Assignment 1

Avnee Satija - 5992424 Harshal Sable - 5949697

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Question 1) Reddit a) If the control and treatment groups are similar across tenure, premium_user, and num_posts_before metrics

```r
data_Q1 <- read.csv("~/Desktop/Semester 2/Casual Inference/Homework/hw1/data_Q1.csv", na.strings = c(""
knitr::opts_chunk$set(echo = TRUE)
t.test(tenure ~ treated, data = data_Q1)
```

```
##
##  Welch Two Sample t-test
##
## data:  tenure by treated
## t = 1.373, df = 1789.6, p-value = 0.1699
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -19.09774 108.23144
## sample estimates:
## mean in group 0 mean in group 1
##        572.1680        527.6011
```

Analysis for Number of Posts before and Treated - H0: True difference in means of between treated and control groups is equal to 0 for number of posts before. H1: True difference in means between treated and control groups is not equal to 0 for number of posts before. Based on p-value of 00.1699 - which is fairly highly, so, we fail to reject H0. Hence, there's no difference between treated and non-treated number of posts before groups.

```r
knitr::opts_chunk$set(echo = TRUE)
t.test(premium_user ~ treated, data = data_Q1)
```

```
##
##  Welch Two Sample t-test
##
## data:  premium_user by treated
## t = 0.95906, df = 1769.9, p-value = 0.3377
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.006928414  0.020188082
## sample estimates:
## mean in group 0 mean in group 1
##      0.02541436      0.01878453
```

Analysis for Premium Users and Treated - H0: True difference in means of between treated and control groups is equal to 0 for premium users group. H1: True difference in means between treated and control groups is not equal to 0 for premium group. Based on p-value of 0.3377 - which is fairly highly, so, we fail to reject H0. Hence, there's no difference between treated and non-treated premium user groups.

```
knitr::opts_chunk$set(echo = TRUE)
t.test(num_post_before ~ treated, data = data_Q1)
```

```
##
##  Welch Two Sample t-test
##
## data:  num_post_before by treated
## t = 0.56253, df = 1796.1, p-value = 0.5738
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.2307971  0.4164325
## sample estimates:
## mean in group 0 mean in group 1
##        1.643094        1.550276
```

Analysis for Number of Posts before and Treated - H0: True difference in means of between group 0 and group 1 is equal to 0 for number of posts before. H1: True difference in means between group 0 and group 1 is not equal to 0 for number of posts before. Based on p-value of 0.5738 - which is fairly highly, so, we fail to reject H0. Hence, there's no difference between treated and non-treated number of posts before groups.

b) b) 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.

```
knitr::opts_chunk$set(echo = TRUE)
reg = lm(posted ~ treated , data = data_Q1)
summary (reg)
```

```
##
## Call:
## lm(formula = posted ~ treated, data = data_Q1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6232 -0.5602  0.3768  0.4398  0.4398
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56022    0.01631   34.34   <2e-16 ***
## treated      0.06298    0.02307    2.73   0.0064 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4908 on 1808 degrees of freedom
## Multiple R-squared:  0.004105,   Adjusted R-squared:  0.003554
## F-statistic: 7.452 on 1 and 1808 DF,  p-value: 0.006396
```

Posted = B0 + B1Treat + Error H0: There is a no linear relationship between posted and treated, B1 = 0. H1: There is a linear relationship between posted and treated, B1 != 0. The p-value is 0.0064 which low. So, we reject the H0, and conclude there is a linear relationship between posted and treated, B1 != 0.

c) What sorts of users are more likely to increase their contribution? (use the tenure and the first_timer variables)

```r
knitr::opts_chunk$set(echo = TRUE)
model_1 <- lm(posted ~ tenure + first_timer + treated, data = data_Q1)
summary(model_1)
```

```
##
## Call:
## lm(formula = posted ~ tenure + first_timer + treated, data = data_Q1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6967 -0.5549  0.3336  0.4070  0.5656
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.332e-01  2.399e-02  26.396  < 2e-16 ***
## tenure      -4.143e-05  1.714e-05  -2.417  0.01576 *
## first_timer -9.348e-02  2.374e-02  -3.937 8.56e-05 ***
## treated      6.351e-02  2.299e-02   2.763  0.00579 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4886 on 1806 degrees of freedom
## Multiple R-squared:  0.01383,    Adjusted R-squared:  0.01219
## F-statistic: 8.441 on 3 and 1806 DF,  p-value: 1.434e-05
```

```r
model_2 <- lm(posted ~ first_timer * treated, data = data_Q1)
summary(model_2)
```

```
##
## Call:
## lm(formula = posted ~ first_timer * treated, data = data_Q1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -0.6370 -0.6120  0.3630  0.3880  0.5031
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.630841   0.023621  26.706  < 2e-16 ***
## first_timer         -0.133986   0.032536  -4.118 3.99e-05 ***
## treated              0.006196   0.033877   0.183   0.8549
## first_timer:treated  0.108949   0.046107   2.363   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4887 on 1806 degrees of freedom
## Multiple R-squared:  0.01369,    Adjusted R-squared:  0.01205
## F-statistic: 8.354 on 3 and 1806 DF,  p-value: 1.623e-05
```

We looked at linear relation between posted (dependent variable) and tenure, first timer and treated (independent variables). We found that tenure has a coefficient of -4.143e-05 with the p-value 0.01576. This makes it statistically insignificant so, we drop it from our model and focus on the interaction relationship between first timer and treated.

For Model 2- H0: There is no interactive relationship between first timer and treated. B3 = 0 H1: There is an interactive relationship between first timer and treated. B3 != 0

We find B3 to be 0.108949 with a p-value of 0.0182. This indicates that the posting effect of being a first_timer is increased by 0.108949 for users who are in the treated group. A p-value of 0.0182 indicates that the relationship is statistically significant.

d) Is the SUTVA assumption likely to be violated in the experiment? The SUTVA assumption does not appear to be violated in this experiment. There is no evidence that receiving gold does influences the posting behavior of non-gold receiving users.

Question 2) Balsakhi program

a) 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).

```r
knitr::opts_chunk$set(echo = TRUE)
data_Q2 <- read.csv("~/Desktop/Semester 2/Casual Inference/Homework/hw1/data_Q2.csv", na.strings = c(""
# Subset for pre-period data
pre_data <- data_Q2 %>% filter (pre == 1)
```

For Math Subset->

```r
#Aggregate Averages for Math
avg_norm_math <- pre_data %>%
  filter(test_type == 0) %>%
  group_by(schoolid, bal) %>%  # Group by school and treatment status
  summarise(avgnorm = mean(norm, na.rm = TRUE), .groups = "drop")
print(avg_norm_math)
```

```
## # A tibble: 193 x 3
##    schoolid   bal avgnorm
##       <int> <int>   <dbl>
##  1      101     0 -0.210
##  2      101     1 -0.206
##  3      103     0 -0.0925
##  4      103     1  0.187
##  5      107     0 -0.273
##  6      107     1 -0.382
##  7      108     0  0.375
##  8      108     1  0.315
##  9      113     0  0.423
## 10      113     1 -0.386
## # i 183 more rows
```

```
# T-test on aggregate averages
t_test_math_pre <- t.test(avgnorm ~ bal, data = avg_norm_math)
print(t_test_math_pre)
```

```
##
##  Welch Two Sample t-test
##
## data:  avgnorm by bal
## t = 0.41122, df = 189.77, p-value = 0.6814
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.09537582  0.14561544
## sample estimates:
## mean in group 0 mean in group 1
##      0.01371641     -0.01140340
```

Analysis for treatment and control groups, for the subset math before Balsakhi was introduced - H0: True difference in means of between treatment and control is equal to 0 for the math subset, before Balsakhi was introduced. H1: True difference in means of between treatment and control is not equal to 0 for the math subset, before Balsakhi was introduced. Based on p-value of 0.6814 - which is fairly highly, so, we fail to reject H0. Hence, there's no difference between treatment and control groups, for the subset language, before Balsakhi was introduced.

For Language Subset->

```
knitr::opts_chunk$set(echo = TRUE)
#Aggregate Averages for Language
avg_norm_language <- pre_data %>%
  filter(test_type == 1) %>%
  group_by(schoolid, bal) %>%
  summarise(avgnorm = mean(norm, na.rm = TRUE), .groups = "drop")
print(avg_norm_language)
```

```
## # A tibble: 193 x 3
##    schoolid   bal avgnorm
##       <int> <int>   <dbl>
##  1      101     0  -0.268
##  2      101     1  -0.104
##  3      103     0  -0.226
##  4      103     1   0.438
##  5      107     0  -0.494
##  6      107     1  -0.272
##  7      108     0   0.375
##  8      108     1   0.494
##  9      113     0   0.431
## 10      113     1  -0.297
## # i 183 more rows
```

```r
# T-test on the aggregated averages
t_test_language_pre <- t.test(avgnorm ~ bal, data = avg_norm_language)
print(t_test_language_pre)
```

```
##
##  Welch Two Sample t-test
##
## data:  avgnorm by bal
## t = -0.66848, df = 186.31, p-value = 0.5047
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.15970007  0.07886297
## sample estimates:
## mean in group 0 mean in group 1
##    -0.009666828     0.030751723
```

Analysis for treatment and control groups, for the subset language before Balsakhi was introduced - H0: True difference in means of between treatment and control is equal to 0 for the language subset, before Balsakhi was introduced. H1: True difference in means of between treatment and control is not equal to 0 for the language subset, before Balsakhi was introduced. Based on p-value of 0.5047 - which is high, so, we fail to reject H0. Hence, there's no difference between treatment and control groups, for the subset language, before Balsakhi was introduced.

b) 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.

For Math Subset ->

```r
knitr::opts_chunk$set(echo = TRUE)
# Subset for post-period data
post_data <- subset(data_Q2, post == 1)

# Filter for math test (test_type = 0)
math_post_data <- subset(post_data, test_type == 0)

# Calculate average test scores for treatment and control groups
avg_math_scores <- aggregate(test ~ bal, data = math_post_data, mean)

print(avg_math_scores)
```

```
##   bal     test
## 1   0 19.78144
## 2   1 21.46939
```

```r
#t-test for math group
math_post_data <- post_data %>% filter (test_type == 0)
t_test_math_post <- t.test(test ~ bal, data = math_post_data)
print("T-Test for Math for Post Period:")
```

```
## [1] "T-Test for Math for Post Period:"
```

```r
print(t_test_math_post)
```

```
##
##  Welch Two Sample t-test
##
## data:  test by bal
## t = -5.807, df = 8391.7, p-value = 6.591e-09
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.257751 -1.118161
## sample estimates:
## mean in group 0 mean in group 1
##        19.78144        21.46939
```

Analysis for treatment and control groups, for the subset math after Balsakhi was introduced - H0: True difference in means of between treatment and control is equal to 0 for the math subset, after Balsakhi was introduced. H1: True difference in means of between treatment and control is not equal to 0 for the math subset, after Balsakhi was introduced. Based on p-value of 6.591e-09 - which is low, so, we reject the H0. Hence, there's a difference between treatment and control groups, for the subset math after Balsakhi was introduced.

For Language Subset ->

```r
knitr::opts_chunk$set(echo = TRUE)
# Filter for language test (test_type = 0)
lang_post_data <- subset(post_data, test_type == 1)

# Calculate average test scores for treatment and control groups
avg_lang_scores <- aggregate(test ~ bal, data = lang_post_data, mean)

# Display the result
print(avg_lang_scores)
```

```
##   bal     test
## 1   0 21.09880
## 2   1 22.11557
```

```r
knitr::opts_chunk$set(echo = TRUE)
#t-test for language group
lang_post_data <- post_data %>% filter (test_type == 1)
t_test_lang_post <- t.test(test ~ bal, data = lang_post_data)
print("T-Test for Language for Post Period:")
```

```
## [1] "T-Test for Language for Post Period:"
```

```r
print(t_test_lang_post)
```

```
##
##  Welch Two Sample t-test
##
## data:  test by bal
## t = -3.773, df = 8407.6, p-value = 0.0001624
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.5450260 -0.4885151
## sample estimates:
## mean in group 0 mean in group 1
##        21.09880        22.11557
```

Analysis for treatment and control groups, for the subset language after Balsakhi was introduced - H0: True difference in means of between treatment and control is equal to 0 for the language subset, after Balsakhi was introduced. H1: True difference in means of between treatment and control is not equal to 0 for the language subset, after Balsakhi was introduced. Based on p-value of 0.0001624 - which is low, so, we reject the H0. Hence, there's a difference between treatment and control groups, for the subset language after Balsakhi was introduced.

c) Can you conclude if the Balsakhi program increase test scores in reading and mathematics?

The Balsakhi program has a statistically significant effect on test scores in both mathematics and language. In both subjects, students who received the Balsakhi program (treatment group) showed significantly different test scores compared to students who did not receive the program (control group) in the post-period.

In the pre-period, there was no significant difference between the groups, indicating that the randomization was done correctly sub-part (a).

Thus, based on the t-test results, we can conclude that the Balsakhi program increased test scores in both language and mathematics after its implementation.

d) Is the SUTVA assumption violated in the example?

The SUTVA assumption was likely violated in this experiment. That there is high probability of indirect spillover effects occurring from treated group to control group. For example, students who did not receive the Balsakhi program might have learned indirectly from treated students, or heard about the supplementary lessons, potentially affecting their own academic performance.