Treatment Effects Practical Session #1: Testing the LATE Model

Frank DiTraglia

Oxford Econometrics Summer School

Introduction

This practical session is based on Huber & Mellace (2015). You may find it helpful to consult the paper and or my lecture notes. See Hands-On Programming with R for a review of basic R that you will need below. My notes on this book are available here.

Exercises

1. Write a function to simulate `n` iid draws from the model given below, with arguments `n`, `alpha` and `beta`. Your function should return a data frame (or tibble) with named columns `D`, `Z`, and `Y`.

$$
\begin{aligned}
Y &= D + \beta Z + U \\
D &= 1\{\alpha Z + \epsilon > 0\} \\
\begin{bmatrix} U \\ \epsilon \end{bmatrix} &\sim \text{Normal}(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\
Z &\sim \text{Bernoulli}(0.5), \text{ indep. of } (U, \epsilon)
\end{aligned}
$$

2. Answer the following questions about the model from the preceding part.
   (a) Is the treatment D endogenous? How can you tell?
   (b) What is the distribution of treatment effects? What is the LATE in this model?
   (c) What is the role of β?
   (d) What is the role of α?
   (e) Which of the LATE assumptions does the model satisfy?

3. Write a function called `get_theta()` to compute the sample analogues of $\theta_1, \theta_2, \theta_3, \theta_4$ defined in Equation (7) of Huber & Mellace (2015). Your function should take a single input argument: a data frame (or tibble) with columns named `D`, `Z`, and `Y` corresponding to the model from above. It should return a vector with four named elements: `theta1`, `theta2`, `theta3`, and `theta4`.

4. Check your function from the preceding part by generating 100,000 observations from the model in part 1 with parameter values $\alpha = 0.6$ and $\beta = 1$. You should detect a violation of the LATE assumptions. Calculate the Wald estimand. Does it equal the LATE? Repeat for $\beta = 0$. How do you results change?

5. Repeat the preceding part for a variety of values of β until you find one for which the LATE assumptions are violated but you *cannot* detect a violation of the inequalities from the paper. Why is this possible?

6. Load the `wooldridge` dataset and read the documentation for the `card` dataset. Once you understand the contents of the dataset, carry out the following steps to construct a data frame (or tibble) called `card_dat`:

   (a) Define the instrument `Z` as a dummy variable for living near a 4-year college in 1966. (The idea here is that living near a college reduces your costs of attending in a way that doesn't affect wages.)
   (b) Define the outcome `Y` as the log of weekly earnings in 1976.
   (c) Construct the treatment `D` as a dummy variable that equals one if a person has completed 16 years of education or more by 1976. This is effectively a proxy for "has a four-year degree."

7. Apply your function `get_theta()` to `card_dat`. Do you detect any violations of the LATE model? Re-read the documentation for `card` to see if you can find any potential explanation for your results. Interpret the IV estimate for `card_dat` in light of this.

8. **Bonus Question:** If you found the preceding parts too easy, here's a challenge for you! We did not consider statistical significance when looking for a violation of the LATE model in the preceding part. Use the function `boot()` from the R package `boot`, along with your function `get_theta()` from above to implement the "simple bootstrap with Bonferroni adjustment" described on page 402 of Huber & Mellace (2015) and apply it to `card_dat`. Briefly discuss your findings.

Solutions

Problem 1

```
library(mvtnorm)
library(tidyverse)
draw_sims <- function(alpha, beta, n = 250) {
  Z <- rbinom(n, 1, 0.5)
  S <- matrix(c(1, 0.5,
                0.5, 1), 2, 2, byrow = TRUE)
  errors <- rmvnorm(n, sigma = S)
  U <- errors[,1]
  epsilon <- errors[,2]
  D <- 1 * (alpha * Z + epsilon > 0)
```

```
  Y <- D + beta * Z + U
  tibble(Z, D, Y)
}
get_IV <- function(dat) {
  cov(dat$Y, dat$Z) / cov(dat$D, dat$Z)
}
```

Problem 2

(a) The treatment D is indeed endogenous since it is correlated with the error term U, as we see from the variance-covariance matrix Σ.

(b) The parameter β controls the strength of the *direct effect* of the instrument Z on the outcome Y. Unless $\beta = 0$, mean exclusion is violated.

(c) The treatment effects in this model are in fact *homogenous* so their distribution is degenerate: $Y_1 - Y_0 = 1$ for everyone, so this is the LATE as well as the ATE! This isn't a problem: the LATE model allows but does not require homogeneous treatment effects.

(d) The parameter α determines the share of compliers, i.e. the share of people who will have $D = 1$ if and only if $Z = 1$. Since $D = 1\{\alpha Z + \epsilon > 0\}$, compliers are individuals with anyone with $-\alpha < \epsilon \leq 0$. Since $\epsilon \text{Normal}(0, 1)$, the share of compliers is $\texttt{pnorm}(0) - \texttt{pnorm}(-\alpha) = 0.5 - \texttt{pnorm}(-\alpha)$.

(e) It depends on the parameter values. The only two assumptions that may be violated are mean exclusion and the existence of compliers. As long as $\alpha \neq 0$, there are some compliers in the population. Mean exclusion is only satisfied if $\beta = 0$. Unconfounded type holds regardless of the values of α and β since ϵ is independent of Z. Monotonicity also holds regardless of the values of α and β because the first-stage take-up model shifts everyone in the same way: it is a threshold crossing model.

Problem 3

```
get_theta <- function(dat) {
  D <- dat$D
  Z <- dat$Z
  Y <- dat$Y

  q <- mean(D[Z == 0]) / mean(D[Z == 1])
  if(q > 1) q <- 1
  if(q < 0) q <- 0

  r <- mean(1 - D[Z == 1]) / mean(1 - D[Z == 0])
  if(r > 1) r <- 1
  if(r < 0) r <- 0

  Y11 <- Y[(D == 1) & (Z == 1)]
```

```
  y_q_bottom <- quantile(Y11, q)
  y_q_top <- quantile(Y11, 1 - q)

  Y00 <- Y[(D == 0) & (Z == 0)]
  y_r_bottom <- quantile(Y00, r)
  y_r_top <- quantile(Y00, 1 - r)

  mu_always <- mean(Y[(D == 1) & (Z == 0)])
  mu_never <- mean(Y[(D == 0) & (Z == 1)])

  return(c(theta1 = mean(Y11[Y11 <= y_q_bottom]) - mu_always,
           theta2 = mu_always - mean(Y11[Y11 >= y_q_top]),
           theta3 = mean(Y00[Y00 <= y_r_bottom]) - mu_never,
           theta4 = mu_never - mean(Y00[Y00 >= y_r_top])))
}
```

Problem 4

```
set.seed(12345)

get_IV <- function(dat) {
  D <- dat$D
  Z <- dat$Z
  Y <- dat$Y
  cov(Z, Y) / cov(Z, D)
}

sims0 <- draw_sims(alpha = 0.6, beta = 0, n = 1e5)
get_IV(sims0) # Should equal one and it does
```

```
## [1] 1.00287
```

```
get_theta(sims0) # Everything should be negative and everything is
```

```
##     theta1     theta2     theta3     theta4
## -0.6467863 -0.3120119 -0.4502310 -0.8680653
```

```
sims1 <- draw_sims(alpha = 0.6, beta = 1, n = 1e5)
get_IV(sims1) # Way too high!
```

```
## [1] 5.413024
```

```
get_theta(sims1) # A violation for both theta1 and theta4
```

```
##     theta1     theta2     theta3     theta4
```

```
##  0.3417739 -1.3105387 -1.4459594  0.1207921
```

Problem 5

```
sims2 <- draw_sims(alpha = 0.6, beta = 0.5, n = 1e5)
get_IV(sims2) # Way too high!
```

```
## [1] 3.230891
```

```
get_theta(sims2) # No violations
```

```
##     theta1     theta2     theta3     theta4
## -0.1460979 -0.8244632 -0.9774652 -0.3703117
```

Problem 6

```
library(dplyr)
library(wooldridge)
data(card) #?card for documentation
card_dat <- card %>%
  mutate(Y = lwage, Z = nearc4, D = 1 * (educ >= 16)) %>%
  select(Y, Z, D) %>%
  tibble()

# Check some values against Huber & Mellace
# Share of compliers matches the paper
card_dat %>%
  group_by(Z) %>%
  summarize(takeup = mean(D)) %>%
  pull(takeup) %>%
  diff()
```

```
## [1] 0.06856902
```

```
get_IV(card_dat)
```

```
## [1] 2.273731
```

The LATE estimate suggests an *insanely* high return to college: remember that the outcome here is on the log scale!

Problem 7

```
# Check for violations of the LATE model
get_theta(card_dat) # There seems to be a violation in theta4
```

```
##      theta1      theta2      theta3      theta4
## -0.09029760 -0.25126153 -0.23175056  0.09925123
```

In the documentation, we see that IQ is a crucial omitted variable and that it is correlated with the instrument. This may be partially responsible for the outrageously high LATE estimate.

Problem 8

```
library(boot)
get_boot_p <- function(dat) {
  boot_results <- boot(dat,
                       function(data, boot_rows) get_theta(data[boot_rows,]),
                       R = 499)
  theta_hat <- boot_results$t0
  theta_hat_boot <- boot_results$t
  colnames(theta_hat_boot) <- names(theta_hat)
  theta_tilde_boot <- sweep(theta_hat_boot, 2, theta_hat)
  colMeans(sweep(theta_tilde_boot, 2, theta_hat, FUN = ">"))
}

pvalues <- get_boot_p(card_dat)
pvalues
```

```
##    theta1    theta2    theta3    theta4
## 0.9759519 1.0000000 1.0000000 0.0000000
```

```
4 * min(pvalues) # Bonferroni correction
```

```
## [1] 0
```