Homework 1

Grant Jackson

2024-09-08

Theoretical Questions

Question 1

i. **D__i is an indicator variable that equal 1 if a worker went to college; y__i is a variable measuring the worker's health.**

Should expect the average untreated potential outcomes for workers that went to college (treated) to be higher than workers who did not (control). Factors that lead people to attend college, such as geographic location, family income, and alcohol/drug use are likely to be associated with better health outcomes. We would find a positive bias in the difference-in-means estimator of the ATT, meaning an overestimate of the effect of college on health.

ii. **D__i is an indicator variable that equal 1 if the Amazon user was experimentally shown a banner of type 'A'; y__i is whether the user purchased the product**.

Due to this treatment being experimental, the randomization should balance out any pre-existing differences between the treated and control group. We should expect the average untreated potential outcomes to be similar for both the Amazon user groups. There should not be any bias in the difference-in-means estimator of the ATT.

iii. **D__i is a variable if a school district contains an arts program; y__i is the school district's average ELA score.**

The average untreated potential outcomes for school districts with art programs (treated) should be expected to be higher than school districts without one (control). Factors that lead to school districts to have an arts program, such as higher funding, school size, and state/local education policy are likely to be associated with higher average ELA scores. We would find a positive bias in the difference-in-means estimator of the ATT, meaning an overestimate of the effect of an arts program on school district test scores.

iv. **D__i is a variable that equals 1 if a census tract has a Dollar General in it; y__i measures the census tract's rate of obesity.**

The average untreated potential outcomes for the census tract with a Dollar General (treated) may be higher than those without one (control). Factors that lead Dollar General to open a store at a certain location, such as lower income and less access to healthy food options are also associated with higher obesity rates. We would find a positive bias in the difference-in-means estimator of the ATT, meaning an overestimate of the effect of a Dollar General store on obesity rates.

v. **D__i is a variable that equals 1 if a driver has tinted windows; y__i is a measure of "response time" that measures visual performance.**

The average untreated potential outcomes for drivers with tinted windows (treated) may be different for drivers without tinted windows (control). Drivers with tinted windows could be more safety-conscious or have better visual ability, leading to faster response times even without the treatment. We may find a negative bias in the difference-in-means estimator of ATT, meaning an underestimate of the effect of tinted windows on response time.

Question 2

A/B testing with a large number of arms and few trails per arm can lead to problems with inference for several reasons. Having few trails per arm means each arm has a small sample size, which can lead to unstable and unreliable estimates of treatment effects. With the many arms, you could run into the multiple comparisons problem, leading to a higher chance of finding a false positive or false negative result. To fix these issues as a tech firm, I would suggest several actions such as, increasing the sample size per arm, limiting the amount of variations being tested, or practice sequential testing.

Coding Exercise

```r
setwd("~/Applied Microeconometrics/Data")
data <- read.csv("national_jtpa_study.csv")
head(data)
```

```
##   foundjob treat age priorearn educ female nonwhite married
## 1        1     1  46         0   12      1        1       0
## 2        1     1  24      3591   11      1        0       0
## 3        1     1  28      6000    9      0        1       1
## 4        1     0  23      2000   11      0        0       1
## 5        1     1  34         0   12      1        1       1
## 6        1     1  31      9476   10      1        0       0
```

1. Assessing Randomization

Using regression to check if key variables are balanced between the treated and control groups, and assessing if the randomization was done properly:

```r
balance_vars <- c("age", "priorearn", "educ", "female", "nonwhite")

balance_tests <- lapply(balance_vars, function(var) {
  formula <- as.formula(paste(var, "~ treat"))
  model <- lm(formula, data = data)
  tidy(model)
})

balance_results <- do.call(rbind, balance_tests)
print(balance_results)
```

```
## # A tibble: 10 x 5
##    term        estimate std.error statistic p.value
##    <chr>          <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)   29.5      0.159     185.     0
##  2 treat         -0.279    0.193      -1.45   0.148
##  3 (Intercept) 2899.      60.2        48.1    0
##  4 treat       -114.      72.9        -1.57   0.117
##  5 (Intercept)   11.5      0.0270    426.     0
##  6 treat         -0.00650  0.0327     -0.199  0.842
##  7 (Intercept)    0.561    0.00748    75.0    0
##  8 treat         -0.00326  0.00906    -0.360  0.719
##  9 (Intercept)    0.435    0.00748    58.2    0
## 10 treat          0.00382  0.00905     0.422  0.673
```

We can confidently say that the treated and control groups “look the same” based on these results. None of the key demographic variables show statistically significant differences between the treated and control groups. This balance also provides strong evidence that the randomization was successful.

2. Difference-in-means Estimator “by hand”

Calculating difference-in-means estimator for the effect of JIPA training on the probability of finding a job:

```
treated_mean <- mean(data$foundjob[data$treat == 1])
control_mean <- mean(data$foundjob[data$treat == 0])
diff_in_means <- treated_mean - control_mean

cat("Difference-in-Means Estimator:", diff_in_means)
```

```
## Difference-in-Means Estimator: 0.0187493
```

3. Difference-in-means Estimator by regression

Calculating the difference-in-means estimator using regression and report standard errors:

```
model <- lm(foundjob ~ treat, data = data)
robust_se <- sqrt(diag(vcovHC(model, type = "HC1")))

results <- tibble(
  estimate = coef(model)["treat"],
  std_error = robust_se["treat"],
  t_value = estimate / std_error,
  p_value = 2 * pt(abs(t_value), df = nrow(data) - 2, lower.tail = FALSE)
)

print(results)
```

```
## # A tibble: 1 x 4
##   estimate std_error t_value p_value
##      <dbl>     <dbl>   <dbl>   <dbl>
## 1   0.0187   0.00676    2.77 0.00555
```

We can conclude that the estimate is statistically significant. The JTPA program appears to increase the job-finding rate by about 1.87%.