PhD/MA Econometrics Examination

August 2018

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)

Q1. Fundamentals of Econometrics

a. Define
    i. Econometrics
    ii. A statistic
    iii. An estimate
    iv. An estimator
b. Write out an OLS model.
    i. Define your variables (include at least 4 independent variables) and describe your model.
    ii. Write out the hypothesis for each coefficient (null and alternative in mathematical notation) and use economic theory to defend your expected result (sentences).
c. Define multicollinearity and describe how you would test for it. If multicollinearity exists in theory but not in practice, what should you do?
d. In words and mathematical notation describe how to deal with measurement error in your data:
    i. When the error is in your dependent variable
    ii. When the error is in one of your independent variables

Q2. Ordinary Least Squares (OLS)

a. Write out the OLS equation in matrix form.
b. Write out the matrices and state their dimensions.
c. State the classical assumptions – in words and equations.
d. Derive the normal equations.
e. Demonstrate that the OLS estimator is BLUE.
f. From $X'e = 0$, we can derive several properties. State these properties. Hint: there are 6 and 5 of them require that the OLS regression includes a constant.

Q3. The Variance of Least Squares

We know $var(b) = \sigma^2(X'X)^{-1}$, but σ^2 is an unknown parameter. Therefore, to find $\widehat{var}(b)$, we need to find a good estimator for σ^2.

a. Derive that estimator.
b. Write down the standard error of b.
c. What is the standard of error of b used for? Or explain "why did we derive it?"

PART B: Answer any Two

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

Q4. Consider the model

$$y_t = X\beta + \varepsilon_t$$

$$\text{where } \operatorname{var}(\varepsilon_t) = \Sigma \neq \sigma^2 I$$

a. Which OLS assumption fails? What are the implications of that failure for the OLS estimator?

b. Derive the properties for the OLS estimator in this scenario (i.e., what is the mean and variance of the OLS estimator for β).

c. Assuming $\Sigma = \sigma_\varepsilon^2 \Omega = (P'P)^{-1}$, derive the GLS estimator for β. Show that $\hat{\beta}_{GLS}$ is unbiased and the variance for the GLS predicted residual is $\sigma^2 I$

d. Suppose the residual ε_t takes the following form

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

$$\text{where } u_t \sim N(0, \sigma_u^2)$$

What is the stationarity assumption? Derive the properties of ε_t assuming the stationarity assumption holds (i.e. derive the mean and variance of ε_t).

e. Derive the correlation between ε_t and ε_{t-s}, where $s \geq 1$ in the scenario presented in c.

f. What do the matrices for Σ and Ω look like in this scenario, where $\Sigma = \sigma_\varepsilon^2 \Omega$.

Q5. Let $\tilde{y}$ be some unobserved latent variable such that

$$\tilde{y} = x\beta + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma^2)$$

You observe y_i and x_i, $i = 1,...,N$, such that

$$y_i = \begin{cases} 1 & \text{if } \tilde{y}_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Define $\phi(\theta)$ as the pdf for a standard normal and $\Phi(\theta)$ as the cdf for the standard normal.

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

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

b.) Derive the probabilities that $y_i = 1$ and $y_i = 0$ for 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.

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).

Q6. A researcher using data for a sample of 3240 female employees 25 years of age and over to investigate the relationship between employees' hourly wage rates Y_i (measured in *dollars per hour*) and their age X_i (measured in *years*). The population regression equation takes the form of equation (1) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Preliminary analysis of the sample data produces the following sample information:

$N = 3240 \qquad \sum_{i=1}^{V}(y_i - \bar{y})^2 = 78434.97 \qquad \sum_{i=1}^{N}(x_i - \bar{x})^2 = 25526.17$

$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = 3666.426 \qquad \sum_{i=1}^{N} y_i = 34379.16 \qquad \sum_{i=1}^{N} x_i = 96143.00$

$\sum_{i=1}^{N} y_i^2 = 443227.1 \qquad \sum_{i=1}^{N} x_i^2 = 2878451.0 \qquad \sum_{i=1}^{N} x_i y_i = 1023825.0$

$\sum_{i=1}^{N} e_i^2 = 77908.35$

g. Use the above information to compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1. *(Hint: Remember the formula for beta based on the normal equations)*

h. Interpret the slope coefficient estimate you calculated in part a—i.e., explain in words what the numeric value you calculated for $\hat{\beta}_1$ means.

i. Calculate an estimate of s^2, the estimated error variance. *(Remember that you technically have 2 explanatory variables since there is a column of ones multiplied by the intercept, meaning k=2)*

j. Calculate the estimate of $\text{var}(\hat{\beta}_1)$.

k. Compute the value of R^2, the coefficient of determination for the estimated OLS sample regression. Briefly explain what the value that you have calculated for R^2 means.

l. Calculate the sample value of the t-statistic for testing the null hypothesis $H_o : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. *(Note: You are not required to obtain or state the inference of this test. Just calculate the test-statistic itself).*

PART C: Answer any Two

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

Q7. Given the following 2-equation system,

$$ChildLabor_t = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * ChildAge_t + \alpha_3 * FatherRemittance_t + u1_t$$
$$FatherRemittance_t = \beta_0 + \beta_1 * Poverty_t + \beta_2 * FatherAge_t + u2_t$$

FatherRemittance variable captures the absence of fathers working overseas sending money back home to the family.

a. How would you estimate the *child labor* equation using a 2-sls method? Briefly discuss the two steps.
b. Likewise, how would you estimate the *Remittance* equation? Explain the necessary steps.
c. Set up this system in a grand matrix notation

$$Y = Z * \beta + U \quad (1)$$

d. Derive the Var-Cov(U). You can use the generic notation, where *u1 and u2* are stacked within *U*.
e. Discuss the iterative 3-SLS estimation method to obtain the estimate of β and its variance-covariance matrix; show all the necessary steps and derivations.
f. What is the benefit of using a 3-SLS method over the 2-SLS?
g. BONUS exercise: how would you set up a model like this as a FIML?

Q8. 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).

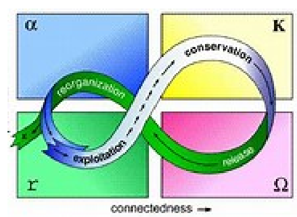
For simplicity, consider only three variables for model set up (age, income, and price).

A) Set up a Random Utility Model (RUM). Show all the steps.
B) Present the log likelihood function. Show all the steps. (You may assume that the income and age have the same impact on the utility functions.)
C) Show an example of the data table for this problem. D) What are differences between the simple logit, the multi-nomial logit, and the conditional logit?

Q9. Consider the following conceptual framework around the concept of boom and bust (collapse-and-regeneration) within a social-ecological systems:

> "Creating institutions to meet the challenge of sustainability is arguably the most important task confronting society; it is also dauntingly complex. Ecological, economic, and social elements all play a role, but despite ongoing efforts, researchers have yet to succeed in integrating the various disciplines in a way that gives adequate representation to the insights of each.
>
> *Panarchy*, a term devised to describe evolving hierarchical systems with multiple interrelated elements, offers an important new framework for understanding and resolving this dilemma. Panarchy is the structure in which systems, including those of nature (e.g., forests) and of humans (e.g., capitalism), as well as combined human-natural systems (e.g., institutions that govern natural resource use such as the Forest Service), are interlinked in continual adaptive cycles of growth, accumulation, restructuring, and renewal. These transformational cycles take place at scales ranging from a drop of water to the biosphere, over periods from days to geologic epochs. By understanding these cycles and their scales, researchers can identify the points at which a system is capable of accepting positive change, and can use those leverage points to foster resilience and sustainability within the system."

Diagrammatically, the collapse function is represented by the *green function* and the regenerative phase is represented by the *grey function*:



Context: In the far north-west region of Mustang Valley near the border of Nepal-Tibet, two villages began facing severe water crisis as a result of the extend drought period, thus causing these villages (and villagers) to relocate (Swiss Study). The *green cycle* represents the collapsing of the social-ecological systems.

The following non-linear model is postulated to represent the collapse phase:

$$P = \frac{\alpha}{[1 + \exp\{\mu * (x - \beta)\}]} + u$$

Where, x = annual drought index. β = delay parameter around which time the collapse begins to transition from bad to worse; P = Agriculture production index (mostly from grazing, apple farming, and high altitude herbal plants, all seem to have been affected by the drought.).

a. Do you know what α represents?

b. If you were to simulate this collapse scenario, what sign (positive or negative) would you assign to the parameter: μ?
c. Present the step-by-step method of estimating this function using the non-linear least squares (use the algorithm of your choice, e.g.: Gauss-Newton).

OR

Set up a maximum likelihood function and discuss the optimization algorithm (derivation of Newton-Raphson).