Quant2 Lab1 Exercise

2026-01-29

Simulated Potential Outcome Data

- Download the file `thescience.tsv` and `lab01_exercise.qmd` from this week's lab folder on GitHub
- Move the file to a "lab01" folder on your own computer
- Install the `pacman`, `tidyverse` and `here` R packages if you don't already have them

```
# install.packages(c('pacman','tidyverse', 'here'))
pacman::p_load(tidyverse, here)
df <- read_tsv(here('lab01/thescience.tsv'))
```

The data contains the following columns:

- Potential Outcomes: `y0` and `y1`
- Observed Outcome: `y`
- Treatment: `t`
- Covariates: `x_1`, `x_2`, `x_3`

```
df %>% head
```

```
# A tibble: 6 x 7
      x1      x2    x3     y0    y1     t     y
   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 0.288  -0.301   1.39 0.0284 0.350     1 0.350
2 0.788  -0.497   2.30 2.35   3.92      1 3.92
3 0.409   0.513   2.34 1.25   1.54      0 1.25
4 0.883   0.376  -1.57 4.53   5.00      1 5.00
5 0.940  -0.755  -3.05 4.13   5.90      1 5.90
6 0.0456 -0.0476  5.41 2.26   2.79      0 2.26
```

How does the observed outcome `y` relate to the treatment `t` and the potential outcomes `y0` and `y1`?

```
df %>% select(y0, y1, t, y) %>% tail
```

```
# A tibble: 6 x 4
     y0    y1     t     y
  <dbl> <dbl> <dbl> <dbl>
1 0.616 0.501     0 0.616
2 4.15  4.92      1 4.92
3 4.46  5.37      1 5.37
4 1.87  2.13      1 2.13
5 4.83  4.45      0 4.83
6 0.995 3.83      0 0.995
```

Calculate the difference in means between the treated and the untreated.

```
mean(filter(df, t == 1)$y) - mean(filter(df, t == 0)$y)
```

```
[1] 2.257254
```

Calculate the true global average treatment effect

```
mean(df$y1 - df$y0)
```

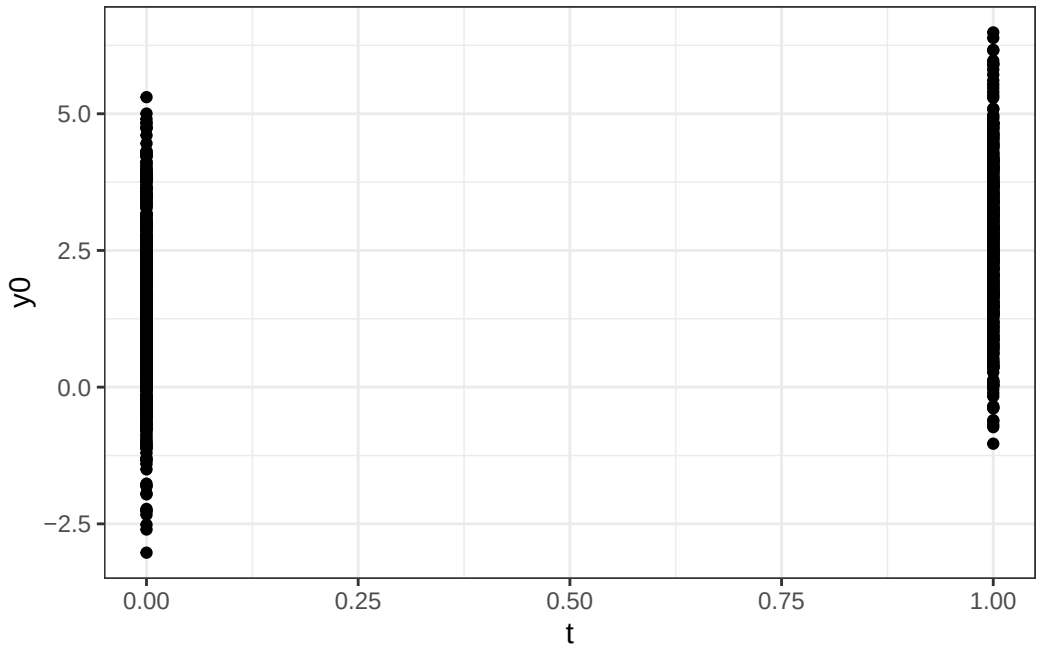
```
[1] 1.000603
```

Why they are different?

```
# Selection bias because there is a difference in the potential outcomes
# of the treatment vs. control
mean(filter(df, t == 1)$y0) - mean(filter(df, t == 0)$y0)
```

```
[1] 1.256952
```

```r
# We can see this graphically by plotting the data
df %>% ggplot(aes(x =t , y = y0)) + geom_point() + theme_bw()
```




```r
# Or we can observe this by noting that treatment and potential outcomes
# are correlated
cor(df$t, df$y0)
```

```
[1] 0.3990395
```

```r
# Selection bias wrt ATE?
rho_i <- df$y1 - df$y0
mean(rho_i[df$t == 1]) - mean(rho_i[df$t == 0])
```

```
[1] -0.0006006687
```

Exercise

Do the next part in pairs. Prepare your work using Quarto

What is the ATE vs the ATC and ATT?

How would we calculate these from the science?

```
# your code here
```

Which of x1, x2, and x3 are associated with treatment assignment? Are they potentially confounders for identifying the ATE?

```
# your code here
```

Fixing the ATE estimation

- You get to play omnipotent being! Create an alternate universe (ie, a new treatment assignment and new outcome variable) such that the difference in means between the treated and the untreated can be reliably estimated.
- Estimate the difference in means and compare it to the true effect.
- Are they different? Why/How?

```
# your code here
```