Public Policy 240B
Spring 2022
Problem Set #2

Question 1:
A researcher estimates a bivariate regression of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, but confides to a colleague that she believes $\text{cov}(x_i, \varepsilon_i) \neq 0$ and therefore, $\hat{\beta}_1$ is a biased estimate. The colleague then asks whether one can test whether $\text{cov}(x_i, \varepsilon_i) \neq 0$. The colleague suggests that the researcher construct $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ then run a regression of $\hat{\varepsilon}_i$ on x_i, that is, a regression of the form $\hat{\varepsilon}_i = \gamma_0 + \gamma_1 x_i + v_i$, then test the null Ho: $\gamma_1 = 0$ to see whether ε_i and x_i are correlated. Is this a good idea or not?

HINT: The OLS estimate of $\hat{\gamma}_1$ would be $\hat{\gamma}_1 = \dfrac{\sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

Question 2: Production data for 22 firms in a certain industry produce the following, where y = output and x = labor hours input:

$\bar{y} = 20, \quad \sum_{i=1}^{22} (y_i - \bar{y})^2 = 100$

$\bar{x} = 10, \quad \sum_{i=1}^{22} (x_i - \bar{x})^2 = 60$

$\sum_{i=1}^{22} (x_i - \bar{x})(y_i - \bar{y}) = 30.$

(A) Compute the least squares estimates of β_0 and β_1 in the model

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

(B) The OLS estimates generate a value of $\hat{\sigma}_\varepsilon^2 = 3$. Calculate the standard errors for the regression coefficient estimate of $\hat{\beta}_1$ (i.e., for the estimates of the slope). Conceptually, why do the regression coefficients have a statistical distribution?

(C) Calculate the R^2 for this model (note: the standard error of the regression can be calculated from the information given in (B)).

(D) Test the hypothesis that there is no relationship between x and y.
(E) Test the hypothesis that $\beta_1 = 0.75$.

Question 3:
Puce polo shirts from Waters' End
You have been hired as a consultant to a new clothing retailer called "Waters' End" or WE for short. The fine people at WE are interested in determining how the price of their new puce polo shirts affects the number of shirts sold. They have collected data on sales and prices from their chain of clothing stores.

You decide to start with a simple linear relationship of the form

$qs_i = \beta_0 + \beta_1 ps_i + u_i$,

where qs_i is the quantity of puce polo shirts sold at store "*i*" and ps_i is the price of the puce polo shirts at store "*i*".

(A) If prices are measured in dollars per polo shirt and quantities are measured in polo shirts, describe in words the meanings of β_0 and β_1. Use one sentence for each coefficient.

(B) Let the estimated coefficients be denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$. In terms of the notation, what is the predicted mean number of shirts sold at a store that sets the price at $40?

(C) Suppose that you estimate the model using prices measured in cents per shirt rather than dollars per shirt, where one dollar equals 100 cents. What effect does this have, if any, on the coefficient estimates?

Consider two random samples: sample A consists of 100 WE stores and sample B consists of 200 WE stores.

(D) What effect, if any, does switching from sample A to sample B have on the expected values of the coefficients from the regression? Justify your answer.

(E) What effect, if any, does switching from sample A to sample B have on the estimated standard errors of the coefficients? Justify your answer. Suppose that you decide to add a variable measuring the mean household income in the census tract (a small local area) in which each store is located to the regression, thereby yielding the model:

$qs_i = \beta_0 + \beta_1 ps_i + \beta_2 hhinc_i + u_i$,

where $hhinc_i$ is the mean household income in the census tract surrounding store "*i*".

(F) Suppose that prices were initially *randomly assigned* to the stores. What is the effect, if any, of adding the mean household income variable to the regression on the coefficient on the price? In particular, indicate whether the expected value of the new coefficient is the same, larger, smaller or different in some unknown way from the old coefficient. Express your answer in terms of the notation.

(G) Suppose that higher prices were assigned to stores in tracts with higher mean

incomes and that stores in higher income areas sell more polo shirts, conditional on price. What is the effect, if any, of adding the mean household income variable to the regression on the coefficient on the price? In particular, indicate whether the expected value of the new coefficient is the same, larger, smaller or different in some unknown way from the old coefficient. Express your answer in terms of the notation.

Question #4: An Empirical Investigation of Economic Growth

In this problem, you will explore the determinants of economic growth using a cross-country data set. A table of definitions of the variables used in this data follows the questions. The data set is available on the class website (bspace).

A. Preliminary analysis

a) Provide scatter plots for a country's annual growth rate against the average number of years of schooling and assassinations
b) Based on the graphs in a), how do years of schooling and number of assassinations appear to affect economic growth?

B. Bivariate regression

a) Write the regression of growth on years of schooling in equation form.
b) What does the subscript *i* index here?
c) Estimate the regression in a). Display the results in a table, including the R-squared and the number of observations.
d) Interpret both the intercept and the slope coefficients. Provide a 95% confidence interval for the slope coefficient. Test the null hypothesis that educational attainment is not a significant predictor of economic growth.
e) Interpret the R-squared. Does the regression provide a good fit?
f) Does the association between a country's economic growth and educational attainment persist once we control for the number of political assassinations?

Documentation for Growth Data

The data set contains data on average growth rates over 1960-1995 for 65 countries, along with variables that are potentially related to growth. These data were provided by used in paper by Ross Levine, Thorsten Beck and Norman Loayza "Finance and the Sources of Growth" *Journal of Financial Economics*, 2000, Vol. 58, pp. 261-300.

Variable	**Definition**
country_name	Name of country
growth	Average annual percentage growth of real GDP from 1960 to 1995
europe	=1 if country is in Europe =0 otherwise
yearsschool	Average number of years of schooling if adult residents in that country in 1960
assasinations	Average number of political assassinations in country from 1960 to 1995 (per million population)

Question #5: The Determinants of Test Scores

In this problem, you will explore the determinants of test scores using the Project Star data set. The project that generated this data is described in detail in a future reading in the coursepack ("How does your kindergarten classroom affect your earnings? Evidence from Project STAR"). The unit of observation is the student. The data for this problem set contains observations on 5,786 kindergarteners who attended kindergarten in Tennessee in the late 1980s. The relevant variables are described below.

Free school lunch and math scores.

a) An economic downturn has led to widespread state cuts. The legislature is considering eliminating the subsidy on lunch for needy students (providing free lunches). One legislator suggests examining the possible effects of this policy by running the regression:

$$Mathscore_i = \beta_0 + \beta_1 FreeLunch_i + \varepsilon_i$$

Provide an estimate of these coefficients.

b) If we interpret these results causally, what is the effect of providing the lunch subsidy to needy students?

c) Another legislator objects to this regression and suggests that there is omitted variable bias. Provide one possible omitted variable and derive the bias associated with its exclusion in the regression you estimated in a).

d) This legislator claims that students in the inner-city are more likely to be on FreeLunch. Investigate this claim by running the regression:

$$Mathscore_i = \beta_0 + \beta_1 FreeLunch_i + \beta_2 InnerCity_i + \varepsilon_i$$

Comparing the coefficient β_1 in d) with that in a) what can you say about the correlation between FreeLunch and Innercity?

Documentation of Test Score Data.
Variable Description

Variable	Description
Experience	Years of total teacher experience
Mathscore	Total scaled math score (measured in points)
Freelunch	=1 if student is receiving a subsidized lunch, 0 else
Masters	=1 if the teacher has a master's (MA), 0 else
Innercity	=1 if the student's residence is located in the inner city