Treatment Effects Practical Session #2: Gaussian MTEs

Frank DiTraglia

Oxford Econometrics Summer School

Introduction

This practical session is based on Heckman, Tobias, & Vytlacil (2001). You may find it helpful to consult the paper and or the lecture slides. 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.

Throughout this session, we will work with the following model:

$$
\begin{aligned}
Y &= (1 - D)Y_0 + DY_1 \\
D &= 1\{\gamma_0 + \gamma_1 Z > V\} \\
Y_0 &= \mu_0 + U_0 \\
Y_1 &= \mu_1 + U_1
\end{aligned}
$$

$$
\begin{bmatrix} V \\ U_0 \\ U_1 \end{bmatrix} \sim \text{Normal}(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \sigma_0\rho_0 & \sigma_1\rho_1 \\ & \sigma_0^2 & \sigma_{01} \\ & & \sigma_1^2 \end{bmatrix}
$$

$$
Z \sim \text{Bernoulli}(q), \text{ indep. of } (V, U_0, U_1)
$$

In real life, we would observe only (Y, D, Z) but in some of the exercises below we will also work with the unobserved variables (Y_0, Y_1, V) directly.

Exercises

1. Write a function to simulate n iid draws of (Y_0, Y_1, V) from the multivariate normal distribution described above, fixing $\mu_0 = \mu_1 = 0$, $\sigma_0 = \sigma_1 = 1$, and $\sigma_{01} = 1/2$. Your function should take three arguments—`n`, `rho0`, `rho1`—and return a data frame (or tibble) with named columns `Y0`, `Y1`, and `V`. You should *not* simulate draws for D or Z at this point; we'll do that in a later exercise. Note that there is no need to store (U_0, U_1) since they coincide with (Y_0, Y_1) when $\mu_0 = \mu_1 = 0$.

2. In the lecture slides we derived a number of analytical expressions for the model given above. These included:

$$\begin{aligned}
\text{TOT}(p) &= \frac{-(\sigma_1\rho_1 - \sigma_0\rho_0)\varphi(\Phi^{-1}(p))}{p} \\
\text{TUT}(p) &= \frac{(\sigma_1\rho_1 - \sigma_0\rho_0)\varphi(\Phi^{-1}(p))}{1-p} \\
\text{LATE}(p_0, p_1) &= -(\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{\varphi(\Phi^{-1}(p_1)) - \varphi(\Phi^{-1}(p_0))}{p_1 - p_0}\right]
\end{aligned}$$

where p denotes the fraction of (eligible) people treated under a hypothetical *status quo* policy, and (p_0, p_1) are the share of people who would take the treatment when $Z = 0$ and $Z = 1$, respectively.

   (a) Use your function from the preceding exercise, generate and store 100,000 simulation draws of (Y_0, Y_1, V) with `rho0 = 0.5`, `rho1 = 0.2`.
   (b) Based on your simulation draws from part (a), who would be more likely to take the treatment when $\rho_0 = 0.5$ and $\rho_1 = 0.2$: someone with a *high* treatment effect or someone with a *low* treatment effect?
   (c) Use your simulation draws from part (a) to numerically verify the expressions for $\text{TOT}(p)$, $\text{TUT}(p)$ and $\text{LATE}(p_0, p_1)$ for a few different values of p, p_0, p_1.
   (d) Repeat parts (a)–(c) with `rho0 = 0.3`, `rho1 = 0.4`. How do your results change? Explain briefly.

3. In exercise 1 you wrote a function that returns simulation draws of (Y_0, Y_1, V). In real life, however, we observe only (Y, D, Z). Write a new function that *builds on* your earlier one but returns a data frame (or tibble) with the observable quantities only: named columns `Y`, `D`, and `Z`. Your new function should take arguments `n`, `rho0`, `rho1`, `gamma0`, `gamma1` and `q` and fix the parameters μ_0, μ_1, σ_0, σ_1, and σ_{01} to the same values as in exercise 1 above.

4. In this exercise you will test your function from exercise 3 by calculating the same LATE in two different ways. These should agree up to simulation error.

   (a) How do (p_0, p_1) relate to (γ_0, γ_1) under the model? Explain briefly.
   (b) Under the model described above, how would you estimate the LATE based on a sample of n iid observations of (Y, D, Z)? Explain briefly.
   (c) Use your function from the exercise 3 to make and store 100,000 simulation draws with `q = 0.5`, `gamma0 = -1`, `gamma1 = 0.5`, `rho0 = 0.5`, and `rho1 = 0.2`. Use them to calculate the LATE two different ways: first using the analytical expression from exercise 2 above, and second using the estimator from part (b).

5. **Bonus Question:** We know how to estimate the LATE from observations of (Y, D, Z), but suppose we instead wanted to calculate the ATE, TOT, or TUT. Under the multivariate normal model described above, this is relatively straightforward. For this exercise, define the shorthand $\delta_0 \equiv \sigma_0\rho_0$ and $\delta_1 \equiv \sigma_1\rho_1$ and recall the following

expressions that we derived in the lecture slides:

$$\begin{aligned}
\mathbb{E}[Y|D=d,Z=z] &= \mu_d + \delta_d \mathbb{E}(V|D=d,Z=z); \quad d,z \in \{0,1\} \\
\mathbb{E}(V|D=1,Z=z) &= \frac{-\varphi(\gamma_0+\gamma_1 z)}{\Phi(\gamma_0+\gamma_1 z)} \\
\mathbb{E}(V|D=0,Z=z) &= \frac{\varphi(\gamma_0+\gamma_1 z)}{1-\Phi(\gamma_0+\gamma_1 z)}
\end{aligned}$$

(a) Propose a way of using observations of (D,Z) to estimate γ_0 and γ_1. Test your approach using the draws you made in exercise 4 above.

(b) By substituting your estimates of γ_0 and γ_1 from part (a), propose a way of using observations of (Y,D,Z) to estimate μ_0, μ_1, γ_0, and γ_1. Test your approach using the simulation draws you made in exercise 4 above.

(c) Based on your answers to parts (a) and (b), propose a way of estimating the ATE, TOT, and TUT from observations of (Y,D,Z) drawn from the model described above. Again, test your approach using the simulation draws from exercise 4.