UtkarshJoshi_Assigment1

Utkarsh Joshi

2025-02-12

SUBMISSION BY - UTKARSH JOSHI UMN ID - 5982808

```
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
```

```
reddit_data <- read.table("C:\\Users\\91788\\Downloads\\data_Q1.csv", header = TRUE, sep = ",")
```

Question 1 -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:

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

```
t.test(tenure ~ treated, data = reddit_data)
```

```
##
##  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
```

```r
t.test(premium_user ~ treated, data = reddit_data)
```

```
##
##  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
```

```r
t.test(num_post_before ~ treated, data = reddit_data)
```

```
##
##  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
```

Hypotheis for Tenure and Treated Variables ->

Null Hypothesis (H0): There is no significant difference in the mean number of posts before treatment between the treated and control groups. Alternative Hypothesis (H1): There is a significant difference in the mean number of posts before treatment between the treated and control groups.

Since the p-value is 0.1699, which is relatively high, we do not have enough evidence to reject the null hypothesis. This suggests that there is no significant difference in the number of posts before treatment between the treated and control groups.

Hypotheis for Premimum Users and Treated Variables ->

Null Hypothesis (H0): There is no significant difference in the mean values between the treated and control groups for premium users. Alternative Hypothesis (H1): There is a significant difference in the mean values between the treated and control groups for premium users.

Given that the p-value is 0.3377, which is relatively high, we do not have sufficient evidence to reject the null hypothesis. This indicates that there is no significant difference between the treated and non-treated premium user groups.

Hypotheis for Number of Posts and Treated Variables ->

Null Hypothesis (H0): There is no significant difference in the mean number of posts before treatment between the treated and control groups. Alternative Hypothesis (H1): There is a significant difference in the mean number of posts before treatment between the treated and control groups.

With a p-value of 0.5738, which is quite high, we do not have sufficient evidence to reject the null hypothesis. This suggests that there is no significant difference in the number of posts before treatment between the treated and non-treated groups.

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.

```
model = lm(posted ~ treated , data = reddit_data)
summary (model)
```

```
##
## Call:
## lm(formula = posted ~ treated, data = reddit_data)
##
## 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
```

Analysis of the Relationship Between Posted and Treated: Model: Posted = B0 + B1Treat + Error Null Hypothesis (H0): There is no linear relationship between "posted" and "treated," meaning B1 = 0 Alternative Hypothesis (H1): There is a significant linear relationship between "posted" and "treated," meaning B1 != 0.

With a p-value of 0.006396, we reject the null hypothesis. This indicates that there is a statistically significant linear relationship between "posted" and "treated," confirming that B1 != 0.

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

```
linear_model <- lm(posted ~ tenure + first_timer + treated, data = reddit_data)
summary(linear_model)
```

```
##
## Call:
## lm(formula = posted ~ tenure + first_timer + treated, data = reddit_data)
##
## 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
```

```
interaction_model <- lm(posted ~ first_timer * treated, data = reddit_data)
summary(interaction_model)
```

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

Analysis of the Linear Relationship Between Posted and Independent Variables We examined the linear relationship between posted (dependent variable) and tenure, first timer, and treated (independent variables). The coefficient for tenure was found to be -4.143e-05 with a p-value of 0.01576, indicating it is not statistically significant. As a result, we excluded tenure from the model and instead focused on the interaction effect between first timer and treated.

Interaction Between First Timer and Treated Null Hypothesis (H0): There is no interaction effect between first timer and treated (B3 = 0)

Alternative Hypothesis (H1): There is a significant interaction effect between first timer and treated (B3 != 0).

The interaction term (B3) was found to be 0.108949 with a p-value of 0.0182, suggesting that the effect of being a first-time poster is increased by 0.108949 for users in the treated group. Since the p-value is statistically significant, we conclude that there is a meaningful interaction effect between first timer and treated on posting behavior.

d) Is the SUTVA assumption likely to be violated in the experiment?

The Stable Unit Treatment Value Assumption (SUTVA) does not appear to be violated in this experiment. There is no evidence suggesting that receiving gold influences the posting behavior of users who did not receive gold. Therefore, the assumption holds in this context.

---------- QUESTION 2------------

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:

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).

```
balsakhi_data <- read.table("C:\\Users\\91788\\Downloads\\data_Q2.csv", header = TRUE, sep = ",")

preperiod_data <- balsakhi_data %>% filter (pre == 1)

#For Math ->
preperiod_math <- subset(preperiod_data, test_type == 0)
math_avgerage <- aggregate(norm ~ bal, data = preperiod_math, mean)
math_avgerage
```

```
##   bal          norm
## 1   0 -6.854839e-09
## 2   1 -6.825465e-03
```

```
# T-test Preperiod for Math group
preperiod_math <- preperiod_data %>% filter (test_type == 0)
t_test_preperiod_math <- t.test(norm ~ bal, data = preperiod_math)
print(t_test_preperiod_math)
```

```
##
##  Welch Two Sample t-test
##
## data:  norm by bal
## t = 0.34151, df = 10159, p-value = 0.7327
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.03235140  0.04600232
## sample estimates:
## mean in group 0 mean in group 1
##   -6.854839e-09   -6.825465e-03
```

```
# T-test Preperiod for Language group
preperiod_lang <- preperiod_data %>% filter (test_type == 1)
t_test_preperiod_lang <- t.test(norm ~ bal, data = preperiod_lang)
print(t_test_preperiod_lang)
```

```
##
##  Welch Two Sample t-test
##
## data:  norm by bal
## t = -1.2176, df = 10140, p-value = 0.2234
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.06397876  0.01495078
## sample estimates:
## mean in group 0 mean in group 1
##   -1.313364e-08    2.451397e-02
```

Analysis of Treatment and Control Groups Before Balsakhi Was Introduced

Math Language -> Null Hypothesis (H0): There is no significant difference in the mean scores between the treatment and control groups for the math subset before Balsakhi was introduced (true difference = 0).
Alternative Hypothesis (H1): There is a significant difference in the mean scores between the treatment and control groups for the math subset before Balsakhi was introduced (true difference != 0).

With a p-value of 0.7327, we fail to reject the null hypothesis. This indicates that there is no significant difference between the treatment and control groups in the math subset before the introduction of Balsakhi.

Language Language -> Null Hypothesis (H0): There is no significant difference in the mean scores between the treatment and control groups for the language subset before Balsakhi was introduced (true difference = 0).
Alternative Hypothesis (H): There is a significant difference in the mean scores between the treatment and control groups for the language subset before Balsakhi was introduced (true difference != 0).

With a p-value of 0.2234, which is still relatively high, we fail to reject the null hypothesis. This suggests that there is no significant difference between the treatment and control groups in the language subset before the introduction of Balsakhi.

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.

```
postperiod_data <- subset(balsakhi_data, post == 1)
#For Math ->
postperiod_math <- subset(postperiod_data, test_type == 0)
math_avgerage <- aggregate(test ~ bal, data = postperiod_math, mean)
math_avgerage
```

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

```
#Math group t-test
postperiod_math <- postperiod_data %>% filter (test_type == 0)
postperiod_math_t_test <- t.test(test ~ bal, data = postperiod_math)
postperiod_math_t_test
```

```
##
##  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
```

```
#For Language ->
postperiod_lang <- balsakhi_data %>% filter (post == 1)
lang_avgerage <- aggregate(test ~ bal, data = postperiod_lang, mean)
lang_avgerage
```

```
##   bal     test
## 1   0 27.25349
## 2   1 29.05664
```

```
# Filter for language test (test_type = 0)
postperiod_lang <- subset(postperiod_data, test_type == 1)
lang_avgerage <- aggregate(test ~ bal, data = postperiod_lang, mean)
lang_avgerage
```

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

```
#Language group t-test
postperiod_lang <- postperiod_data %>% filter (test_type == 1)
postperiod_lang_t_test <- t.test(test ~ bal, data = postperiod_lang)
postperiod_lang_t_test
```

```
##
##  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 of Treatment and Control Groups After Balsakhi Was Introduced

Math Subset:
Null Hypothesis (H0): There is no significant difference in the mean scores between the treatment and control groups for the math subset after Balsakhi was introduced (true difference = 0).
Alternative Hypothesis (H1): There is a significant difference in the mean scores between the treatment and control groups for the math subset after Balsakhi was introduced (true difference is not 0).

With a p-value of 6.591e-09, we reject the null hypothesis. This indicates that there is a significant difference between the treatment and control groups for the math subset after Balsakhi was introduced.

Language Subset:
Null Hypothesis (H0): There is no significant difference in the mean scores between the treatment and control groups for the language subset after Balsakhi was introduced (true difference = 0).
Alternative Hypothesis (H1): There is a significant difference in the mean scores between the treatment and control groups for the language subset after Balsakhi was introduced (true difference is not 0).

With a p-value of 0.0001624, we reject the null hypothesis. This suggests that there is a significant difference between the treatment and control groups for the language subset after Balsakhi was introduced.

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

The results indicate that the Balsakhi program had a statistically significant impact on test scores in both mathematics and language. After the program was implemented, students in the treatment group showed significantly different test scores compared to those in the control group.Prior to the program (pre-period), there was no significant difference between the treatment and control groups, confirming that the randomization was conducted correctly, as established in part (a).

Therefore, based on the t-test results, we can conclude that the Balsakhi program led to an increase in test scores in both language and mathematics.

d) Is the SUTVA assumption violated in the example?

The Stable Unit Treatment Value Assumption (SUTVA) was likely violated in this experiment. There is a high probability of indirect spillover effects from the treated group to the control group. For instance, students who did not receive the Balsakhi program may have indirectly benefited by learning from their peers in the treatment group or by becoming aware of the supplementary lessons, which could have influenced their academic performance.