Ecn 102 - Analysis of Economic Data
University of California - Davis
John Parman

March 5, 2011

Problem Set 6 - Solutions

This problem set will not be collected. It is for extra practice to help you prepare for the final. Solutions to the problem set are available on Smartsite.

1. Suppose that the size of the squirrel population in Davis depends on the amount of rain Davis receives and the amount of nuts the trees produce. Squirrels dislike rain, so more rain means fewer squirrels. Squirrels like nuts, so more nuts means more squirrels. The true relationship between the squirrel population and rainfall and amount of nuts is given by:

$$P = 1000 - 20R + 4N + \varepsilon \tag{1}$$

where P is the number of squirrels, R is the average monthly rainfall in inches, N is the average number of nuts produced by a tree in Davis, and ε is an error term that satisfies all of our assumptions. The number of nuts each tree produces depends on how much rain Davis gets. The true relationship between rain and nuts is given by:

$$N = 100 + 2R + \nu \tag{2}$$

where ν satisfies all or our assumptions.

(a) Suppose that we run a regression with squirrel population as the dependent variable and rainfall as the independent variable. Will the errors be correlated with the regressor? If so, will they be positively or negatively correlated?

If we regress P on R but omit N, the $4N$ term will go into the error term. Since N is positively correlated with R, this means that the error term will now be positively correlated with regressor R. This will lead to an omitted variable bias.

(b) What is the expected value of the estimated slope coefficient? What would the expected value of the slope coefficient be if the correlation between N and R was zero?

The expected value of the slope coefficient will be equal to the true value of the slope coefficient in the population model plus a term capturing the omitted variable bias. This bias term will be equal to the coefficient relating N to P multiplied by the coefficient relating R to N:

$$E(\tilde{b_R}) = \beta_R + \beta_N \cdot \gamma_R$$

In the above equation, β_R is the coefficient for rainfall in equation (1), β_N is the coefficient for nuts in equation (1) and γ_R is the coefficient for rain in equation (2). Plugging in the appropriate values gives us:

$$E(\tilde{b_R}) = -20 + 4 \cdot 2$$

$$E(\tilde{b_R}) = -12$$

If the rainfall and nuts were uncorrelated, then γ_R would be zero and we would not have an omitted variable bias, so the expected value of $\tilde{b_R}$ would be equal to 20.

(c) One way to get a precise value for N is to count the number of nuts produced by every tree in Davis and then divide by the number of trees. Suppose that to save time we decided to randomly sample just a few trees (say 10 trees), count the number of total nuts produced and divide by the number of trees in the sample to get $\tilde{N}$. How does the expected value of $\tilde{N}$ compare to the true value of N? If we used $\tilde{N}$ instead of N to estimate equation (1), would the expected value of the slope coefficient on $\tilde{N}$ be greater than, less than or equal to 4?

> If we truly take a random sample of trees, then the expected value of $\tilde{N}$ will be equal to N (this is just saying that on average, the sample mean will be equal to the population mean). However, sometimes $\tilde{N}$ will be a bit bigger than N and sometimes it will be a bit smaller depending on which trees we happen to sample. This makes $\tilde{N}$ a noisy measure of N and will lead to a measurement error problem, biasing the coefficient on the nuts term toward zero.

(d) Suppose we were only interested in the relationship between nuts and rainfall. If we used $\tilde{N}$ instead of N to estimate equation (2), how would the estimated slope coefficient be affected?

> The problem is still about measurement error but now it is measurement error in the dependent variable, not the independent variable. This will make are estimates less precise, increasing the standard error of the slope coefficient on rainfall but it will not bias the slope coefficient, so the expected value of the estimated slope coefficient on rainfall will still be 2.

2. For each scenario below, determine whether the estimated coefficient will be biased. If there will be a bias, determine the sign of the bias. In every scenario, the researcher is running a regression with SAT score as the dependent variable and a dummy variable for whether a person takes an SAT prep course as the independent variable to test whether taking a prep course is associated with a higher SAT score. The dummy variable is equal to one if a person takes a prep course and zero otherwise. Treat each scenario separately (in other words, when answering one part, ignore any of the omitted

variables mentioned in the other parts). There may be multiple correct answers for a part, the key is to be able to explain the economic intuition behind your chosen answer.

(a) Individuals who take prep courses are more likely to have parents that will buy them other test prep materials which help improve scores.

> We are mistakenly omitting a variable capturing the use of test prep materials. This variable would be postively correlated with the prep course dummy variable and positively correlated with SAT score leading to a positive bias for the prep course coefficient. Our regression will overstate the change in SAT score associated with taking a prep course.

(b) Individuals who take prep courses do so because they tend to do poorly on standardized tests and need extra help.

> Our error term includes test taking ability which is negatively correlated with the prep course dummy variable (lower ability will be associated with higher likelihood of taking a course) but postively correlated with SAT score. This will produce a negative bias for the prep course coefficient leading us to underestimate the change in SAT score associated with taking a prep course.

(c) All students sign up for prep courses but due to space limitations, only a random subset of students are admitted into the prep course.

> If students are chosen at random, then there shouldn't be any unobserved characteristics that are correlated with the course prep dummy variable. Our error term will be uncorrelated with the dummy variable and we should be able to get an unbiased estimate of the slope coefficient.

(d) All students sign up for prep courses but due to space limitations, only those students willing to pay the most are admitted into the prep course.

> Now the students are no longer being chosen randomly. Instead, the value of the dummy variable will be positively correlated with student's willingness to pay. This can create a bias, the direction of which depends on how willingness to pay is correlated with SAT score. If willingness to pay is negatively correlated with SAT score (say because students that do badly on tests are desperate to increase their scores), then there will be a negative bias on the dummy variable coefficient. If willingness to pay is positively correlated with SAT score (say because wealthy parents tend to invest a lot in their children which leads to good academic outcomes) then there will be a positive bias on the dummy variable.

3. Suppose that the number of individuals who smoke (N) is related to the the number of ads for cigarettes (A), the number of studies about the link between smoking and cancer (S) and a random error term (ε). When the number of studies is held constant

at 100, one extra ad is associated with a 5 percent increase in the number of smokers. When the number of studies is held constant at 200, one extra ad is associated with a 1 percent increase in the number of smokers. When there are 100 ads for cigarettes, an additional study is associated with a 5 percent decrease in the number of smokers. If there are no ads or studies, the expected number of smokers is 10,000. Write down an equation that captures the relationship between N, A and S. N should be your dependent variable.

The basic model we should be thinking about will have relate the natural log of the number of smokers (because everything was phrased in terms of percent changes in the numer of smokers) to A, S and and interaction term $A \cdot S$ because the marginal effect of one of the independent variables depends on the value of the other independent variable. So our basic model should be:

$$ln(N) = \beta_1 + \beta_2 A + \beta_3 S + \beta_4 A \cdot S + \varepsilon$$

Now we need to figure out values for the coefficients. Let's start with what we are told about the impact of one extra ad on the number of smokers. Based on the equation above, the change in the log of the number smokers with an additional ad is:

$$\frac{dln(N)}{dA} = \beta_2 + \beta_4 S$$

We are given two pieces of information about this equation relating to two different values for S:

$$.05 = \beta_2 + \beta_4 100$$

$$.01 = \beta_2 + \beta_4 200$$

Subtracting the second equation from the first gives us:

$$.04 = -100\beta_4$$

$$-.0004 = \beta_4$$

To get a value for β_2, we can plug this value for β_4 back into either one of the equations:

$$.05 = \beta_2 - .0004 \cdot 100$$

$$.05 = \beta_2 - .04$$

$$.09 = \beta_2$$

Now let's move on to using our information about impact of one additional study:

$$\frac{dln(N)}{dS} = \beta_3 + \beta_4 A$$

Plugging in the information from the problem and the value we calculated for β_4 gives us:

$$-.05 = \beta_3 - .0004 \cdot 100$$

$$-.01 = \beta_3$$

Finally, we can get the value of β_1 using the information about the number of smokers when A and S are both zero:

$$ln(N) = \beta_1 + \beta_2 A + \beta_3 S + \beta_4 A \cdot S + \varepsilon$$

$$ln(10000) = \beta_1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0$$

$$9.2 = \beta_1$$

So our final model is:

$$ln(N) = 9.2 + .09A - .01S - .0004A \cdot S + \varepsilon$$