ECO 5720
Problem Set 1

1. Please provide an example of a causal question that you would like to answer using a simple regression model such as

$$y = \beta_0 + \beta_1 x + u.$$

(i) What do x and y denote? What unobserved factors are in u?

Answer:

Example 1:

Using individual-level (i.e., child-level) data, I want to estimate the effect of exposure to pollution on health. x denotes air quality around the child's residence; y is a measure of respiratory illness. u includes other determinants of y such as family income, family education, availability of doctors, and other characteristics.

Example 2:

Using country-level data, I want to analyze the effect of tariffs on foreign direct investment (FDI). x denotes a measure of tariff rates in a country; y is the amount of foreign investment inflows for that nation. u represents other factors affecting FDI, e.g., a country's non-tariff barriers, share of skilled labor, labor policies, and other factors.

(ii) What does the assumption of zero correlation between x and u imply in your context? Is this likely to be satisfied?

Answer:

Example 1:

The assumption of zero correlation implies that exposure to pollution is uncorrelated with factors such as family income and education. This is not likely to be satisfied since the exposure could be correlated with these characteristics. For example, education and income may be correlated with choice of location and thereby exposure to pollution.

Example 2:

Here, zero correlation implies that factors such as a country's non-tariff barriers and labor policies are uncorrelated with tariff rates. However, it may well be the case that tariff policies are related to labor regulations and other non-tariff policies.

(iii) Based on your answer to part (ii), is x exogenous or endogenous?

Answer: x is thus endogenous.

(iv) What does the property of unbiasedness of the OLS estimator imply? In your context, is the OLS estimator of β_1 unbiased?

Answer: Unbiasedness implies the average or expected value of $\hat{\beta}_1$ will be equal to the true value of β_1. This is not likely in our context since x and u are correlated.

(v) What does the assumption of homoskedasticity imply in your context? Is this likely to be satisfied?

Answer:

Example 1:

Suppose we look at locations with high, medium, and low levels of ambient pollution. Homoskedasticity implies that for children in such locations, the variance of factors such as family income, family education, and environmental regulations is constant. I don't think this is likely. For example, it could that be in locations with low levels of pollution, there is less variation in availability of doctors.

Example 2:

Suppose we look at countries with low, medium, and high tariff rates. Homoskedasticity implies that for countries with such policies, the variance of factors such as non-tariff barriers and labor policies is constant. Again, this isn't necessarily true. For example, in countries with high tariff rates, there could be less variation in non-tariff policies.

(vi) In general, if x is endogenous, does OLS still minimize the sum of squared residuals (SSR)?

Answer: Yes, by design OLS minimizes SSR and fits the best line. It doesn't mean that the estimators are unbiased (and provide the true causal effect) on average.

2. Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log (price)} = 9.40 + 0.312 \log (dist)$$

(i) Interpret the coefficient on log(*dist*). Please explain the sign and magnitude in terms of *price* and *dist*.

Answer: The interpretation relates to the elasticity of price with respect to distance. If distance from an incinerator increases by 1%, housing prices increase by 0.31%.

(ii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

Answer: Other factors include size of the house, number of bathrooms, size of the lot, age of the property, and quality of the neighborhood (including school quality). These could be correlated with distance since the city may choose to locate incinerators in areas with low quality houses.

3. The data set in CEOSAL2 contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

Note: Please see the R script from February 2.

```
####################
#Data
data(ceosal2)
```

```
summary(subset(ceosal2, ceoten==0))
stargazer(subset(ceosal2, ceoten==0), type="text")
(regceo <- lm(lsalary ~ ceoten, data=ceosal2))
stargazer(regceo, type="text")
#####################
```

(i) Find the average salary.

Answer: The average salary is $865,864.

(ii) How many CEOs are in their first year as CEO (that is, $ceoten = 0$)?

Answer: Five.

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u.$$

What is the (approximate) percentage increase in salary given one more year as a CEO?

Answer: The estimated equation is

$$\widehat{\log(salary)} = 6.505 + 0.010 ceoten.$$

	Dependent variable:
	lsalary
ceoten	0.010
	(0.006)
Constant	6.505^{***}
	(0.068)
Observations	177
R^2	0.013
Adjusted R^2	0.008
Residual Std. Error	0.604 (df = 175)
F Statistic	2.334 (df = 1; 175)
Note:	$^{*}p<0.1;\ ^{**}p<0.05;\ ^{***}p<0.01$

One more year as CEO is predicted to increase salary by 100(0.010)% or 1%.