Introduction to econometrics
Economics 4811: Spring2018

Problem Set 1

Due: *Beginning of class, January 31st, 2018*

You may collaborate with other students. Please hand in your own work. At the top of your answer sheet, please identify the students with whom you have consulted to complete this problem set.

The data for this problem set are available on the class Canvas page. Click on the link to the data file and save the data on your local disk.

STATA commands that may be useful for this problem set include:

summarize provides summary statistics. The option **detail** provides order statistics (median, 25% percentile, 75% percentile etc.)
e.g. **summarize** X Y**, detail**
correlate X Y**, covar** calculates the covariance between X and Y.
twoway scatter Y X will draw a graph of Y against X.
reg computes the OLS regression estimates

generate will generate a new variable. For example, **generate** x1sq = x1 * x1
drop if <some condition> will delete observations from your data if that condition is met.
To restrict attention to a subset of the data you can use the **if** command.
For example, **reg** Y X **if** Z~=20 runs the regression with Y, the dependent variable, regressed on X, the independent variable for the observations that satisfy the condition Z is not equal to 20. This does not delete the observations from the data set.

Question 1

A growing literature in Economics and policy-makers have been interested in studying the impact of average number of cigarettes smoked during pregnancy (*cigs*) on infant birth weight in ounces (*bwght*). The dataset ps1q1.dta contains data on 1,388 births to women in the United States. This dataset allows you to explore this question in the context of the United States.

Consider the basic regression model

$$bwght_i = \alpha_0 + \alpha_1 cigs_i + u_i \qquad (1)$$

where i denotes that each observation is at the individual birth level.

(i) Explain what u_i represents and if different newborns will have different values of u_i.

(ii) Please give 3 examples, specific to this model, of things that might be in u_i.

(iii) Give an intuitive argument for why:

a. α_1 might be positive in this model?

b. α_1 might be negative in this model?

(iv) Do you think this simple regression solves the problem of omitted variable bias and reverse causality? Why or why not? (I want you to provide intuition in your answer. No math is necessary.)

(v) Estimate quation (1) using the data in ps1q1.dta.

Interpret $\hat{\alpha}_0$ and $\hat{\alpha}_1$

(vi) What is the predicted birth weight when *cigs* = 0? What about when *cigs* = 20 (one pack per day)? Comment on the difference.

(vii) To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.

(viii) A colleague tells you that she thinks this is the wrong econometric model for the relationship between *cigs* and *bwght*:

a. What strong assumption about the relationship between *cigs* and *bwght* does your model make because it is linear?
b. Your colleague tells you that actually, the effect of *cigs* on *bwght* diminishes as *cigs* increases. How would you model that using the natural log?
c. Do you have any concerns with using that model?

Question 2

Four firms want you to use your data skills to help them understand the relationship between the hourly wages and years of education of their workers. For each firm (i=1–4) you have the years of education, "yeduc_*i*", and the hourly wages, "hrwage_*i*", for 11 of their workers. This data is available in **ps1q2.dta**. <u>Round all your answers to 2 decimal places.</u>

(i) Using these data for each firm calculate:

a. the mean and standard deviation of each variable
b. the covariance and correlation between years of education and hourly wages for each firm.

(ii) For each firm, using OLS regress hourly wages on years of education:

For example for firm 1:

hrwage_1 = $\alpha_0 + \alpha_1$ yeduc_1 + ε_1

What are the estimates of the α_1 coefficient for each firm?
(Round your answer to 2 decimal places)

(iii) For each firm, what would you tell them is your prediction from OLS of the added value of one year of education to their hourly wage?

(iv) Graph the relationship between hourly wages and years of education for each firm.

(v) Looking at the graphs would you conclude that your prediction is equally good for each firm? Why or why not?

(vi) Looking at the graphs would you conclude that the relationship between hourly wages and years of education is the same for each firm? Why or why not?

(vii) What does this suggest about the choice of model in OLS estimation?

Question 3

(Problem 2,3 of Introductory Econometrics by J. Wooldridge)

For each of the following questions show your work. You cannot use STATA to solve this problem.

The following table contains the *ACT* scores and the *GPA* (grade point average) for eight college students. Grade point average is based on a four-pint scale and has been rounded to one digit after the decimal.

Student	GPA	ACT
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(i) Estimate the relationship between GPA and ACT using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \widehat{\beta_0} + \widehat{\beta_1} ACT$$

(ii) Does the intercept have a useful interpretation here? Explain.

(iii) How much higher is the GPA predicted to be if the ACT score increases by five points?

(iv) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

(v) How much of the variation in GPA for these eight students is explained by ACT? Explain.

Question 4

The data set **ps1q4.dta** contains data on course evaluations and professor “beauty” from a sample of courses at University of Texas. A professor's beauty is determined by the average ranking of 6 students.[1]

Using Stata:

(i) Find $\overline{beauty}$

(ii) Calculate $(beauty_i - \overline{beauty})$ for each observation, and $\sum_{i=1}^{n}(beauty_i - \overline{beauty})$.

(iii) Calculate the covariance between course evaluations and beauty. What are the units of this measure? Do those units have a real world interpretation?

(iv) What is the correlation between course evaluations and beauty? Show using both the correlation function in Stata and the following definition of correlation:

$$\rho_{XY} = \frac{cov(X,Y)}{sd(X)sd(Y)}$$

(v) Plot the data with beauty on the x-axis.

(vi) Using the definition of $\widehat{\beta}_1 = \frac{cov(X,Y)}{var(X)}$, calculate the regression slope coefficient you would get from a regression of course evaluations on beauty.

(vii) Calculate the estimated intercept. How does it relate to the mean of $course\_eval$? Why?

(viii) By estimating the regression:

$$course\_eval_i = \beta_0 + \beta_1 * beauty_i + u_i$$

calculate the ordinary least squares (OLS) estimates. Compare your estimate of β_1 with the results you found in (vi).

(ix) Give an intuitive explanation of what R^2 measures. What does the standard error (RMSE) of a regression measure? Do you prefer it to R^2? Why or why not?

(x) What is the R^2 from the previous regression. Interpret this number.

[1] This is a subset of data used in Hammermesh and Parker, ``Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review*, 2005.