Problem Set 3

SOC-GA 2332 Intro to Stats (Spring 2021)

Due: Saturday, Apr. 17th, 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.

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

# Load packages here
```

Part 1 Assumptions of OLS Regression

Recall that in our first lecture on regression, we talked about the **Gauss-Markov Assumptions**. If all these assumptions are met, the OLS estimator is the **Best Linear Unbiased Estimator** (**BLUE**). In a simple bivariate case, if the "true" data-generating process is $Y = \beta_0 + \beta_1 X + \epsilon$. The Gauss-Markov Assumptions can be state as the following:

(a) **Linearity**: A linear relationship between X and Y hold in the sample.

(b) **Exogeneity of Predictors**: The conditional mean of the error term, given the predictor, is zero ($\mathbf{x} = [x_1, x_2, ..., x_n]^\top$ is the value vector of X):

$$E[\epsilon_i|\mathbf{x}] = 0, \text{ for all } i = 1, 2, ..., n.$$

(c) **No Perfect Collinearity**: Explanatory variables cannot be perfectly correlated.

(d) **Homoskedasticity**:

- No Heteroskedasticity: The conditional variance of the error term, given the predictor, is constant: $Var[\epsilon_i|\mathbf{x}] = \sigma^2$, for all $i = 1, 2, ..., n$.

- No Autocorrelation: Conditional on the predictor, the error terms are uncorrelated across the observations: $Cov[\epsilon_i, \epsilon_j|\mathbf{x}] = 0$, $i \neq j$.

1. [15pts] For each of the assumptions, discuss what will go wrong when the assumption is violated. Be brief in your answers. *Note*: In addition to class materials, you can learn more about these assumptions in the Wikipedia article on the Gauss–Markov theorem, particularly the "Gauss–Markov theorem as stated in econometrics" section. (You can skip all the mathematical proofs and remarks.)

[Your Answer Here]

2. [5pts] Let $\beta_0 = -0.25$, $\beta_1 = 1.2$, $X \sim \Gamma(5, 4)$, and $\epsilon \sim Normal(0, 1)$. Here, $\Gamma(\alpha, \psi)$ denotes the Gamma distribution with shape parameter α and rate parameter ψ. (You can search how to use R to simulate from this distribution.)
Simulate a dataset of size $n = 3,000$ from this process in which all of the assumptions you've discussed above hold. Estimate a OLS model and plot regression diagnostics of this model.

```
# Your code here
```

Bonus Question [10pts]: From assumption (a), (b) and (d), choose one assumption and simulate a data that violates that assumption (all other assumptions should be satisfied). Create a plot which illustrates how the violation of the assumption affects the regression results. This can be a scatterplot with both the "true" and "false" OLS lines, a sampling distribution of the OLS estimator (comparing your estimate model results with actual simulations), or anything that shows how the violation leads us to false decisions if we assume the assumption is true. (The point is to demonstrate a contrast between the "true" and the "false", not just diagnostics of the "false".)

When simulating data, you don't have to use the parameters set in the previous problem.

Hint: You can search how to use `+ stat_function()` to plot a nonlinear line when plotting with `ggplot()`, or search how to use the base R functions such as `plot()` and `curve()`.

```
# Your code here
```

Part 2 Causality

A study on COVID-19 constructed a "COVID risk factor" score based on the COVID infection rate of a given area (defined by zip code).

A researcher wants to estimate the effect of having a vaccination center in the area on that area's COVID risk factor score. She compiled a dataset that contains each area's COVID risk factor score and whether the area has a vaccination center. She then estimated the effect of having a vaccination center using the "naive estimator" we discussed in class.

You noted that the quality of information residents have about COVID and the vaccine can be a confounding variable that affects both the area's infection rate and whether there is a vaccination center in the area. Assume that you are able to estimate the relationships this "informedness" confounder (`info`) and the original "vaccination center" predictor (`vaccine`) have with the COVID risk factor score (`covid_risk`), which can be simulated using the following code (`n` is sample size):

```
set.seed(1234) # set the same seed to ensure identical results
e = rnorm(n, 0, 0.5)
covid_risk = rescale( 0 - 7*vaccine - 2*info + e, to = c(0, 100))
```

1. [5pts] Import the data `covid.csv`, according to the counterfactual framework, constructing a counterfactual "risk factor" in the dataframe.

```
# Your code here
```

2. [10pts] Fill out the table below (round to 1 decimal points):

Group	Y^T	Y^C
Treatment Group ($D = 1$)	$E[Y^T \mid D = 1] = ?$	$E[Y^C \mid D = 1] = ?$
Control Group ($D = 0$)	$E[Y^T \mid D = 0] = ?$	$E[Y^C \mid D = 0] = ?$

3. [15pts]Estimate the following:

(a) The Naive Estimator of ATE

(b) Treatment Effect on the Treated

(c) Treatment Effect on the Control

(d) Selection Bias

4. [15pts] Write a non-technical, short summary reporting your results in response to the above mentioned researcher who used the naive estimation. Imagine that you are explaining this to an audience who may not be familiar with the specific terminologies of the counterfactual framework (such as ATE or Treatment Effect on the Treated), but is interested in your substantive findings.

[Your Answer Here]

Part 3 Linear Probability Model and Logistic Regression

`admin.csv` contains a dataset of graduate school admission results with the following variables:

Variable Name	Variable Detail
`admit`	Admission Dummy (Admitted is 1)
`gre`	GRE score
`gpa`	GPA
`rank`	Institution Tier (Tier 1 to 4)

1. [10pts] Import `admin.csv` to your R environment. Estimate (a) a linear probability model and (b) a logistic regression model to predict the probability of being admitted based on the applicant's GRE, GPA, and institution tier. Display the two modeling results in a table.

```
# Your code here
```

2. [10pts] In one or two paragraphs, summarize your modeling result for each model.

[Your Answer Here]

3. [15pts] Plot the predicted probability of admission based on one's GPA percentile and institution rank (holding GRE at the mean) for the logistic regression model. For the purpose of this exercise, please set the value of `gpa` to range from 1 to 4. Make sure to add appropriate title and labels to your figure.

```
# Your code here
```

Part 4 (Not Graded) Final Replication Project

At this point, you should complete most of the data cleaning and start replicating the descriptive tables and figure. You can submit an additional PDF file if you have made progress in replication Table A1a, Table A1b, and Figure 1.