PhD/MA Econometrics Examination

January, 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)

Q1. Use the table below. "HWSEI is a constructed variable that assigns a Hauser and Warren Socioeconomic Index (SEI) score to each occupation using the modified version of the 1990 occupational classification scheme available in the OCC1990 variable. The HWSEI variable is a measure of occupational status based upon the earnings and educational attainment associated with each category in the 1990 occupational scheme."

Source	SS	Df	MS	Number of obs	=	12104
				F(???,???)	=	498.53
Model	1662649.59	13	127896.1	Prob > F	=	0
Residual	3101628.43	12090	256.5449	R-squared	=	???
				Adj R-squared	=	0.3483
Total	4764278.02	12103	393.6444	Root MSE	=	16.017

HWSEI	**Coef.**	**SE.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Age	1.4378	???	20.55	???	???	???
Age^2	-0.0174	0.0009	-20.44	0.0000	-0.0190	-0.0157
Female	-2.2691	???	???	???	-2.8422	-1.6959
Nursery school to grade 4	-1.4976	2.4374	-0.61	0.5390	-6.2754	3.2801
Grade 5, 6, 7, or 8	-2.3946	1.7404	-1.38	0.1690	-5.8061	1.0169
Grade 9	-1.9035	1.7752	-1.07	0.2840	-5.3831	1.5761
Grade 10	-1.0185	1.7148	-0.59	0.5530	-4.3798	2.3429
Grade 11	0.8919	1.6985	0.53	0.6000	-2.4373	4.2211
Grade 12	8.0669	1.5618	5.17	0.0000	5.0056	11.1283
1 year of college	12.1170	1.5784	7.68	0.0000	9.0230	15.2109
2 years of college	17.0264	1.6285	10.46	0.0000	13.8342	20.2186
4 years of college	26.0401	1.5930	16.35	0.0000	22.9176	29.1626
5+ years of college	35.9438	1.6119	22.3	0.0000	32.7843	39.1033
Constant	-8.8311	2.0393	-4.33	0.0000	-12.8284	-4.8337

Omitted groups: "male" and "no school or N/A"

a) Interpret the coefficient (and other relevant information) on “Grade 11.” Grade 11 is a dummy variable for people who have completed through 11th grade, i.e. one year short of finishing high school.

b) Now, looking at the coefficient on the variable Grade 12, is there an important lesson that we can learn?

c) Calculate R^2. Explain your answer. What does the number mean?

d) Find the degrees of freedom for the F.

e) What is the F value? What does it tell you? Can you calculate it from the information given above? If you can calculate F, interpret the value.

f) Calculate the missing values in the row labeled “Age.”

g) Calculate the marginal effect of age from the regression results.

h) Calculate the missing values in the row labeled “Female.”

i) What is the “Dummy Variable Trap”?

*** You should have filled in 10 blanks (**???**) in the table. ****

Q2. Fundamentals of OLS

a. Write out the OLS equation in matrix form. Also, write out the matrices and state their dimensions.
b. State the OLS assumptions in mathematical statements and in sentences (words).
c. Show that the OLS estimator is BLUE and define BLUE. Show all parts: B, L, U, and E.
d. What are the properties (hint: there are six) of the OLS estimator? State them in mathematics and words. Also, state any requirements which are necessary for these properties to hold.
e. Given the properties in part d, what can you infer about the disturbances from the residuals?
f. Write out a simple OLS model. Define your variables and describe how your model might meet or not meet all the assumptions you stated above.

Q3. 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 central moments – *name them do not calculate them.*
i. Define and draw the PDF and CDF of the distribution.

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.]

Q4. You have been commissioned to investigate the relationship between the birth weights of newborn females and the number of prenatal visits to a physician or midwife that their mothers made during pregnancy. The dependent variable is $bwght_i$, the birth weight of the *ith* newborn female measured in *grams*. The explanatory variable is $pnvisits_i$, the number of prenatal visits of the *ith* newborn's mother during pregnancy, measured in *number of visits.* The model you propose to estimate is given by the population regression equation

$$bwght_i = \beta_0 + \beta_1 pnvisits_i + u_i \,.$$

Your research assistant has used 857 sample observations on $bwght_i$ and $pnvisits_i$ to estimate the following OLS sample regression equation, where the figures in parentheses below the coefficient estimates are the *estimated standard errors* of the coefficient estimates:

$$\underset{(65.6909)}{bwght_i = 3199.02} + \underset{(5.36347)}{14.1219}\, pnvisits_i + \hat{u}_i \qquad i = 1,\ldots,N \text{, N=857}$$

a. Perform a test of the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis: $H_A : \beta_1 \neq 0$ at the 1% significance level (i.e., for significance level $\alpha = 0.01$). Note, the critical value for the t-distribution at the one percent level is 2.58. Show how you calculated the test statistic. State the decision rule you use, and the inference you would draw from the test. What would you conclude form the results of this test?

b. Compute the two-sided 95% confidence interval for the intercept coefficient β_0. Use this two-sided 95% confidence interval for β_0 to test the hypothesis that the mean birth weight of newborn females whose mothers made no prenatal visits to a physician or midwife equals 3,000 grams. State the null hypothesis H_0 and the alternative hypothesis H_A. State the decision rule you use, and the inference you would draw from the test. Note, at the 95% confidence level, the critical value for the t-distribution is 1.96.

c. What is the interpretation of the coefficient on $pnvisits_i$?

d. What assumption is necessary in order to interpret the coefficient on What is the interpretation of the coefficient on $pnvisits_i$ as causal? Do you think that assumption is satisfied? Why or why not?

Q5. Suppose $y_1, y_2, \ldots, y_n$ be iid with a poisson distribution such that

$$f(y_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

a. What is the likelihood of observing your data (i.e., what is the likelihood function for your sample)?

b. Derive the log likelihood and score functions for estimating the parameter λ.

c. Derive the Maximum Likelihood Estimator for λ.

Derive the asymptotic variance for $\hat{\lambda}_{MLE}$ using the information matrix method *(Hint: Remember that $E[\theta] = \theta_{MLE}$ because MLE gives a consistent estimate of the parameter. Also use the relationship derived in part c to substitute out all random variables from the asymptotic variance).*

Q6. 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 $\Sigma = \sigma_\varepsilon^2 \Omega$.

PART C: Answer any Two

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

Q7. 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).

Q7.1. For simplicity, consider only three variables for model set up (age, income, and price/premium) and a binary decision –Basic versus Otherwise options.

Y(t)* = a+ b*age(t) + c*income(t)+ u(t)

where if y*(t) >0 => y(t) = 1 (Non-basic option chosen) else 0 if option Basic chosen.

7.1.1 Set up a RUM framework with linear indirect utilities to analyze this model. Show all the steps and utilities etc..

7.1.2 Derive the log-likelihood function using the logistic assumption.

Q7.2 Now, add the premium variable to this model and consider all three options: basic, general, and comprehensive.

Y(t)* = a+ b*age(t) + c*income(t)+ d*premium(t) +u(t)

7.2.1 Set up a RUM framework with linear indirect utilities to analyze this model. Show all the steps in details. (You may assume that the age and price have the same impact on the preference functions, but allow income to have different impact.)

7.2.2 Present the corresponding data table in the wide form as a set up for a long-hand mle script.

7.2.3 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?

Q8. Discuss the following with proper notations (whenever appropriate) and examples (NOT JUST WORDS and DEFINITIONAL SENTENCES):

- 8.1 Mulitinomial Logit versus conditional logit.
- 8.2 Extreme value distribution and logistic distribution.
- 8.3 Multinomial logit versus Ordered logit.
- 8.4 Multinomial logit and its limitation, and the Nested Logit and its benefit.
- 8.5 Poisson, Negative Binomial, Zero-Inflated, and Hurdle

Q9. Consider the following 2-equation model for a two-country tariff retaliation:

$$Tariff\_US(t) \quad = C1 + \phi 11 * Tariff\_US(t-1) \quad + \phi 12 * Tariff\_China(t-1) + u1(t)$$
$$Tariff\_China(t) = C2 + \phi 21 * Tariff\_US(t-1) \quad + \phi 22 * Tariff\_China(t-1) + u2(\text{t})$$

a. Derive the var-cov matrix of the error vector of this 2-equation model. Show steps in detail.
b. Discuss the step-by-step GLS method of estimation.
c. Why is it called a seemingly unrelated regression (SUR), and not a simultaneous model?
d. Show that the OLS and GLS estimators become identical when 1) the cross-error co-variances between u1 and u2 are zero OR 2) the right hand variables are identical.

Set up the FIML log-likelihood for this model.