ECON207 Session 3: Review Exercise B

AY2024/25 Term 1

Question 1 Consider the data on the variables X, Y and Z in `multireg_eg.csv`. The variables X and Y are continuous variables, whereas Z takes values 1, 2, 3, 4, and 5. The data are displayed in Session 3 Slide 60. The data can be considered to be a random sample from a population where

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon\,,\; E(\epsilon \mid X, Z) = 0\,,\; \beta_1 > 0, \beta_2 > 0, Cov(X, Z) < 0\,.$$

Suppose your interest is in measuring the effect of X on Y, controlling for Z. As illustrated in Session 3 Slide 60, and in line with our assumptions about the population, for any fixed Z the effect of X on Y is positive. However, because Z is positively correlated with Y and negatively correlated with X, a simple linear regression of Y on X produces a negative coefficient estimate on X.

We argued in class that we cannot hold Z fixed, and argued that the solution was to use multiple linear regression, regressing Y on X and Z together. For estimating the effect of X on Y, this has the effect of stripping out the effect of Z on both Y and X, thereby "controlling" for Z.

(a) One can argue that because Z is integer valued, we *are* in fact able to "fix" Z at some value. Suppose we regress Y on X (using simple linear regression) using only the data for $Z = 3$ and disregard the rest. Explain why this gives you an unbiased estimate of the effect of X on Y, controlling for Z.

(b) An obvious disadvantage of the proposal in (a) is that we are "wasting" a lot of data. What is the consequence of this?

(c) *Bright idea!* Suppose we regress Y on X for $Z = 1$ to get $\hat{\beta}_{1,Z=1}$, then repeat this process for each of the other values of Z. Then estimate the effect of X on Y controlling for Z using

$$\tilde{\beta}_1 = \frac{1}{5}\sum_{j=1}^{5} \hat{\beta}_{1,Z=j}\,.$$

Show that $\tilde{\beta}_1$ is a linear estimator, and explain why this estimator is unbiased.

(d) Explain why the estimator obtained by multiple linear regression of Y on X and Z should still be preferred to the estimator proposed in (c).

(e) Suppose you are now interested in measuring the effect of Z on Y, controlling for X. Are you able to use a similar method to the one proposed in (c) to achieve this?