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?

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

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?

(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?

(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.

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.)

(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.

(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.

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:

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

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.