PhD/MA Econometrics Examination

August 2019

Total Time: 8 hours

MA students are required to answer from A and B.

PhD students are required to answer from A, B, and C.

The answers should be presented in terms of equations, statistical details, and with necessary proofs and statistical deduction. Verbal and brief descriptive discussions will not be sufficient.

PART A

(Answer any TWO from Part A)

1. ***Probability Theory, Distributions, and More***
    a. State Bayes' Theorem
    b. In words, describe what Bayes' Theorem means
    c. What distribution is below:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

    d. Find the first and second moments of this distribution directly from the distribution itself.
    e. Find the first and second moments using the moment generating function. Also, discuss if these moments are the same or different from the moments you found in **part d** and why or why not.
    f. This distribution has a property called "memoryless." Prove that this distribution is memoryless.
    g. Name a practical application of this distribution, i.e., where does it naturally occur?
    h. Name the third and fourth moments – *name them do not calculate them.*
    i. Define and draw the PDF and CDF of the distribution.

2. ***Estimator:*** Estimating the Variance of the LS Estimator
    a. We know $var(b) = \sigma^2(X'X)^{-1}$, but σ^2 is an unknown parameter. Therefore in order to find $var(b)$, we need to find a good estimator for σ^2. *Derive an estimator of* σ^2.
    b. Is it unbiased? Prove this is an unbiased estimator?
    c. If you take the square root of this estimator what is that called?
    d. How and why do we use the estimator you just derived?

3. Ordinary Least Squares (OLS)

- **a.** State the classical assumptions – in words and equations.
- **b.** Derive the normal equations.
- **c.** Demonstrate that the OLS estimator is BLUE.
- **d.** From $X'e = 0$, we can derive a number of properties. State these properties. Hint: there are 6 and 5 of them require that the OLS regression includes a constant.

Part B: Answer any two of the following three questions

[Short verbal descriptive answer without mathematical proofs, steps, and necessary derivation will not earn you full credit.]

4. The results reported in Table 1 were generated from a regression of a sample of workers' hourly wage rates on the following control variables: *education, years of experience, years of experience squared, a dummy variable equal to one if female (male is the base category), a dummy variable equal to one if looks are rated as above average, and a dummy variable equal to one if looks are rated as below average (looks rated as average is the base category).*

Table 1: Determinants of Hourly Earnings

	Model I (1)	**Model II (2)**
Education	0.412***	0.413***
	(0.046)	(0.046)
Experience	0.256***	0.259***
	(0.038)	(0.039)
Experience squared	-0.004***	-0.004***
	(0.001)	(0.001)
Female	-2.49***	-2.89***
	(0.255)	(0.337)
Above Average	0.075	-0.284
	(0.268)	(0.332)
Female X Above Average		1.011*
		(0.551)
Below average	-0.915**	-1.147**
	(0.370)	(0.465)
Female X Below Average		0.673
		(0.766)
Constant	-0.704	-0.611
	(0.695)	(0.698)

Standard errors in parentheses

Please answer the following questions regarding the results reported in Table 1.

a. If the dummy variable for female in the regression provided in Column 1 of Table 1 was replaced with a dummy variable for male (=1-female), what would be the new intercept (i.e., the coefficient on the constant).

b. Based on the regression in column 1, what is the predicted hourly wage rate for a male with 12 years of education, average looks (not above or below average) and 10 years of experience?

c. Based on the regression in column 1, what is the wage rate of someone with below average looks relative to a person with above average looks?

d. Based on column 1, the marginal effect of experience on wages is positive until experience reaches how many years?

e. Suppose that full-time workers (i.e., those that work 35 or more hours per week) are paid higher hourly wages than part-time workers. Also, suppose that more educated workers are more likely to work full-time jobs. If a dummy variable for full-time employment was added to the regression in column 1, then would the coefficient on education *increase, decrease,* or *either depending on other factors*?

The regression reported in column 1 of Table 1 was re-estimated with interactions between the female dummy variable and the dummy variables for above and below average looks. These results are reported in column 2 of Table 1.

f. Based on the regression results reported in column 2, what can we conclude about the effects of looks on the wages of men versus women?

g. Based on the regression in column 2, all else equal, what is the predicted difference in the wage rate of males with above and below average looks?

h. Based on the regression in column 2, all else equal, what is the predicted difference in the wage rate of females with above and below average looks?

i. Suppose that you believe that the effect of below average looks on wages depends on the type of job you are in. For example, you might thing that below average looks matters less for highly educated people. Describe a regression model that you could estimate to test the null hypothesis that the effect of below average looks on wages is independent of education. Write out the regression equation and describe any new variables that you would add to the earlier regressions. Also, give a precise definition of the hypothesis you would test based on the coefficients that are estimated in your regression.

j. Explain how you could tell if looks has a *smaller* effect on the earnings of college graduates based on your model.

5. Let $\widetilde{y}$ be some unobserved latent variable such that

$$y_i = \widetilde{y}_i = X_i'\beta + \varepsilon_i \text{ if } \widetilde{y}_i > 0$$

y_i is unobserved otherwise

and $\varepsilon \sim N(0, \sigma^2 I)$

Note: $\frac{\partial \Phi(z)}{\partial \theta} = \phi(z)\frac{\partial z}{\theta}$ and $\frac{\partial \phi(z_i)}{\partial \theta} = -z_i \phi(z_i)\frac{\partial z_i}{\partial \theta}$

a. What is θ, the identifiable parameter of interest in this problem?

b. Derive the probability that you observe an individual *i.*

c. Derive the contribution of each individual in your sample to the overall likelihood function (i.e., derive $L_i(\theta)$) and the individual log-likelihood function. *(5 points)*

d. Derive the score function needed to identify $\hat{\theta}_{MLE}$.

e. Explain what is implied by the simplified form of the Score function (i.e., what is the implied orthogonality condition).

6. Suppose you want to estimate the effect of childbearing (motherhood status) on labor force earnings for a sample of women in the U.S. using the following model

(1) $$Y_i = \alpha + \beta D_i + \varepsilon_i ,$$

where Y_i is the labor market earnings of woman *i,* and D_i is a dummy variable equal to one if woman *i* has had at least one child. In this way you are hoping to estimate the average treatment effect of being a mother on female labor market earnings.

a. Which of the Ordinary Least Squares assumption is likely to fail when estimating this model? Explain why? What does the mean for your estimate of the average effect of motherhood on earnings?

b. What is the "Fundamental Problem of Causal Inference"?

c. Define Y_{1i} as the earnings of woman *i* if she is a mother and Y_{0i} as the earnings of that same woman *i* if she is not a mother. If motherhood status was randomly assigned across women in the population then the treatment effect of motherhood on earnings would be equal to the following

(2) $$E_i[Y_{1i} - Y_{0i}] = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0].$$

However, motherhood status is not randomly assigned. Using the Potential Outcomes framework, decompose the expectation in (2) into the *"average treatment effect on the treated"* and *"selection bias"*.

d. You decide to instrument for motherhood using the instrumental variable Z. Which two assumptions are necessary for this instrument to be valid?

e. Would the following variables be plausible instruments for motherhood. Explain why or why not.

i. The quality of each woman's health insurance coverage
ii. An indicator of whether or not the woman has experienced infertility
iii. Regional differences in abortion laws (e.g. the oldest gestational age a woman can legally obtain an abortion in her region).
iv. Availability of family planning in a local area.
v. Number of siblings the woman has
vi. The woman's marital status

Q7. Consider the following 2-equation model:

$$InfDI(t) = C1 + \phi 11 * InfDI(t-1) + \phi 12 * InPC\_GDP(t-1) + u1(t)$$
$$InPC\_GDP(t) = C2 + \phi 21 * InfDI(t-1) + \phi 22 * InPC\_GDP(t-1) + u2(t)$$

Where *PC_GDP* = Per capita GDP, *fDI* = foreign direct investment.

a. Why is it called a seemingly unrelated regression (SUR), and not a simultaneous model?
b. Show that the OLS and GLS estimators become identical when i) the cross-error co-variances between u1 and u2 are zero, OR ii) the right hand side variables are identical
c. Derive the variance-covariance matrix of the two error vectors, u1 and u2.
d. Derive the feasible GLS estimator (explain each estimation iteration.).
e. How would you perform the Granger-Causality test to see if *fDI* causes GDP?
f. Discuss the concept of impulse response function in this setting.

Q8. The Nepal Study Center is planning to conduct a study to help a clinic in a rural village in Nepal's Gulmi District to implement a micro health insurance program. It plans to use a dichotomous choice experiment design to carry out the study. The plan is to sample 420 households randomly from the three communities that lay around the clinic --its catchment area. Each community has nine wards. The sampling will be performed by using the proportional sampling design representing all the wards from each of the clinic. The households are presented with options to enroll in one of three micro health insurance plans: Basic (clinic visits), General (clinic + plus pharmacy), Comprehensive (clinic visits, pharmacy + minor surgery). The three alternatives are presented below:

c=(Comprehensive, General, Basic)

We would expect a person's utility related to each of the three alternatives to be a function of both personal characteristics (such as income, age etc..) and characteristics of the health care plan (such as its price/premium).

We collected data would look like the table below: person's age (divided by 10), the person's household income (in Rs10,00 / month), and the price of a plan (in Rs100 / 6 months). The first three cases from the data are shown below. It is in the long form.

	HHid	MH_Alt ch	Choice	hhinc	age	Premium
1	1	Comprehensive	1	3.66	2.1	2
2	1	General	0	3.66	2.1	1
3	1	Basic	0	3.66	2.1	0.5
4	2	Comprehensive	0	3.75	4.2	2
5	2	General	1	3.75	4.2	1
6	2	Basic	0	3.75	4.2	0.5
7	3	Comprehensive	0	2.32	2.4	2
8	3	General	0	2.32	2.4	1
9	3	Basic	1	2.32	2.4	0.5

Additionally, we will also collect information on the following variables: **Receive Remittance (yes/no), No Of Children, No of Clinic Visits Per Six Month, and Distance to Clinic (minutes of walking distance).** These variables are not shown in the table to save space.

Taking the first case (**id==1**), we see that the case-specific variables **hhinc, age, Remittance, NoChildren, and Distance** are constant across alternatives, whereas the alternative-specific variable **price** varies over alternatives. Additionally, we also collected information on the following variables: **Receive Remittance, No of Children, No of Clinic Visits Per Six Month, and Distance to Clinic**

The variable **MHalt** (micro health insurance alternatives) labels the alternatives, and the binary variable **choice** indicates the chosen alternative (it is coded 1 for the chosen plan, and 0 otherwise).

1. For simplicity, consider only three variables for model set up (age, income, and price/premium).

A) Set up a Random Utility Model (RUM). Show all the steps.

B) Present the corresponding data table in the wide form as a set up for a long-hand mle coding.

C) Present the log likelihood function. Show all the steps.

D) Write the STATA codes (long-hand).

(You may assume that the income and age have the same impact on the choice functions.)

2. As you recall, clogit automatically suppresses alternative specific constants, whereas asclogit allows the constants. In the DC modeling community, there is no consensus regarding the preference for an ASC (alternate specific constant approach versus the non-ASC option). In this case, which option may make more sense and why?

3. Usually, we use some sort of clustering adjustment for the standard errors vce (cluster id). In this case, which clustering id would you use – individual id, community id, or ward id – and why?

Q9. A group of 540 pregnant women from several rural villages were sampled to study their prenatal visit to clinics. The independent variables were: age, education, distance to clinic, household income, and presence of husband at home. The dependent variable – visits—was recorded as the count values (0, 1, 2, 5 etc.). There were many 0's too.

a. Write out the regression function (equation).
b. What are the expected signs on the independent variables?
c. Is this a demand function? Why or why not?
d. Set up a Poisson modelling framework, and spell out the log likelihood function. Show all the steps.

There were many women with a 0 visit (with no visit recorded over the last six months), leading a problem of "excess zeros". This causes a problem known as "over dispersion." You have a few options to deal with this situation:

e. Given the count nature of the data you could pick a) Zero inflated Poisson framework OR, Negative Binomial (Type II). Alternatively, although not perfect, c) a censored method (Tobit) or d) hurdle method could also be used to address the problem with 0's.

**Choose one of the four options and present your derivations (log-likelihood expression).

**Given the following information: c = 1.31, e = .67; Average of Visit = 1.82, Average of age = 32, average of distance = 31 minutes, Average of Female = .51

f. What is the interpretation of the coefficient "c?"
g. What is the interpretation of the coefficient "e?"
h. Show the calculation / formula and the estimated marginal effect of income on visit.
i. Now, calculate the income elasticity of visit.