Introduction to Statistics (SOC-GA 2332)

Assignment III

Due Date: September 6th, 11:59pm

Instructions

Submit a pdf-file that answers the following problems together with the `R` script that is used in your analysis. Make sure that the script runs without error from start to end before submitting it.

The files you submit should be named as follows: `[Last Name]_Assignment3.pdf` and `[Last Name]_Assignment3.R`. If you are Bob Smith, your document should be named `Smith_Assignment3.pdf` and your `R` script `Smith_Assignment3.R`. The pdf-file should be created with LaTeX or Rmarkdown.

You are encouraged to work with others on the problem sets. However, the documents and `R` scripts that you submit have to be created on your own. Both the pdf-file and the `R` script must be submitted to me via e-mail **before September 6th, 11:59pm**.

Important Notes:

1. These problems are designed to help you figure out what to do when you are confronted with a statistical problem that is not obvious at first sight, rather than to test your knowledge. You've learned all the relevant concepts to answer these questions, either in the lecture or in the math camp. The difficult part is putting these things together.

2. Some of these problems are easy, while others might be rather challenging. In either case, you probably won't be able to solve all of the problems in one day. So, please **plan ahead**. Read through the problems and take your time thinking about them.

3. To reiterate, you are encouraged to **discuss the problems with your classmates**. Some of you will be better in dealing with probability, others will be more advanced in coding; so I hope that you will learn from each other. 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. Also, in the title section of the report you submit **write down the names of all people with whom you've worked together**.

4. When answering the questions, try to be as detailed as possible. Otherwise, it is difficult to see exactly where you did a mistake, in the case your submitted answer is incorrect.

5. If there are one or more problems for which you remain completely clueless by **September 4rd**, send me an e-mail (`bp1094@nyu.edu`). In this e-mail, **please**

tell me what you have tried so far, so I can understand exactly why you are struggling and help you efficiently. If you forget to include your prior trials, my first response will be most likely "can you tell me what you have tried so far?"

6. On September 7th, I will share a solution manual that explains the answers to the problems. In the lab following that date, we will have some time to discuss the problems. So, **please read the solution manual before you come to class**.

Less Important:

1. In all numerical answers, round the numbers to three decimal digits. That is, if you get 2.2014542, you should report 2.201.
2. When you include figures in your document, include them within the text rather than pushing them to the end of the document.
3. When creating your `R` script, please keep in mind the **Style Suggestions** that I've provided throughout the Math Refresher and the labs.
4. Make sure that your `R` script is well-documented with comments. In addition, make clear which parts of your code are used for what problems in the problem set. I have numbered all problems, so this should not be difficult.

PROBLEMS

Problem 1 (Assumptions)

Consider the "true" data-generating process (DGP)

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

1. Suppose we obtain a sample, $\{Y_i, X_i\}_{i=1}^n$, from this process. State the assumptions under which the OLS estimator $\hat{\beta}_1$ is BLUE (Best Linear Unbiased Estimator) for β_1.
2. For each of the assumptions, discuss what will go wrong when the assumption is violated.
3. Let $\beta_0 = -0.25, \beta_1 = 1.2, X \sim \Gamma(5, 4)$, and $\epsilon \sim \text{Normal}(0, 1)$. Here, $\Gamma(\alpha, \psi)$ denotes the Gamma distribution with shape parameter α and rate parameter ψ. A simple Google search will lead you to a way to simulate from this distribution (Don't get scared away from the new distribution and the fancy Greek letters!).

   Simulate a dataset of size $n = 3,000$ from this process in which all of the assumptions you've discussed above hold.
4. For each of the assumptions, create a plot which illustrates how the violation of the assumption affects the regression results (all other assumptions should be satisfied). This can be a scatter plot, a plot of the regression line, the sampling distribution

of the OLS estimator (comparing what the regression table suggest with actual simulations), or anything else. The important point is to show how the violation *leads us to false decisions* if we assume the assumption is true, not just showing that the assumption *is* violated. When simulating data, you don't have to use the parameters set in the previous problem. Also, you can skip the "no autocorrelated errors" assumption, which goes beyond the course materials.

Here is a big hint for this exercise: Let `x` be a length-`n` vector of simulated values. To simulate `n` values from a Normal distribution which 1) has a mean equal to zero, 2) is uncorrelated with `x`, 3) but has a variance that depends on `x` by the function `g()`, you can use the code

```
rnorm(n, mean = 0, sd = g(x))
```

Problem 2 (Back to the GSS)

Recall that we have created two abortion attitude variables in the lab: the Rossi scale and a new measure that was introduced in 2018 (new measure hereafter). We have analyzed the Rossi scale but not the second measure.

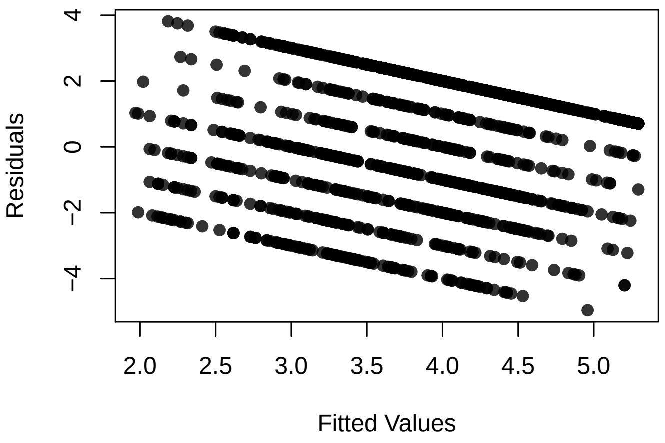
1. Plot the distribution of the new measure using a bar plot
2. Using the new measure as the outcome variable fit both Model 1 and Model 2. In these models include also a new variable which is 1 if the respondent is currently married and 0 otherwise.
3. Calculate heteroskedasticity-robust standard errors for Model 2; call this model Model 3.
4. Create a regression table with three columns: Model 1, Model 2, and Model 3.
5. Using the new measure create the same type of plot as that shown right above the "diagnostic" section. Be clear about which variables you fix at what values when generating the plot.
6. Using both the tables and the plot, interpret the results.

Problem 3 (Stripes)

The fitted vs residual plot we created in the last lab is reproduced below:

You will be able to create this plot by following the code of the lab. You see that the plot consists of a number of parallel "stripes." You'll see this kind of pattern quite often when analyzing survey data. In answering the following questions, it might be helpful to play around with the data a little bit, by drawing lines through the plots, trying to color code the points, and so on (although, I believe that the questions can also be answered by thinking carefully about how the residuals and predicted values are produced).

1. Explain why the number of stripes we observe is equal to seven and determine the vertical distance between the stripes. Be specific in your reasoning.



2. Determine the slope of each of these stripes. Explain why this has to be the case.

3. Explain why the stripes are are parallel.

Problem 4 (Replication Exercise)

1. Reproduce Tables A1a and A1b. The authors decided to show only means and proportions. You should produce tables that also show the standard deviation for each variable (show it in parentheses below or next to the mean). You also need to do some research and learn how to calculate the indexes of dissimilarity at the second to last row. Use stargazer to automate the process of generating the different columns of the table.

2. Make a similar plot as Figure 1 to describe the trends in average weekly earnings by gender and race.

* For more detailed instructions, please refer to the file "Replication Project Instructions" (Pages 2 – 3).