Introduction to econometrics
Economics 4811: Spring2018

Problem Set 1
Solutions

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.

u_i represents all factors that affect a newborn's birthweight, *bwght_i*, other than the number of cigarettes the mother smoked during pregnancy, *cigs_i*. An example would be mother's health status, which varies across newborns, and therefore we expect *u_i* to vary for between newborns. In addition, chance/luck is always in the error term as well as any measurement error.

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

1) **Parents' characteristics**: having richer, healthier, better educated parents will be positively correlated with a newborn's birth weight.
2) **Environmental characteristics**: being closer to prenatal care, living in a wealthier society with better sanitation and nutrition will be positively correlated with a newborn's birth weight.
3) **Random chance**: different newborns' birth weights can differ from just good or bad luck.

(iii) Give an intuitive argument for why:

a. α_1 might be positive in this model?

A positive α_1 means that smoking *increases* newborns' birth weight. Many arguments are possible to explain this positive effect. These are a few examples:

1. One possibility is that smoking *increases* newborns' birth weight for some unknown medical reason.

2. A positive α_1 could be explained by omitted variable bias if there is some positive correlation between the number of cigarettes a mother smokes and other unobserved characteristics also positively correlated with birth weight.

3. A positive α_1 could be explained by sample selection. Although a mother's smoking reduces birth weight, smoking causes more miscarriages for those who would have otherwise had low birth weight newborns. Smoking mothers who deliver a baby might be the healthiest mothers and this is why we see the positive effect.

b. α_1 might be negative in this model?

A negative α_1 means that smoking *decreases* newborns' birth weight. The medical literature has found that the nicotine from a cigarette gets into the placenta which affects the oxygen and nutrient that arrives to the baby. In general, any argument based on medical reasons is acceptable.

(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.)

It does not solve the problem of omitted variable bias, because there are potentially other variables related both with birthweight and smoking while pregnant. For example, smoking while pregnant could be related with a lower health status of the mother, which could affect the newborn's weight. Without adding those additional controls we cannot distinguish the causal effect of smoking while pregnant from the causal effect of the health of the mother.

Reverse causality does not seem to be a problem, since the newborn's birthweight is determined after the decision of smoking during pregnancy.

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

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

$$\widehat{bwght}_i = 119.77 - \underset{(0.09)}{0.51} cigs_i$$

$\hat{\alpha}_0 = 119.77$: The birth weight when the mother does not smoke is predicted to be 119.77 oz.
$\hat{\alpha}_1 = -0.51$: One additional cigarette per day smoked is associated with a 0.51 ounce decrease in birth weight.

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

$119.77 - .51(0) = 119.77\text{oz}$

$119.77 - .51(20) = 109.57\text{oz}$

We predict a birth weight of a little more than 10 oz less when a mother smokes a pack a day than a mother who does not smoke.

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

$119.77 - 0.51cigs = 125 \Longrightarrow cigs \approx -10$

A woman would have to smoke *negative* 10 cigarettes each day for us to predict a birth weight of 125 oz. It is clearly impossible to smoke a negative number of cigarettes.

(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?

Every additional cigarette smoked has the same correlation with birth weight, regardless of the number smoked.

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?

In this case, both $bwgt_i = \beta_0 + \beta_1 \ln cigs_i + w_i$ and $\ln(bwgt_i) = \beta_0 + \beta_1 cigs_i + w_i$ would give you a decreasing effect of an additional cigarette on birthweight.

c. Do you have any concerns with using that model?
We should prefer $\ln(bwgt_i) = \beta_0 + \beta_1 cigs_i + w_i$, given that $bwgt_i = \beta_0 + \beta_1 \ln cigs_i + w_i$ model would excludes non-smokers, because ln 0 is undefined.

The limitation of $\ln(bwgt_i) = \beta_0 + \beta_1 cigs_i + w_i$ is that it assumes that each additional cigarette leads to a constant *percentage* change in birth weight.

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**. Round all your answers to 2 decimal places.

(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.

```
summarize y* h*

    Variable |  Mean   Std. Dev.             Variable |  Mean   Std. Dev.
-------------+-----------------          -------------+-----------------
     yeduc_1 |  9.00    3.32                 hrwage_1 |  7.50    2.03
     yeduc_2 |  9.00    3.32                 hrwage_2 |  7.50    2.03
     yeduc_3 |  9.00    3.32                 hrwage_3 |  7.50    2.03
     yeduc_4 |  9.00    3.32                 hrwage_4 |  7.50    2.03

For each firm:
corr hrwage_1 yeduc_1, cov
corr hrwage_1 yeduc_1

firm | cov(y,hrw)  corr(y,hrw)
-----+-------------------------
   1 |   5.50         0.82
   2 |   5.50         0.82
   3 |   5.50         0.82
   4 |   5.50         0.82
```

(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

```
For each firm:
reg hrwage_1 yeduc_1
```

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

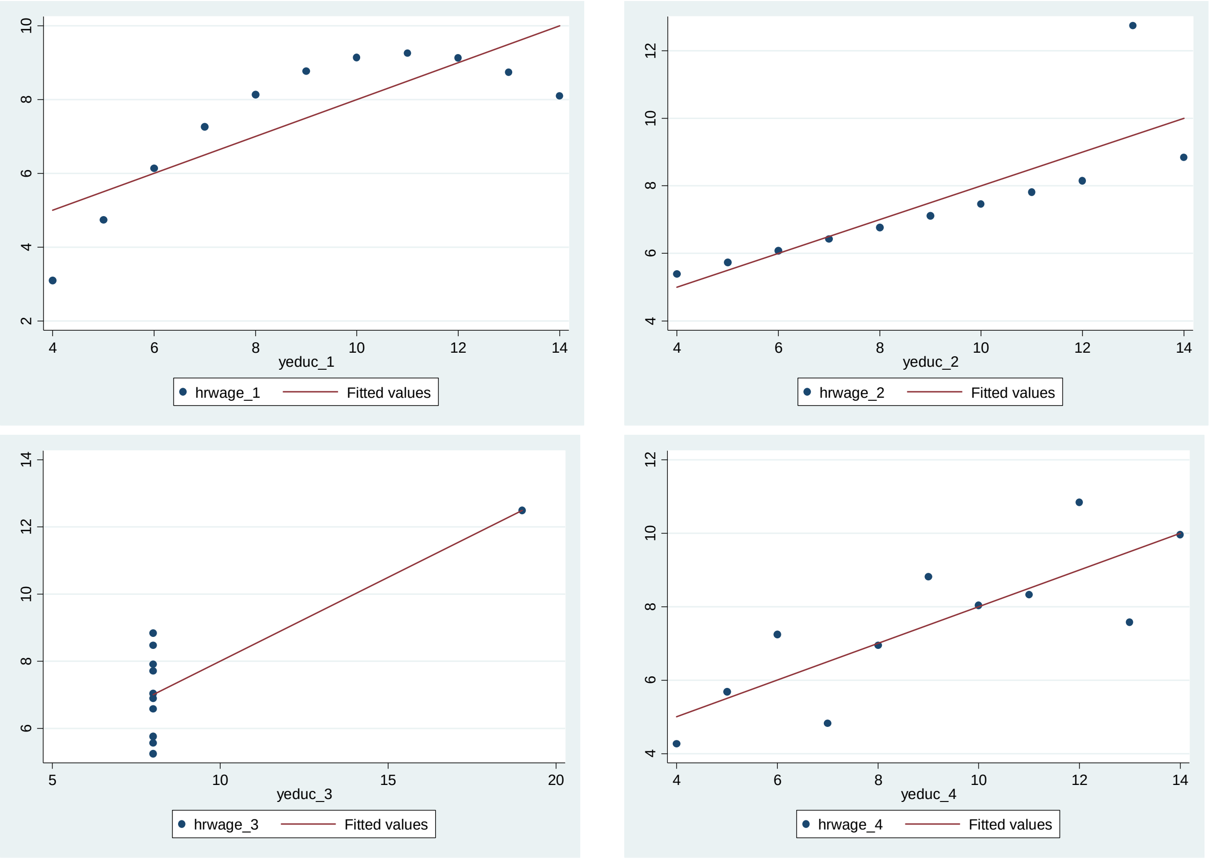
```
firm |    $\hat{\alpha}_1$
-----+-------
   1 |   0.50
   2 |   0.50
   3 |   0.50
   4 |   0.50
```

(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?

For each firm, an additional year of education is predicted to increase the hourly wage by $0.50.

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

```
For each firm:
twoway (scatter hrwage_1 yeduc_1) (lfit hrwage_1 yeduc_1)
```



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

No, my predictions are not equally good. Although from (ii) I am predicting the same effect for all the firms, it looks like for Firm 1, education has diminishing returns, therefore a linear model is not the best. For Firm 2, except for one outlier, the predicted slope should be smaller than 0.5; and for firm 3, it looks like there is no relationship between wages and education, except for one outlier. Only for Firm 4, the prediction estimated in (ii) appears reasonable.

(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?

No, the relationship is not the same for each firm. For firm 1, there appears to be a quadratic relationship, with a decreasing marginal association with education, and a *negative* marginal association after about 11 years of education. For firm 2, the marginal association appears to be less than \$0.50/year, with one outlier affecting the estimate. For firm 3, except for one outlier, there appears to be no marginal association. Only for firm 4 does the relationship look like a linear relationship with errors.

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

OLS presumes a linear model. If the actual relationship is not linear, OLS does not yield good predictions.

Recall that:

$$\hat{\beta}_1 = \frac{\text{cov}\, X, Y}{\text{var}\, X}$$

Since the covariances and the standard deviations are almost identical in all pairs, the OLS estimates will also be very similar. However, if we look at the scatter plot of each pair we note that the relationship between X and Y is in fact very different in each case.

This finding clearly illustrates how basic summary statistics are usually not enough to capture all the richness of the data, and thus, any OLS estimate (based on these statistics) will not necessarily give us an accurate idea of the relationship between two variables. This highlights the relevance of being cautious when specifying our econometric model.

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-point 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

$$\widehat{GPA} = \widehat{\beta}_0 + \widehat{\beta}_1 ACT$$

$$\overline{GPA} = \frac{1}{8}(2.8 + 3.4 + 3.0 + 3.5 + 3.6 + 3.0 + 2.7 + 3.7) = 3.2125$$

$$\overline{ACT} = \frac{1}{8}(21 + 24 + 26 + 27 + 29 + 25 + 25 + 30) = 25.875$$

$$\text{var } ACT = \frac{1}{8-1}[(21 - 25.875)^2 + \cdots + (30 - 25.875)^2] = 8.125$$

$$\text{cov } G, A = \frac{1}{8-1}[(2.8 - 3.2125)(21 - 25.85) + \cdots] = 0.8304$$

$$\hat{\beta}_1 = \frac{\text{cov } G, A}{\text{var } ACT} = \frac{0.8304}{8.125} \approx 0.10$$

$$\hat{\beta}_0 = \overline{GPA} - \hat{\beta}_1\overline{ACT} = 0.57$$

(ii) Does the intercept have a useful interpretation here? Explain.
The intercept does not have a useful interpretation. It predicts the GPA of someone who scores zero on the ACT, which is not a possible score.

(iii) How much higher is the GPA predicted to be if the ACT score increases by five points?
$\Delta\widehat{GPA} = \hat{\beta}_1(5) = 0.5$, so a 5 point ACT score increase is predicted to increase GPA by 0.5 points.

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

GPA_i	*ACT_i*	$\widehat{GPA_i}$	$GPA_i - \widehat{GPA_i}$
2.8	21	2.7143	0.0857
3.4	24	3.0209	0.3791
3	26	3.2253	-0.2253
3.5	27	3.3275	0.1725
3.6	29	3.5319	0.0681
3	25	3.1231	-0.1231
2.7	25	3.1231	-0.4231
3.7	30	3.6341	0.0659
		SUM	0.0000

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

$$R^2 = \sum \frac{\left(\widehat{GPA_i} - \widehat{\overline{GPA}}\right)^2}{(GPA_i - \overline{GPA})^2} = 0.5774$$

The variation in *ACT* explains 57.74 percent of the variation in *GPA*. Knowing a student's ACT score gives us 58 percent of the information we need to predict their GPA.

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 $\overline{beauty}$

sum beauty

$\overline{beauty} \approx 0$

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

$beauty_i - \overline{beauty} = beauty_i$ for all observations.

$$\sum_{i=1}^{n}(beauty_i - \overline{beauty}) = 0$$

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

corr beauty course_eval, cov

$\mathrm{cov}\ beauty, course_{eval} = 0.0973$

The units of measure are *beauty* units × *course_eval* units, which do not 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:

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

sum beauty course_eval

sd(*beauty*)= 0.7212

sd(*course_eval*)=0.5986

$$\rho_{beauty,eval} = \frac{0.0973}{(0.7212)(0.5986)} \approx 0.2254$$

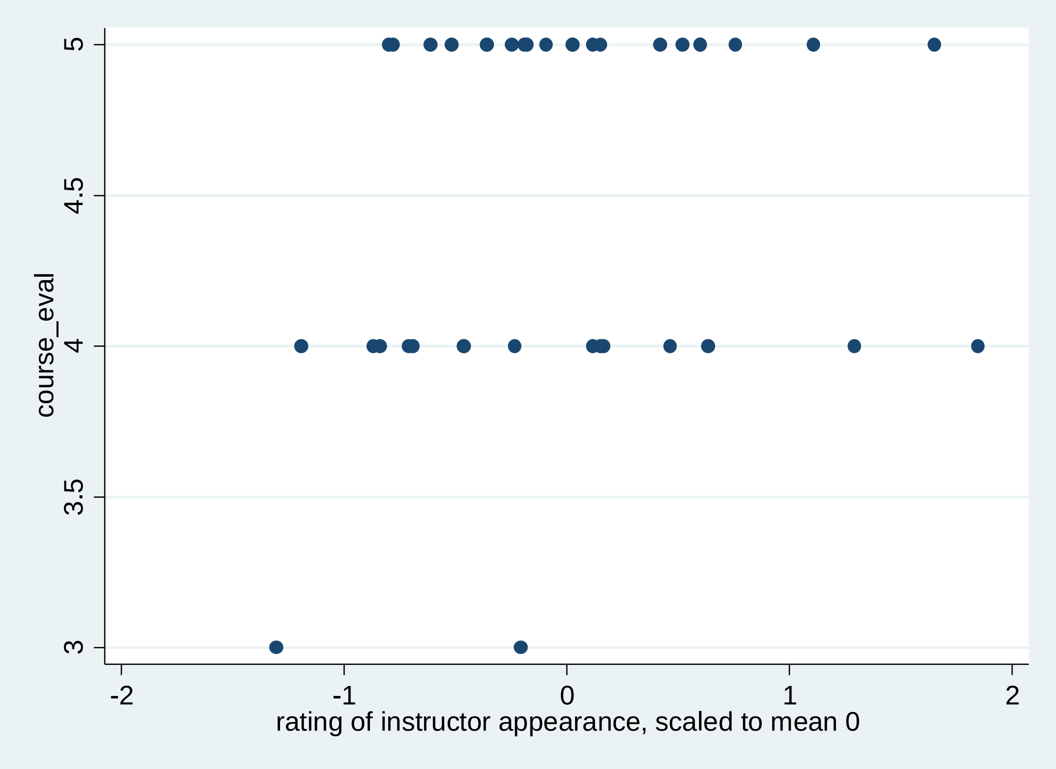
corr beauty course_eval

$\rho_{beauty,eval} = 0.2254$

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

twoway (scatter course_eval beauty)

[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.



(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.

$$\hat{\beta}_1 = \frac{0.0973}{0.7212^2} \approx 0.19$$

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

$\hat{\beta}_0 = \overline{eval} - \hat{\beta}_1\overline{beauty} = 4.525 + 0.19(0) = 4.525$

The regression line always passes through the mean of the *X*s and *Y*s. Since the mean of the *X*s, *beauty*, is 0, the y-intercept is $\overline{eval}$.

(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).

$\widehat{eval}_i = \underset{(0.13)}{4.525} + \underset{(0.09)}{0.19} beauty_i$

They are identical except for rounding errors.

(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?

R^2 measures the *percentage* of variation in *Y* that is explained or predicted by variation in *X*. The Root Mean Squared Error, i.e. $\sqrt{\Sigma(Y_i - \hat{Y}_i)^2}$, measures the *amount* of variation in *Y* (in units of *Y*) not predicted by variation in X. R^2 is preferable because it is scaled by the total amount of variation in Y, and is easier to interpret.

(x) What is the R^2 from the previous regression. Interpret this number.
$R^2 = 0.0508$: About 5 percent of the variation in course evaluations is predicted by variation in beauty ranking.