Applied Class #5 - Gaussian MTE Model

Introduction

This practical session is based on Heckman, Tobias, & Vytlacil (2001) and Angrist (2004). Throughout this class, 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 an R function that uses `rmvnorm()` from the `mvtnorm` package to simulate n iid draws of (Y_0, Y_1, V, Z, D) from the multivariate normal distribution described above, fixing $\mu_0 = \mu_1 = 0$, $\sigma_0 = \sigma_1 = 1$, $\sigma_{01} = 1/2$, and $q = 1/2$. Your function should take five arguments—`n`, `rho0`, `rho1`, `gamma0`, and `gamma1`—and return a data frame with named columns `Y0`, `Y1`, `V`, `Z`, `D`, and `Y`. In real life we can only observe (Y, D, Z) but fortunately for us, simulations aren't real life! In this example 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. Use your function from the preceding part to make and store 1,000,000 simulation draws with $\rho_0 = 0.5$, $\rho_1 = 0.2$, $\gamma_0 = -1$ and $\gamma_1 = 1.5$. Use your simulation draws to calculate the LATE, TOT, and TUT at these parameter values.

3. In the previous exercise you used simulation to approximate the values of the LATE, TOT, and TUT at particular parameter values. Use the following formulas from the lecture slides to check your simulations against the *analytical* formulas that apply in the case where $q = 1/2$ and $\sigma_0 = \sigma_1 = 1$. Recall that we use the shorthand $\delta = (\rho_1 - \rho_0)$.

$$\text{LATE} = -\delta \left[\frac{\varphi(\gamma_0 + \gamma_1) - \varphi(\gamma_0)}{\Phi(\gamma_0 + \gamma_1) - \Phi(\gamma_0)} \right]$$

$$\text{TOT} = -\delta \left[\frac{\varphi(\gamma_0) + \varphi(\gamma_0 + \gamma_1)}{\Phi(\gamma_0) + \Phi(\gamma_0 + \gamma_1)} \right]$$

$$\text{TUT} = \delta \left[\frac{\varphi(\gamma_0) + \varphi(\gamma_0 + \gamma_1)}{\{1 - \Phi(\gamma_0)\} + \{1 - \Phi(\gamma_0 + \gamma_1)\}} \right]$$

4. Consult Section 2.1 of Angrist (2004). Angrist's notation is slightly different from ours: what he calls η we call V, what he calls ρ_{01} we call γ, and what he calls TT we call TOT. Other than that, everything is the same. Figure 1 of this paper plots the TOT and LATE over a range of values for $\mathbb{P}(D = 1|Z = 0)$. Use the formulas from the previous question to reproduce panel (a) of this figure. Add in the TUT effect for good measure if you're feeling ambitious! You'll need to read a few short extracts of the paper to determine how to set the parameters γ_0 and γ_1 when making your plot.

5. In this problem you will apply the Heckman two-step estimator to the simulated values of (Y, D, Z) from question 2 above to estimate the parameters μ_1, μ_1, $\delta_0 \equiv \sigma_0\rho_0$ and $\delta_1 \equiv \sigma_1\rho_1$. In the simulation we know that $\mu_1 = \mu_0 = 0$, $\sigma_0 = \sigma_1 = 1$, $\rho_0 = 0.5$ and $\rho_1 = 0.2$, so you'll check the estimator against these values. Follow these steps:
   a. Use the simulated values of (D, Z) to estimate (γ_0, γ_1). Call your estimates $(\widehat{\gamma}_0, \widehat{\gamma}_0)$. Check that your estimates match the true values that you used to generate the data: $\gamma_0 = -1$ and $\gamma_1 = 1.5$.
   b. Define the shorthand $\widehat{\lambda}(z) = \varphi(\widehat{\gamma}_0 + \widehat{\gamma}_1 z)/[1 - \Phi(\widehat{\gamma}_0 + \widehat{\gamma}_1 z)]$. Add a column called `lambda` to the dataframe containing your simulated values of (Y, D, Z) that evaluates the function $\widehat{\lambda}(\cdot)$ at the observed values of Z.
   c. Define the shorthand $\widehat{\kappa}(z) = -\varphi(\widehat{\gamma}_0 + \widehat{\gamma}_1 z)/\Phi(\widehat{\gamma}_0 + \widehat{\gamma}_1 z)$. Add a column called `kappa` to the dataframe containing your simulated values of (Y, D, Z) that evaluates the function $\widehat{\lambda}(\cdot)$ and at the observed values of Z.
   d. For the subset of observations with $D = 0$, run a regression of Y on `lambda` and a constant. The intercept should be approximately equal to μ_0 and the slope approximately equal to $\delta_0 = \sigma_0\rho_0$.
   e. For the subset of observations with $D = 1$, run a regression of Y on `kappa` and a constant. The intercept should be approximately equal to μ_1 and the slope approximately equal to $\delta_1 = \sigma_1\rho_1$.