Exercise session 9

a) Multiple choice questions

1. The term 'u' in an econometric model is usually referred to as the _____.

a. error term

b. parameter

c. hypothesis

d. dependent variable

2. Which of the following is a nonlinear regression model?

a. $y = \beta_0 + \beta_1 * x^{1/2} + u$

b. $\log y = \beta_0 + \beta_1 * \log x + u$

c. $y = 1 / (\beta_0 + \beta_1 * x) + u$

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

3. If a change in variable x causes a change in variable y, variable x is called the _____.

a. dependent variable

b. explained variable

c. explanatory variable

d. response variable

4. Analyzing the behavior of unemployment rates across U.S. states in March of 2006 is an example of using

a. time series data.

b. panel data.

c. cross-sectional data.

d. experimental data.

5. A data set that consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time, is called a(n) _____.

a. cross-sectional data set

b. longitudinal data set

c. time series data set

d. experimental data set

6. Econometrics can be defined as follows with the exception of
a. the science of testing economic theory.
b. fitting mathematical economic models to real-world data.
c. a set of tools used for forecasting future values of economic variables.
d. measuring the height of economists.

7. The cumulative probability distribution shows the probability
a. that a random variable is less than or equal to a particular value.
b. of two or more events occurring at once.
c. of all possible events occurring.
d. that a random variable takes on a particular value given that another event has happened.

8. An estimator is
a. an estimate.
b. a formula that gives an efficient guess of the true population value.
c. a random variable.
d. a nonrandom number.

9. The correlation between X and Y
a. cannot be negative since variances are always positive.
b. is the covariance squared.
c. can be calculated by dividing the covariance between X and Y by the product of the two standard deviations.
d. is given by $corr(X, Y) = \frac{\text{cov}(X,Y)}{\text{var}(X)\text{var}(Y)}$.

10. A type II error
a. is the error you make when not rejecting the null hypothesis when it is false.
b. is the error you make when choosing type II or type I.
c. is typically smaller than the type I error.
d. cannot be calculated when the alternative hypothesis contains an "=".

Problem 1. In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not.

(i) Suppose you can collect random samples of the driving-age population in both states, for 1985 and 1990. Let arrest be a binary variable equal to unity if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?

$$\boldsymbol{arrest = \alpha_0 + \alpha_1 FL + \alpha_2 Y90 + \alpha_3 FL * Y90 + \varepsilon}$$

where *FL* is a binary variable equal to one if a person lives in Florida, and zero otherwise, Y*90* is a year dummy variable for 1990. The effect of the law is measured by α_3, which is the change in the probability of drunk driving arrest due to the new law in Florida. Including Y*90* allows for aggregate trends in drunk driving arrests that would affect both states; including *FL* allows for systematic differences between Florida and Georgia in either drunk driving behavior or law enforcement.

(ii) Why might you want to control for other factors in the model? What might some of these factors be?

It could be that the populations of drivers in the two states change in different ways over time. For example, age, race, or gender distributions may have changed. The levels of education across the two states may have changed. As these factors might affect whether someone is arrested for drunk driving, it could be important to control for them. At a minimum, there is the possibility of obtaining a more precise estimator of α_3 by reducing the error variance. Essentially, any explanatory variable that affects *arrest* can be used for this purpose.

Problem 2. Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u$$

where PC is a binary variable indicating PC ownership.

(i) Why might PC ownership be correlated with u?

It has been fairly well established that socioeconomic status affects student performance. The error term *u* contains, among other things, family income, which has a positive effect on *GPA* and is also very likely to be correlated with PC ownership

(ii) Explain why PC is likely to be related to parents' annual income. Does this mean parental income is a good IV for PC? Why or why not?

Families with higher incomes can afford to buy computers for their children. Therefore, family income certainly satisfies the second requirement for an instrumental variable: it is correlated with the endogenous explanatory variable. But as we suggested in part (i), *faminc* has a positive affect on *GPA*, so the first requirement for a good IV fails for *faminc*. If we had *faminc* we would include it as an explanatory variable in the equation; if it is the only important omitted variable correlated with *PC*, we could then estimate the expanded equation by OLS.

(iii) Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC.

This is a natural experiment that affects whether or not some students own computers. Some students who buy computers when given the grant would not have without the grant. (Students who did not receive the grants might still own computers.) Define a dummy variable, *grant*, equal to one if the student received a grant, and zero otherwise. Then, if *grant* was randomly assigned, it is uncorrelated with u. In particular, it is uncorrelated with family income and other socioeconomic factors in u. Further, *grant* should be correlated with *PC*: the probability of owning a PC should be significantly higher for student receiving grants. Incidentally, if the university gave grant priority to low-income students, *grant* would be negatively correlated with u, and IV would be inconsistent.

Problem 3. Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coed schools. You have a random sample of senior high school girls from a state in the United States, and score is the score on a standardized math test. Let girlhs be a dummy variable indicating whether a student attends a girls' high school.

(i) What other factors would you control for in the equation? (You should be able to reasonably collect data on these factors.)

Family income and background variables, such as parents' education.

(ii) Write an equation relating score to girlhs and the other factors you listed in part (i).

$$MathSC = \gamma_0 + \gamma_1 girlhs + \gamma_2 faminc + \gamma_3 meduc + \gamma_4 feduc + \varepsilon$$

(iii) Suppose that parental support and motivation are unmeasured factors in the error term in part (ii). Are these likely to be correlated with girlhs? Explain.

Parents who are supportive and motivated to have their daughters do well in school may also be more likely to enroll their daughters in a girls' high school. It seems likely that *girlhs* and ε are correlated.

(iv) Discuss the assumptions needed for the number of girls' high schools within a 20-mile radius of a girl's home to be a valid IV for girlhs.

Let *numghs* be the number of girls' high schools within a 20-mile radius of a girl's home. To be a valid IV for *girlhs*, *numghs* must satisfy two requirements: it must be uncorrelated with ε and it must be partially correlated with *girlhs*. The second requirement probably holds, and can be easily tested. The first requirement is more problematic. Girls' high schools tend to locate in areas where there is a demand, and this demand can reflect the seriousness with which people in the community view education. Some areas of a state have better students on average for reasons unrelated to family income and parents' education, and these reasons might be correlated with *numghs*. One possibility is to include community-level variables that can control for differences across communities.

Problem 4. Let grad be a dummy variable for whether a student-athlete at a large university graduates in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following logit model is obtained:

$$\hat{P}(grad = 1|hsGPA, SAT, study) = \Lambda(-1.77 + .24hsGPA + 0.00058SAT + 0.073study),$$

Where, $\Lambda(\mathrm{z}) = \frac{\exp(\mathrm{z})}{1+\exp(\mathrm{z})}$ is the logit function. Holding hsGPA fixed at 3.0 and SAT fixed at 1,200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

We need to compute the estimated probability first at *hsGPA* = 3.0, *SAT* = 1,200, and *study* = 10 and subtract this from the estimated probability with *hsGPA* = 3.0, *SAT* = 1,200, and *study* = 5. To obtain the first probability, we start by computing the linear function inside $\Lambda(\cdot)$:

$-1.77 + .24*3.0 + .00058*1,200 + .073*10 = .376$. Next, we plug this into the logit function:

$\exp(.376)/[1 + \exp(.376)] \approx .593$. This is the estimated probability that a student-athlete with the given characteristics graduates in five years.

For the student-athlete who attended study hall five hours a week, we compute:

$-1.77 + .24*3.0 + .00058*1,200 + .073*5 = 0.011$. Evaluating the logit function at this value gives $\exp(.011)/[1 + \exp(.011)] \approx .503$. Therefore, the difference in estimated probabilities is $.593 - .503 = .090$, or just under 0.1. [*Note how wrong the calculation would be if we simply use the coefficient on study to conclude that the difference in probabilities is .073*(10 – 5) = .365.*]