Advanced Microeconometrics - Spring 2022
Problem Set 2

Instructor: Cristián Sánchez

Due date: April 27, 2022

NOTE: You are allowed to use already-written commands that are available in statistical softwares (e.g. regress in STATA, lm in R) to perform linear estimation. However, whenever I ask you to perform nonlinear estimation, I want you to explicitly code the estimation routine, and not to use user-written commands that may be available in some softwares.

This set of questions is inspired by Rau, Sanchez and Urzua (2019). The file *data.csv* contains real data for Chilean individuals, that track their schooling and labor market decisions and outcomes from their early teens through their late twenties.

The goal is to estimate a generalized Roy model for the decision to attend a private high school in lieu of a public high school, and labor market outcomes, i.e. wages. The model includes an unobserved factor that approximates a combination of individuals' scholastic abilities.

A description of the variables in the data is as follows:

- *test_X* is students' performance on standardized test X. The unit measure is standard deviations.
- *privateHS* is a dummy indicating having attended a private high school.
- *wage* is the natural log of wage.
- *male*, *momschoolingX*, *dadschoolingX*, *broken_homeX*, *incomehhX*, *north*, *center* are demographic variables.
- *share_private* and *avg_price* are instruments for the decision of attending a private high school, and denote the local share of private high schools and the average fees local private high schools charge.

Formally, the model includes potential outcomes as follows,

$$\begin{aligned} Y_1 &= X\beta_1 + \theta\alpha_1 + U_1 \\ Y_0 &= X\beta_0 + \theta\alpha_0 + U_0, \end{aligned}$$

where Y_1 is the potential outcome (i.e. wages) in the counterfactual of attending a private high school, and Y_0 is similarly defined for the counterfactual of attending a public high school. All relevant observable demographics are included in X, while θ is the unobserved one-dimensional factor (i.e. ability) determining labor market outcomes. The unobserved factor is normally distributed with mean zero and standard deviation σ_θ. The terms U_1 and U_0 are idiosyncratic error terms, that are normally distributed with mean zero and standard deviations σ_1 and σ_0, respectively.

Individuals decide whether or not to attend a private high school based on a latent variable I:

$$I = Z\gamma + \theta\alpha_I + V,$$

where Z include observable demographics and instruments, and V is an idiosyncratic error term with zero mean and unit variance. Note that the unobservable factor is also present in this part of the model. We can thus define a binary variable D indicating treatment status,

$$D = \mathbb{1}\left[I \geq 0\right].$$

The model includes a measurement system, that helps with the identification of the distribution of θ. Specifically,

$$T_k = W\omega_k + \theta\alpha_{T_k} + \varepsilon_k \qquad \text{for each } k = 1, 2, 3, 4,$$

where T_k is test score k, W include demographics determining test scores, and ε_k is a normally distributed error term with mean zero and standard deviation σ_{ε_k}.

Finally, we assume that the error terms in the model are all independent from each other conditioning on the observables and the unobserved factor, i.e. $U_1 \perp\!\!\!\perp U_0 \perp\!\!\!\perp V \perp\!\!\!\perp \varepsilon \mid X, Z, W, \theta$, and θ is independent of all observables.

For the empirical implementation of the model, in X include *male*, *north*, and *center*. In Z include all variables in X plus *share_private* and *avg_price*. In W include *male*, *momschoolingX*, *dadschoolingX*, *broken_homeX*, *incomehhX*, *north*, and *center*. The outcome variable is *wage*. D is *privateHS*. And, the measurement system is comprised by the four test scores *test_X*.

1. Normalize $\alpha_{T_1} = 1$. Discuss and show a sketch of identification of the remaining loadings in the measurement system (i.e. α_{T_2}, α_{T_3}, α_{T_4}) and of the distribution of θ. Why do we need to assume that one of the loadings in the measurement system is equal to one?

2. Run an OLS regression of Y $(= DY_1 + (1 - D)Y_0)$ on D and X. Show your results. Are they biased? Why?

3. Using Z as instruments, estimate by 2SLS the "effect" of D on Y. Show and comment your results.

4. Now, run your OLS regression in 2. but adding T_1–T_4 as additional controls. What is the effect of D on Y?

5. Define the MTE, ATE, and TT parameters in this framework.

6. Write down the likelihood function of the model, integrating out the factor θ over its distribution.

7. Estimate the model by maximum likelihood. Show your estimated coefficients and corresponding standard errors. *HINT: I suggest you use a Gauss-Hermite quadrature for numerical integration; however, you are free to use any other numerical integration routine (e.g. Monte Carlo, Quasi-Monte Carlo).*

8. Using your estimates, simulate 10,000 observations from your model. Compare simulated and actual data averages for test scores, the decision of attending a private high school, and wages conditional on the decision of attending a private high school.

9. Using your simulated data, plot the distribution of θ.

10. Using your simulated data, plot $Y_1 - Y_0$.

11. Using your simulated data, compute and show the ATE and TT parameters.

12. Using your simulated data, plot the TT parameter as function of the unobserved ability, θ.