Problem Set 2

SOC-GA 2332 Intro to Stats (Spring 2021)

Due: Saturday, Mar. 20th, 11:59 pm

Instructions

1. Submit two files for each problem set. The first is a **R Markdown** (`.Rmd`) file that can be run without error from start to end. The second is a **PDF** rendered from your R Markdown file or created using LaTeX.
2. Name your files following this convention: `[Last Name]_ps1.Rmd` and `[Last Name]_ps1.pdf`.
3. Both files should be submitted to the TA via e-mail (di.zhou@nyu.edu) before the time specified above.
4. You are given plenty of time to work on the problem set. Please plan ahead and start early. **Except for special circumstances, the TA will not accept last-minute questions asked on the day when the problem set is due**.
5. You are encouraged to discuss the problems with your classmates. Notice as well that we have students in this class who are not in your cohort. It would be great if you could reach out to them and work together. But **the R Markdown and PDF files that you submit have to be created on your own**.
6. Comment on your code wherever possible and explain your ideas in detail. You will get credit for showing the steps you take and for explaining your reasoning, even if you do not get the correct final result.

Prerequisite

Load multiple packages to your environment using the following code (you can add more packages to the current list as per your need):

```
knitr::opts_chunk$set(echo = TRUE)

library(pacman)
p_load(tidyverse, foreign, corrplot, stargazer, coefplot, effects)
```

Part 1: The Replication Project

This week, we will begin familiarizing ourselves with the replication exercise. Download and save the paper that we will be replicating. Make sure to download and save the **Online Appendix** (you will find it on the ASR website).

1. Read the paper closely and respond to the following questions:

(a) What are the authors' research questions?

(b) What is the gap in the literature that the authors aim to fill? How does their analysis advance the literature?

(c) What is the population that they are making inferences about? Be specific and make sure to identify the geographical region that they are focusing on, the time period, demographic characteristics, and so on.

(d) Do they have data on all individuals in the population? If they don't, how do they solve this?

(e) Did the authors collect the data themselves? If they did, describe their sampling procedure. If they didn't, identify and describe the data source and discuss the sampling procedure that was used to collect this data.

2. As you read the text, make a list of all variables or characteristics of the population that the authors mention throughout the paper (e.g., gender, age, wages, occupation, . . .). Submit a list with all the variables and characteristics that you have identified in the text. *Note:* Make sure to read footnotes and table notes; they contain important information to understand who is in the sample.

3. Based on your answers to prior questions, select the samples and variables that you think you will need to replicate the paper in your IPUMS account. Submit a screenshot of the page where you can see the samples and variables that you have selected. *Note:* you can obtain this from your "Data Cart".

Part 2: Regression

Import the dataset `sat_math.dta` to your R environment and examine the effect of IQ and other variables on SAT math score. *Hint:* use `read.dta()`

Variable Name	Variable Detail
`sat_math`	SAT Math Score
`female`	The Female Dummy (Male = 0)
`black/other`	Two Racial Dummies (White as the Reference Group)
`meduy`	Mother's Years of Schooling
`feduy`	Father's Years of Schooling
`hours`	Average Weekly Study Hours
`IQ`	IQ Score (0 to 100)

1. Report descriptive statistics:

(a) Create a table that reports descriptive statistics (you should at least report the means) of all the variables grouped by gender. Its structure should look similar to the "1970" column of Table A1a in the replication paper, except that we are using gender instead of race as the grouping variable. *Note:* You can either use `tidyverse` (`group_by()` and `summarise`) then transpose the table, or search other R packages that can generate descriptive statistics table for you. At this point, don't worry too much about the specific formatting issues of the table. In practice, we will format the table in LaTeX instead of R Markdown.

(b) Create a correlation matrix and display it. Then visualize the matrix using any one plot option that you think can effectively communicate the result from the `corrplot` package as demonstrated in this webpage.

2. Create scatter plots:

- Besides the key dependent variable (DV) `sat_math`, choose one numeric independent variable (IV) that seems to have a meaningful relation to the DV based on the correlation matrix you created, and then create the following plots. *Note:* Make sure to add appropriate titles, axis labels, and plot legend for each plot.

(a) A scatter plot of the DV and IV

(b) A scatter plot of the DV and IV with a fitted linear regression line

(c) A scatter plot of the DV and IV, and each observation is color coded by gender (make sure to coerce your gender variable to a character or factor using `as.character()` or `as.factor()`, or recode your gender variable to a character variable with values "Male" and "Female" using `mutate()` and `ifelse()`.)

(d) On top of plot (c), fit a linear regression line for each gender group, the lines should also be color coded

3. Additional exploratory data analysis:

(a) What are your preliminary findings/reflections on the data based on the descriptive statistics, the correlation matrix, and the scatter plots?

(b) What other exploratory data analysis will be useful for you to better understand the data before modeling? Please implement some additional exploratory data analysis and discuss your preliminary findings.

4. Nested models:

- Build five nested models that use `sat_math` as the DV and report regression results in a table using `stargazer()` from the `stargazer` package.*Note:* You can set `type = "text"` in `stargazer()` when viewing the table in R, but before knitting to PDF you should set `type = "latex"`

(a) Model 1: Baseline (only add "IQ" as the independent variable)

(b) Model 2: Model 1 + Demographic Characteristics

(c) Model 3: Model 2 + Parental Education

(d) Model 4: Model 3 + Weekly Study Hours

(e) Model 5: Model 4 + An Interaction Between IQ and the Female Dummy

5. For the result of Model 1:

(a) What are the hypotheses that you are testing in this model with your t-values in the `(Intercept)` and `IQ` row of the modeling results?

(b) Create a 95% confidence interval for the parameter β_{IQ} based on Model 1 result.

6. Interpret regression coefficients:

(a) How does the coefficient of "IQ" change across models? What could be the possible reason(s) for such changes?

(b) Interpret the coefficient of "black" in Model 4.

(c) Interpret the coefficient of "meduy" in Model 4.

(d) Interpret the coefficient of the interaction effect between IQ and the Female Dummy in Model 5.

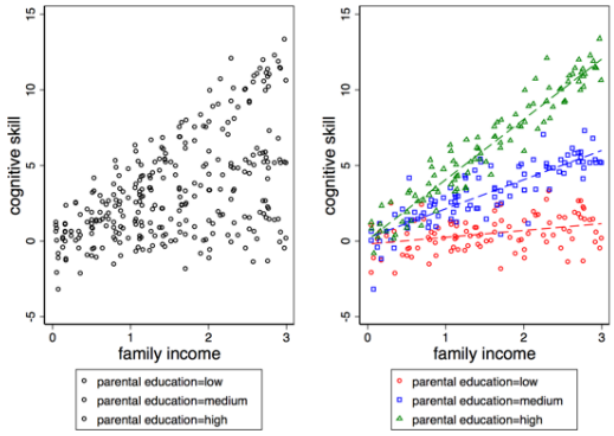
7. Create a coefficient plot for Model 5 with appropriate title and labels.

8. On the basis of Model 5, by holding other variables at their means, create a figure demonstrating the predicted SAT math score by gender and IQ levels (with confidence interval).

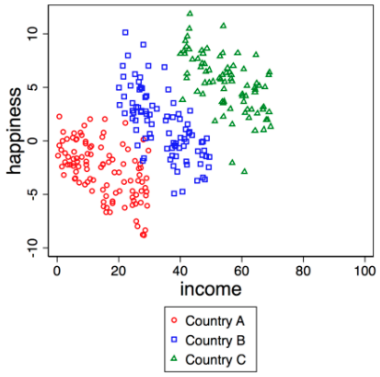
Part 3 (Bonus) Data Simulation

Simulation is a fun and effective way to learn about statistical inference. You will get a better understanding of how each population parameter affects the shape of the distribution.

Now that we have learned about how to identify interactions from a given sample, you can try simulate a data whose true data generating process involves interaction between two variables. For example, you can try to reproduce a similar scatter plot we saw in class (the right panel) by simulating a data whose variables have such associations:



Or, you can try to reproduce a scatter plot that demonstrates the Simpson's Paradox:



Note: Your output does not need to replicate the exact layout of the example graphs. You will get extra credit as long as you generate a similar graph that illustrates the relationship (either a positive or negative interaction, or the Simpson's Paradox) clearly. Remember to use `set.seed()` for any random process.