**Assignment 1 - Identifying causality using experiments**

*For both questions, create a pdf using R Markdown that contains the code that you have run and the conclusions you arrive at after running your code. You may do the assignment in groups of up to two members. Each group should only submit one copy of the assignment. Make sure you put the names and ID numbers of both students on the top of the assignment.*

Q1) Platforms use various methods to stimulate user’s content creation. This includes paying users for reviews and providing awards and badges to users. Reddit is one of the largest platforms for creating and sharing content. On Reddit, users can recognize other contributions by providing gold to each other. However, does getting Reddit gold actually increase the receiver’s content generation?

To find out, researchers gave 905 random users reddit gold. Data is included for a similar number of users in the control group who did not receive gold during the time of the experiment. Import the data and examine:

1. If the control and treatment groups are similar across *tenure, premium\_user, and num\_posts\_before*  metrics.
2. Does getting reddit gold increase likelihood that the user will post (use the *posted* metric as the dependent variable and treated as the independent variable)? Use a simple linear model (not a logit) for the analysis.
3. What sorts of users are more likely to increase their contribution? (use the *tenure* and the *first\_timer* variables)
4. Is the SUTVA assumption likely to be violated in the experiment?

Q2) In 2019, Esther Duflo and Abhijeet Banerjee won the Nobel Prize in Economics for their research on experiments on education and poverty. In one of their experiments, they aimed to increase the academic performance of children in public schools in Vadodara (a town in India).

Duflo and her co-authors examined the impact of the Balsakhi program. In the program, the weakest academic students in Grade 3 were pulled out of their classroom and provided with supplementary classes, during school hours, provided by a Balsakhi, a young woman from the community who would work with the children on basic skills. Schools that did not receive the program formed the comparison group.

Data is provided for the period prior to the introduction of Balsakhis. This is known as the pre-period. Data is provided for math and language tests.

Using the data provided:

1. Use a t-test to see if there is a statistical difference in the pre-period between schools in the treatment (bal = 1) and control (bal = 0). This will check if randomization has been done correctly. To do this, calculate the average normalized test score(*norm*) for the pre period (pre = 1) for math (test\_type = 0). Is there a statistical difference between students who got the Balsakhi program and did not get the program? Perform the same test for language (test\_type = 1).
2. Calculate the average test scores for the post period (post = 1) for math for treatment and control. Is there a statistical difference between students in the two groups of schools? Use a t-test model to test the increase. Perform the same analysis for language test scores.
3. Can you conclude if the Balsakhi program increase test scores in reading and mathematics?
4. Is the SUTVA assumption violated in the example?

Column description for question 1

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| user\_id | ID | user id |
| intervention\_post\_score | Control | Net votes of post created by user as of the time the treatment is assigned |
| tenure | Control | The number of days elapsed after user registered on Reddit |
| premium\_user | Control | Dummy variable, one if user is a premium user, zero otherwise |
| comment\_karma | Control | Comment karma represents user's cumulative comment score on Redit |
| link\_karma | Control | Link karma represents user’s cumulative post score on Reddit |
| num\_post\_before | Control | The number of posts created by user during the two weeks prior to the experiment |
| num\_post\_before\_ss | Control | The number of posts created by user under the treated subreddit during the two weeks prior to the experiment |
| num\_sub\_before | Control | The number of subreddits user's posts content on during the two weeks prior to the experiment |
| first\_timer | Control | users who were posting to their intervention subreddit for the first time. |
| treated | Treatment | binary variable marking treated users |
| posted | Outcome | a binary indicator of whether users created any new posts during the two weeks after the treatment |
| posted\_ss | Outcome | a binary indicator of whether users created any new posts in the same sub-Redit during the two weeks after the treatment |
| posted\_oths | Outcome | a binary indicator of whether users created any new posts in the other sub-Redit during the two weeks after the treatment |
| n\_posts | Outcome | a count measure of the number of posts generated by a user over the two weeks following the experimental intervention |
| l\_n\_posts | Outcome | a count measure of the Ln (number of posts) generated by a user over the two weeks following the experimental intervention |
| n\_posts\_ss | Outcome | a count measure of the number of posts in the same sub-Redit generated by a user over the two weeks following the experimental intervention |
| l\_n\_posts\_ss | Outcome | a count measure of the Ln (number of posts in the same sub-Redit) generated by a user over the two weeks following the experimental intervention |
| n\_posts\_oths | Outcome | a count measure of the number of posts in the other sub-Redits generated by a user over the two weeks following the experimental intervention |
| l\_length | Outcome | average Ln (length) of posts in 2 weeks after the experiment |
| l\_length\_ss | Outcome | average Ln (length) of posts in the same sub-Redit in 2 weeks after the experiment |
| l\_length\_oths | Outcome | average Ln (length) of posts in other sub-Redits in 2 weeks after the experiment |