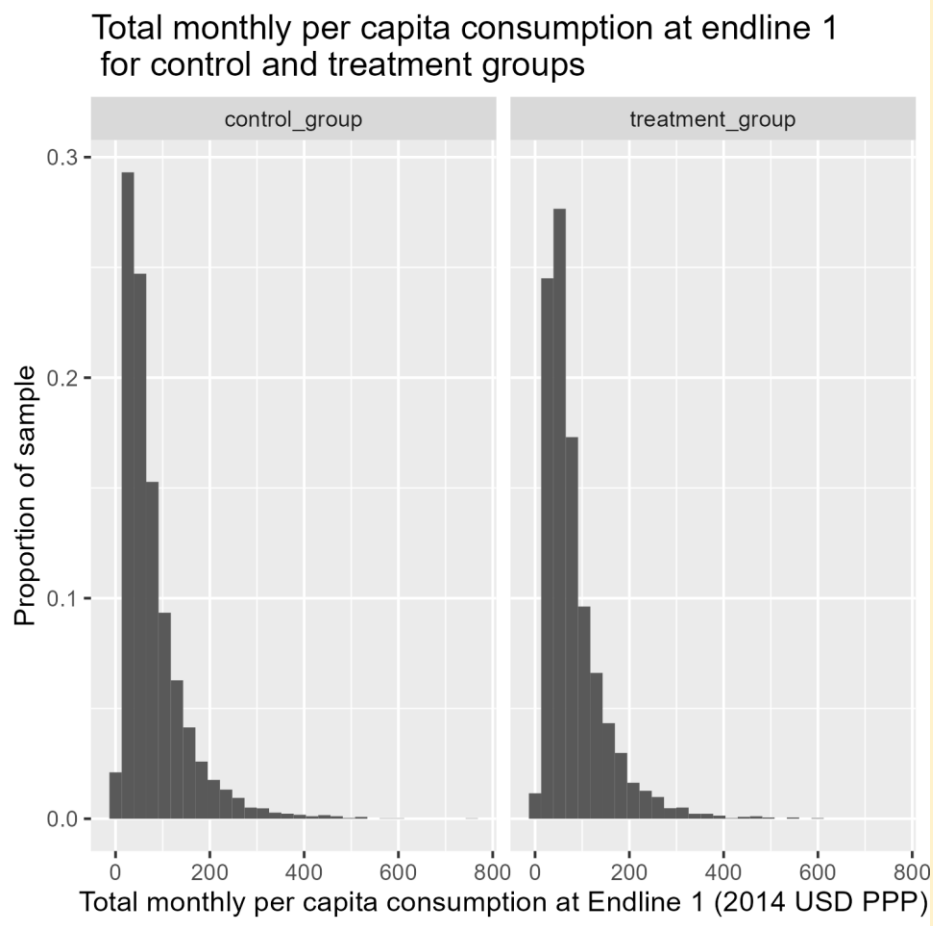
The figure should be labelled properly, and the following is also recommended to make the figure professional:

- Similar to the previous problem set, try adjusting the histogram so that the bins do not appear to take on negative values, with `boundary = 0`[3].

[2] Rds folders are read in via the `read_rds` function in tidyverse, and By saving it in this way, attributes like factor orderings are saved instead of being removed in plain-text formats like a csv.

[3] The values of boundary can range from 0 to 1. Think of the 0, 0.5, 1 as a position on a number line from 0 to 1. 0 (to the left) is left alignment of the bins, 0.5 is center alignment at 0. 1 is right alignment.

- You can also specify how finely binned you want the data to be by the `binwidth` argument (for example `binwidth = 20` means that each bin's width is set to 20 units of the outcome).[4]
- Finally, try showing *fractions* instead of counts on the y-axis by setting the `y` aesthetic to `y = stat(width*density)`. Showing fractions is better suited to compare the SATE because groups of different sample sizes will be converted to the same proportions.



2.2

The previous histogram showed the entire distribution of consumption with all six countries mixed together. Now we will create a table showing the average endline 1 consumption in *each assignment group* in *each country*.

[4] Both `binwidth` and `boundary` are not aesthetics in this example, so they are set outside of the `aes` function.

To do this, use `group_by()` and `summarize()` to and take the average of the endline 1 consumption variable in each of those 12 combinations. When doing this, make sure to handle missing data correctly:

- For some households, the consumption is not recorded because the respondent could not be reached. Missing values are labelled as `NA` in R. You can see which rows have missing values by running, e.g., `dat |> filter(is.na(ctotal_pcmonth_fup))`.

- When taking summary statistics like means and standard deviations, we need to specify the function to **remove missing values** from the data before taking the summary, otherwise the mean of a vector with at least one `NA` is also `NA`. Do this by setting `na.rm = TRUE` within the `mean` function.

All this should give you a summary tibble with 12 rows. (You do not need to show this output in the yellow box, but you will want to save this as its own object to use in the next question).

Next, use `filter` and `left_join` to create a table with only 6 rows, one for each country. **Name this `c_avg` for the autograder.**

- There should be three columns (and no more than three): one column for country, one column for the per capita consumption in the control group, and another for that in the treatment group.

- Reproduce this table in the yellow box (either use `gt()` or a screenshot).

This sets your table up well to compute the sample average treatment effect within each country via difference-in-means of treatment.

Country	mean_control	mean_treatment
Ethiopia	41.71	48.08
Ghana	40.81	42.52
Honduras	79.56	80.16
India (Bandhan)	47.81	54.48
Pakistan	86.89	96.11
Peru	147.40	142.93

2.3

Interpret the difference in means estimate for Peru substantively. In this class, by "interpret substantively", we will mean providing an interpretation of the number in units that is understandable to a policymaker who is not well-versed in statistics and who has not seen the raw dataset. For example, something like this:

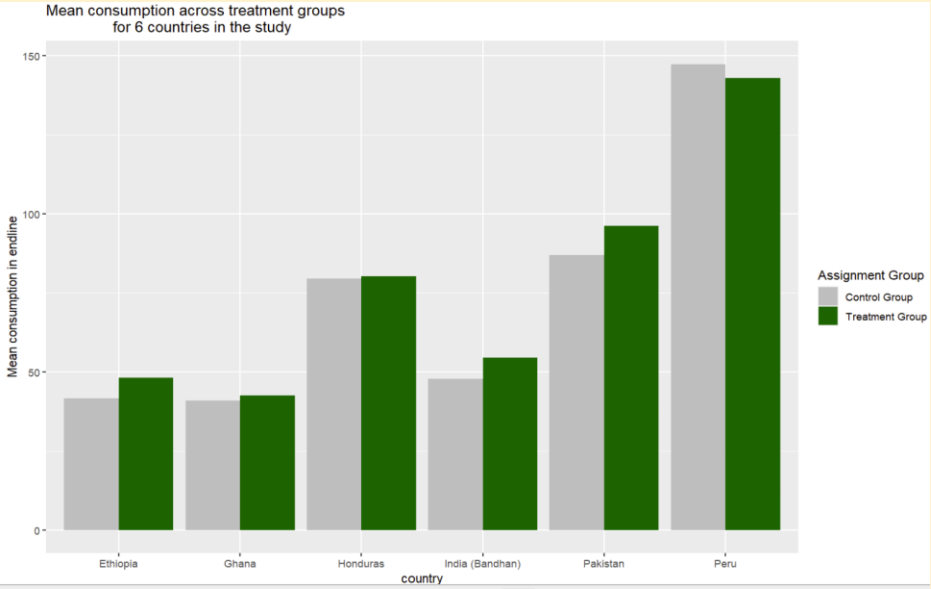
> "The difference in means estimate for the call-back rate of a Black sounding name is -0.02. This means that compared to résumés with a White sounding name, the Black sounding name is 2 percentage points less likely to receive a call back."

The difference in mean estimates of per capita consumption between control and treatment groups in Peru at endline is about $5. This means that when compared to control group, those households that received the graduation program) group consumed $5 less per month on an average as per endline 1 estimates.

2.4

Present the same statistics of country-level means by treatment group as a barchart. There should be 12 bars (6 countries, 2 groups). Let's spend some time making this barchart presentable:

- The bars should come in pairs: a pair of bars for each country, side-by-side.
- Color the bars by treatment group, e.g. control is gray and treatment is dark green. Label the legend too, so that the control is labelled "Control" and not merely "0", etc..
- Axes should be labelled with human-readable titles.



For a figure like this, using the "long" shaped dataset is more amenable than a wide dataset with treatment groups sorted across columns. This is why the long format is what tidyverse refers to as "tidy data".

Problem 3

We will now report the treatment effect via linear regression.

3.1

Run a single regression equation regressing the outcome on the treatment assignment in the Peru subset of the data. **Name this object `fit_peru` for the autograder**. Report the coefficient estimate, and whether you can verify that your conclusion matches from that of the previous difference in means.

The coefficient estimate for the regression equation is -4.466. This verifies the conclusion that matches from the previous difference in means in the table above. While we can verify that per capita consumption of treatment group is about $5 lesser than control group at endline 1, these results are not significant

3.2

Use the `modelsummary` package to create a regression table with six columns, one for each country. The table should show the SATE, intercept, and sample size of each country. The standard errors of each coefficient will be shown in parentheses – you should keep them, but don't worry about interpreting these for now if you are unfamiliar. Do not display summary statistics other than the sample.

	Peru	Ghana	India	Ethiopia	Honduras	Pakistan
(Intercept)	147.395	40.809	47.809	41.708	79.558	86.893
	(2.298)	(0.668)	(1.300)	(1.372)	(1.388)	(2.482)
assignment	-4.466	1.706	6.668	6.376	0.606	9.218
	(3.925)	(1.318)	(1.793)	(1.950)	(2.396)	(3.462)
Num.Obs.	2104	2525	816	915	2287	1041

3.3

Imagine a critic who sees your regression results and complains that you should not be estimating treatment effects by linear regression because the distribution of income is clearly skewed and not normally distributed, as shown by the histogram you showed first. How would you respond?

- Large enough sample size – A large enough sample size for each country that is significant can reduce the impact of a not so normally distributed income variable.
- Log transformation – We can also transform the income variable through logarithmic transformations which can make it more symmetric and normally distributed
- Randomization – Even though the income variable is not normally distributed, since the treatment assignment is randome, this will ensure that the control and treatment groups are similar in all characteristics. Through this randomization, we can arrive at causal inferences.

Hint

Problem 1

Answering this problem confidently requires knowing the mathematical properties of Expectation in statistics. Specifically, expectation is a linear operator such that $E(X + Y) = E(X) + E(Y)$ for any random variables X and Y. Also, expectation affects only random variables, such that if a is a constant, $E(a) = a$. For a condensed summary of expectation in the context of this problem, read 1.1 of *Mastering Metrics* carefully. If you have not encountered this in your past statistics classes, that's ok, but please still attempt to provide a verbal reasoning. We will not remove points for errors in math. This type of understanding is not the modal assessment in this course, but some statements like this in our text is important.

Problem 2

2.1

- Some examples of `facet_wrap()` are in the web version of the help page website: https://ggplot2.tidyverse.org/reference/facet_wrap.html
- Some examples of small multiples and how it can be an effective visualization tool are discussed in Fundamentals of Data Visualization ch. 21 https://clauswilke.com/dataviz/multi-panel-figures.html

2.2

- When you join, remember that you want to join countries together, but nothing else. You will need to specify this in the `by` argument.
- Suppose dataset `X` has columns `a`, `b` and dataset `Y` has columns `a`, `b`, if we left join `X` and `Y` by the only by `a`, then we will get a dataframe that has columns `a`, `b.x` (for the b variable from the left hand side) and `b.y`. You will want to rename the two using `rename` or `select` to make clear which one is control and which one is treatment. `.x` and `.y` are the default suffixes that are put on but you can change them with the `suffix` argument in `left_join`.

- If you apply `gt()` to a grouped dataset, it will by default produce subtables by group. While this is sometimes helpful, it is not here. If your tibble is grouped and you want to use `gt()` on it, ungroup the tibble first by applying `ungroup()`.
- An alternative way to make this 6-row table, is to use `pivot_wider()`. If taking this route, you will want to reshape the tibble into a "wide" table with 6 rows, one for each country.
- `pivot_wider()` is a common function changing the shape of a dataset from long to wide format, and we will be using it even more when we get to panel data in week 6. In this case it will not change the information in the dataset but it will rearrange the cells. The Imai and Williams textbook chapter 2 has various examples of using `pivot_wider()`. A full example can be read through the help page or the website: https://tidyr.tidyverse.org/reference/pivot_wider.html.

2.4

- To dodge bars by a variable that is not the x-axis, specify the `position` argument within the `geom_col()` geom. In the full list of ggplot functions

https://ggplot2.tidyverse.org/reference/index.html, find the set of functions that start with `position_` and look for the one that dodges the bars by a variable.

- Use the layer `scale_fill_manual` to manually set a fill color for each value of the data.
- Like the previous problem set, inside the `scale_fill_manual`, you can also specify the labels for each color. In addition to the `values` argument, you can specify a `labels` argument in exactly the same way, except now the colors are replaced by labels for the colors in the legend.

Problem 3

3.1

- To check that you did your regression right, you can check that your reported treatment effect matches with your table above.
- Remember to subset your dataset so that you only estimate the SATE within Peru

3.2

- The `modelsummary` package is one of several R packages that offer to turn regression output into tables, and can create them into tables in HTML, plain text, MS Word, or LaTeX. The package website https://vincentarelbundock.github.io/modelsummary/articles/modelsummary.html has some helpful examples.
- When showing multiple regression objects as columns of a table, `modelsummary` needs all the objects to be bundled together. R does this most generally by what is called a "list", using the `list` function. This is similar to a vector using the `c` function (e.g. `c("A", "B", "C")`) but lists can be a combination of more complex and heterogeneous objects that are not numbers of characters. The item of each list can be named by using the left handside of an equals operator, e.g. `list(Ethiopia = fit_eth, Peru = fit_peru)`) .The first example in the modelsummary website linked above has an example.
- `modelsummary` reports a couple of diagnostic statistics in each column as a default but many of these diagnostics are rarely interpreted in a paper. To remove everything but the sample size, you can add an argument `gof_map = "nobs"` to your modelsummary function. GOF is the package author's abbreviation for goodness of fit, and `"nobs"` indicates the number of observations in the table.
- If you know how to do a for loop, you can do that and come up with shorter code. A good guide to doing for loops in R is https://rstudio-education.github.io/hopr/loops.html.