PhD/MA Econometrics Examination
January 2012

PART A

ANSWER ANY <u>TWO</u> QUESTIONS IN THIS SECTION

NOTE:

(1) The indicator function has the properties: $I_{(a,b)}(x) = \begin{cases} 1, & x \in (a,b) \\ 0, & x \notin (a,b) \end{cases}$

(2) $\int x e^{ax}\, dx = (e^{ax} / a^2)(ax - 1)$

<u>Question 1</u>

Let $F(x) = 1 - e^{-x}$, $0 \leq x < \infty$ [defined as $F(x) = 1 - e^{-x}\ I_{[0,\infty)}(x)$ if using the indicator function]

(a) Show that $F(x)$ is a CDF for the given range of x. Provide evidence to support your conclusion.

(b) Let $f(x) = e^{-x}, 0 \leq x < \infty$. Is $f(x)$ a PDF, and if so, is it the PDF of $F(x)$?
Provide evidence to support your conclusions.

(c) Find the mean of the random variable X that has $F(x)$ as its CDF.

<u>Question 2</u>

The average price, P, and total quantity (in tens of thousands) sold, Q, of an economy brand ballpoint pen in a large southwestern retail market during the month of August is represented by the outcome of a bivariate random variable having a probability density function:
$f(p,q) = 10pe^{-pq} I_{[0.10,0.20]}(p) I_{(0,\infty)}(q)$.

(a) Find the marginal distribution of P.

(b) Find the conditional PDF of Q on P.

(c) Define the regression curve of Q on P and find the expected quantity of pens sold, given that the price is equal to $0.50.

PART A, Continued

Question 3

(a) State the Gauss Markov theorem in words and mathematically (scalar form is sufficient here).

Suppose Y_1, Y_2 is a sample of two observations that are independently distributed as normal random variables. Both the random variables have mean $E(Y_i) = \beta \quad for \quad i = 1,2$ but they have different variances. In particular: $\text{var}(Y_1) = \sigma^2$ and $\text{var}(Y_2) = a\sigma^2$, where a is a scalar.

(b) Show that the least squares estimator $b = \sum_{i=1}^{3} y_i / 2$ is an unbiased estimator of β.

(c) Find the variance of the least squares estimator b. Is the following weighted estimator $b_w = \frac{1}{2}Y_1 + \frac{1}{4}Y_2$ unbiased? Is $\text{var}(b_W) < \text{var}(b)$? Comment.

PART B
ANSWER ANY <u>TWO</u> QUESTIONS IN THIS SECTION

<u>Question 4</u>

a. Consider the simple linear regression model, $y_i = \alpha + \beta x_i + u_i$. State the objective function for ordinary least squares, using summation notation. Then, derive algebraically the first order conditions (i.e., the normal equations) for OLS estimation of β, assuming a sample size of n.

b. Using the normal equations, show algebraically that $\sum_i \hat{u}_i = 0 \quad and \quad \bar{y} = \bar{\hat{y}}$.

c. Now, let $\mathbf{y} = \mathbf{Xb} + \mathbf{u}$ (in matrix notation) be a valid model for the population. State, both in words and in standard mathematical notation, the classical linear modeling (CLM) assumptions required to prove the OLS estimator $\hat{\mathbf{b}}$ is an unbiased estimator for $\mathbf{b}$. Then, show the proof in matrix notation, clearly indicating where and why each of these assumptions is required.

d. Derive, in matrix notation, the variance of $\hat{\mathbf{b}}$ under the CLM assumptions.

<u>Question 5</u>

a. Define heteroskedasticity, both in words and using standard matrix notation. One potential cause of heteroskedasticity is provided in part (d), below. Provide two additional real-world examples of different causes of heteroskedasticity, then discuss the consequences of heteroskedasticity for OLS estimation.

b. Describe an algorithm (i.e., a formal testing procedure, such as BP, White's, or Wooldridge's alternative to White's) to detect presence of heteroskedasticity.

c. What is GLS, and how does it differ from OLS? Assuming again the linear model (in matrix notation) $\mathbf{y} = \mathbf{Xb} + \mathbf{u}$, derive the GLS estimator.

d. Suppose we want to estimate the effect of a federal grant program on students' successfully completing a college degree, using panel data from a large, representative random sample of high school graduates. Let y be a binary dependent variable which, for each individual i in n, takes on the value $y_i = 1$ if the individual completed a degree, and $y_i = 0$ if not. Let x_1 also be a dummy variable, which takes on the value $x_{1i} = 1$ if the individual received a federal grant, and $x_{1i} = 0$ otherwise.

- **i.** Considering the CLM assumptions, discuss under what circumstances (if any) OLS estimation would provide an <u>unbiased</u> estimate of the effect of the grant program.
- **ii.** As noted above in part (a), the linear probability model (LPM) proposed here could suffer from heteroskedasticity. Show and explain why this is true.
- **iii.** Describe <u>two</u> (better!) alternatives to OLS estimation in this context, to correct the issues you've noted in parts (i-ii) above. Specifically, for each alternative estimator you propose, describe the estimation procedure, and discuss what issue(s) it resolves (i.e., why is it better than OLS?).

PART B, Continued

Question 6

If one views education as an investment in human capital, where students face time costs, forgo earnings, and endure stress, then students should stay in school only if their anticipated gains are large enough to offset these costs. However, individuals' education is also presumed to provide positive externalities for society, such as better civic participation and greater economic development. Acemoglu & Angrist (2001) attempt to estimate separately the magnitude of these private and external returns, using Census wage data from 1960-1980. Their repeated cross-sectional data (note, these are <u>not</u> panel data) include the following variables for each individual observed: weekly wage for individual i in state j at time t (Y_{ijt}), state of birth, state of residence, survey year, and the individual's years of schooling (s_i). With these data, the authors construct state-level averages for years of schooling for each state and year, $\overline{S}_{jt}$, which they use to estimate the external returns to schooling.

a. Write the regression equation for a pooled cross-sectional analysis to estimate the effect of increased years of schooling on log(weekly wage). Define all variables, and be careful to use appropriate subscripts (indexes provided above). For each coefficient to be estimated, explain what the term is controlling for, and how the coefficient should be interpreted.

b. State university systems rely on tax revenues, so states may decrease investments in schooling when the economy is in recession. When the economy slows, average wages may also decline. Explain and show mathematically how (in which direction? to what extent?) omitting a measure of state-year macroeconomic conditions would affect your parameter estimate on $\overline{S}_{jt}$.

   Would state-specific fixed effects (i.e., a vector of state-level dummy variables) solve this problem? Why or why not?

c. One approach to mitigate omitted variable bias is instrumental variables (IV) estimation. State and define the two conditions required for a "good" instrument, describe the IV/2SLS approach to estimation, and discuss how and whether each of these conditions can be formally tested.

d. Even if omitted variable bias as in part (b) were not a problem, positive state-specific shocks to wages should attract more highly-educated workers to a state in future periods, thereby raising $\overline{S}_{jt}$. Discuss the potential implications (bias? variance?) of these theoretical results for empirical (OLS) estimates of the external returns to schooling using the $\overline{S}_{jt}$ variable.

PART C

ANSWER ANY <u>TWO</u> QUESTIONS IN THIS SECTION

<u>Question 7</u>

Consider the following structural model

y1(t) = a0 + a1*x1(t) + a2*y1(t-1) + u1(t)
y2(t) = b0 + b1*x1(t) + b2*y1(t) + b3*y2(t-1) + u2(t)

a. List the endogenous variables. List the predetermined variables.
b. Write the above model in a general matrix notation:

   $Y\Gamma + X\Delta + U = 0$

c. Derive the Var-Cov(U).
d. Derive the 3 SLS estimator.
e. Present the 3SLS method of estimation (iterative GLS).
f. Why is it called a FIML method?
g. Is this a recursive model? Why?

<u>Question 8</u>

Assume that the following tobit model is postulated:

y(t)* = β0 + x(t)' β + z(t)'*δ + u(t)

Where:

y(t)* = y(t) if y(t)* > 0
0 if y(t)* ≤ 0

x(t)' is the row vector of variables: income, home ownership;
z(t)' is the row vector of race variables capturing the two 1/0 categorical variables: Back and Hispanics (with the base category being Whites). The sample consists of only three race categories.

β and δ are column vectors representing slope parameters.

The credit score has numerous missing values replaced by zeros.

PART C, Continued

Question 8, Continued

a. Set up the likelihood function for this model using the normal distribution. (Show all the steps.)

b. Now assume that the data is censored as follows:

$$y(t)^* = \begin{cases} y(t) & \text{if } y(t)^* \leq 750 \\ 0 & \text{if } y(t)^* > 750 \end{cases}$$

That is, the credit score information for people with score greater than 750 has been censored. Set up the likelihood function. (Show all the steps)

c. Show in details the method of testing the hypothesis about the race variable in such disclosures method (Show null, alternate, likelihood ratio test, critical graph, etc.).

Question 9

Consider the following AR(1) model:

$$y_t = \emptyset_1 \; y_{t-1} + e_t$$

where the series is stationary. For simplicity, the constant is suppressed by assuming that the series is mean-deviated.

a. Derive the autocorreation functions for up to lag 3.

b. Assume that the value of the autocorreation parameter, $\emptyset_1$ = -.62. Plot the acf graph.

c. What is the meaning of the following terms: long-run memory, unit root. Compare and contrast the two with examples.

d. How would you test the presence of the unit root in the above mode?

e. Assuming stationarity, derive the moving average representation of the above AR(1) model. Solve: $y_t = (1 - \emptyset_1 B)^{-1} e_t$

f. What is the economic meaning of this MA representation? (assume the y_t represents the per capita GDP)