Applied Class #4 - Testing the LATE Assumptions

Introduction

This practical session is based on Huber & Mellace (2015). You may find it helpful to consult the paper and or my lecture notes.

Exercises

1. Write an R function that uses `rmvnorm()` from the `mvtnorm` package to simulate `n` iid draws from the model given below, with arguments `n`, `alpha` and `beta`. Your function should return a data frame 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.