PhD/MA Econometrics Examination

August 2020

Total Time: 8.5 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 suffice.

PART A

(Answer any TWO from Part A)

1. ***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?

2. 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 several properties. State these properties. Hint: there are 6 and 5 of them require that the OLS regression includes a constant.
e. Write out a simple OLS model. Define your variables and describe how your model might meet or not meet the assumptions you stated above.

Question 3

Table 1

Source	SS	df	MS	Number of obs	*blank*
				F(*blank*)	*blank*
Model	2.1523e+12	4	5.3807e+11	Prob > F	0.000
Residual	1.1246e+14	83,081	1.3536e+09	R-squared	*blank*
Total	1.1461e+14	83,085	1.3794e+09	Root MSE	34523

Income	Coef.	Std.Err.	t	P>t	[95% Conf.	Interval]
Age	2437.933	137.7278	17.70	0.000	2167.987	2707.878
Age^2	-24.74765	1.658123	-14.93	0.000	-27.99756	-21.49774
Family Size	-4644.332	238.2059	*blank*	*blank*	*blank*	*blank*
Family Size^2	231.6115	17.81798	13.00	0.000	196.6884	266.5346
Constant	-15827.59	2886.193	-5.48	0.000	-21484.5	-10170.67

'*blank*' is intentional.

Table 2

Source	SS	df	MS		Number of obs	****??????
					F(5, 12396)	****??????
Model	2.5761e+12	5	5.1522e+11		Prob > F	0.000
Residual	1.1203e+14	83,080	1.3485e+09		R-squared	****??????
Total	1.1461e+14	83,085	1.3794e+09		Root MSE	34203

Income	Coef.	Std.Err.	t	P>t	[95% Conf.	Interval]
Age	2257.733	137.8441	16.38	0.000	1987.559	2527.906
Age^2	-22.88542	1.658335	-13.80	0.000	-26.13574	-19.63509
Family Size	-5493.036	242.5295	****??????	****??????	****??????	****??????
Family Size^2	286.9267	18.0561	15.89	0.000	251.5368	322.3165
Married	****??????	281.937	17.73	0.000	4445.788	5550.977
Constant	-12485.69	2886.927	-4.32	0.000	-18144.04	-6827.334

3. Regression Output Interpretation, Calculations, Definitions, and More

REFERRING ONLY TO TABLE 2 (parts a – f)

Use mathematical statements and words in your answers.

a. Fill in the missing values ***??????*** in Table 2. (There are 8 numbers to calculate.)
b. Define and interpret R-squared
c. Explain the relationship between F and R-squared
d. Define and interpret the 95% CI for "Family Size"
e. Interpret the coefficient on "Married"
f. Calculate the marginal effect of being one year older than the mean age (40.54746) of the sample.

COMPARING TABLES 1 & 2 (part g)

g. Why does adding "Married" create a relatively large change on the coefficients on Family Size and Family Size^2 but cause relatively small change the coefficients on Age and Age^2?

MORE (part h)

h. If you were going to add a variable to this regression, what variable would you add? Explain the variable choice (why?), what the expected sign would be, and how you interpret the coefficient on your newly added variable.

Did you calculate the missing numbers (??????**) in Table 2 (part a)?*

Part B: Answer Any Two

Question 4:

4. Suppose $\tilde{y}$ is an unobserved latent variable that measures an individual's economic productivity (which can be proxied by hourly earnings), such that:

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

However, you do not observe earnings in your data. You only observe, y_i, which indicates whether an individual is working or not. $y_i = 1$ if an individual participates in the labor force and $y_i = 0$, otherwise. An individual participates in the labor force if he/she is able to earn wages above some reservation wage, w .

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. Define y_i in terms of $\tilde{y}_i$ and w.

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

c. Derive the probabilities that $y_i = 1$ and $y_i = 0$ for individual *i.*

d. *D*erive the contribution of each individual in your sample to the overall likelihood function (i.e., derive $\mathcal{L}_i(\theta)$) and the individual log-likelihood function.

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

Now suppose you observe the earnings of each individual *only if* he/she participates in the labor market, such that $y_i = \tilde{y}_i$ if the individual participates in the labor market. However, if the individual does not participate in the labor market, then all you know is that she/he is not working but not what her/his earnings would have been if she/he did participate. So, for all

individuals not participating in the labor market, $y_i = 0.$, since all you know is that this individual is not working but not what her/his earnings would have been if she/he did work.

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

a. Now redefine y_i in terms of $\widetilde{y}_i$ and w under this new setup.

f. Derive the probability that you observe each individual *i.* Assume $w = 0$, from this point forward.

g. *D*erive the contribution of each individual in your sample to the overall likelihood function (i.e., derive $\mathcal{L}_i(\theta)$) and the individual log-likelihood function.

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

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

Question 5

5. Suppose the government is concerned about increasing rate of adolescent vaping in the U.S. and is thinking about implementing a tax on vape juice to combat it. However, first the government would like to know how responsive teen vaping is to changes in price and is asking you to estimate the vape juice demand function among the teenaged population.

   Teen demand for vape juice in as follows:

$$q_j^d = \gamma_0 + \gamma_1 p_j + \varepsilon_j, \quad (1)$$

   where q_j^d is teen demand for vape juice in county j, p_j is the price, and ε_j is the error term, which also includes demand shifters.

   a. Define and explain the five assumptions required to interpret an Ordinary Least Squares (OLS) estimate of a slope parameter as "BLUE."
   b. When the exogeneity assumption fails, we say that the estimate is endogenous. What are the four sources of endogeneity bias we discussed in class? Explain each of them.
   c. If you were to regress (1) using OLS, which of the OLS assumptions is likely to fail? Explain why? What does this mean for your estimate for the slope of your demand equation?

   Supply for vape juice is as follows:

$$q_j^s = \delta_0 + \delta_1 p_j + u_j, \quad (2)$$

   where q_j^s represents vape juice supply in county j, p_j is the price, and u_j is the error term, which also includes supply shifters.

   Note: supply and demand shifters are independent, thus, $cov(\varepsilon_j, u_j) = 0$

   d. Assuming market equilibrium (i.e., $q_j^d = q_j^s = q_j$), using equations (1) and (2), derive an expression for P_i as a function of the slope parameters and error terms. How does this expression relate to the failed OLS assumption you named in (c)? What is the underlying source of this failure?

e. Using the expression for P_i derived in (d), derive an expression for $cov(P_i, \varepsilon_i)$ as a function of the slope parameters and the variance of u_i. What is the sign of $cov(P_i, u_i)$? What does this mean about the correlation between price and the error term? Hint*: substitute the expression you derived in (c.) for p. Another Hint:* $cov(aX + bY, X) = cov(aX, X) + cov(bY, X) = acov(X, X) + bcov(Y, X) = avar(X) + bcov(Y, X)$.

f. Now show that $cov(p_j, q_j)$ can be written as follows: $cov(p_j, q_j) = \gamma_1 var(p_j) + cov(p_j, \varepsilon_j)$. To do so, use the expression for demand denoted by equation (1) and substitute it for q_j in $cov(p_j, q_j)$. Note, you do not need anything you derived in (e) to answer this question.

g. The probability limit of the OLS estimate for γ_1 is as follows: $plim(\gamma_1^{ols}) = \frac{cov(p_j, q_j)}{var(p_j)}$. Using the expression for $cov(p_j, q_j)$ introduced in (f), calculate the asymptotic bias of $plim(\gamma_1^{ols})$ (i.e., calculate $plim(\gamma_1^{ols}) - \gamma_1$). Give what we know from part (e), what is the sign of this bias? What does this mean about the OLS estimate of γ_1 (i.e., does OLS over- or underestimate γ_1)?

h. The nicotine found in vape juice is sourced from tobacco. China produces approximately 40% of the world's tobacco. Suppose you decide to instrument for p_j in equation (1), using growing season temperature in the tobacco producing regions of China as your instrument. Do you think this is a valid instrument for vape juice price? Explain why or why not and be sure to include a discussion of each of the requirements for an instrument to be valid.

i. Assume that temperature is a valid instrument for p_j in equation (1). Denote temperature as Z. Using the variable names in equation (1) and Z, describe the two-stage least squares process. Clearly define your first and second stages. Derive and expression for γ_1^{IV} based on these two stages

Question 6

6. You are examining the correlation between witnessing domestic violence among parents on an individual's experience of domestic violence as an adult—otherwise known as the intergenerational correlation of intimate partner violence. The Philippines is an interesting place to study intimate partner violence (IPV) where reported female perpetrated IPV is as common as male perpetrated IPV. You are therefore interested in comparing the intergenerational correlation of IPV across men and women. Using data on 477 Filipino men and women who are all either married or cohabitating you estimate the following model:

$$IPVindex = \beta_0 + \beta_1 PV + \beta_2 male + \beta_3 PV \times male + \varepsilon \quad (3)$$

where
$IPVindex =$ an index of the level of violence experienced by individual i perpetrated by her/his partner in the last year. The index increases with the level of violence
$PV =$ 1 if individual *i* remembers violence between her/his parents as a child and is 0 otherwise
$male =$1 if individual *i* is male and 0 if female.

You run an OLS regression on equation (3). The results of this regression are reported in column 1 of Table 1 below.

Table 1: OLS Regression Results

	Coefficients	
	Model 1	**Model 2**
Witnessed Parental Violence	0.2662831	0.154842
	(0.0740957)	(0.057451)
Male	0.1531792	0.0224423
	(0.0802545)	(0.0584512)
Witnessed Parental Violence X Male	-0.2754482	
	(0.1164904)	
Constant	-0.1318646	-0.0814228
	(0.04985)	(0.0452727)
N	477	477
R-squared	0.0271	0.0156

Standard errors are in parentheses

a. What does the R-squared tell you about the regression?
b. Please interpret the values of each of the estimated coefficient other than the constant (i.e., $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\beta_3}$) reported in column 1 of Table 1.
c. Construct a 95% confidence interval for $\widehat{\beta_1}$. The critical value is 1.96. *Please note that the standard errors reported in Table 1 are already normalized by the square root of the sample size. In other words, the reported standard errors for each coefficient are equal to* $\sigma/\sqrt{N}$.
d. Derive the marginal effect of PV on the $IPVindex$. What is the marginal effect for males and what is it for females? What does this tell you about the intergenerational correlation of IPV for this sample in the Philippines?
e. Now suppose you wanted to test that there is not actually a differential effect across sexes of parental violence on experience of IPV as adult. If this is the case, then you can drop the interaction variable from your model. You therefore estimate the following restricted model:

$$IPVindex = \alpha_0 + \alpha_1 PV + \alpha_2 male + \alpha_3 PV \times male + \omega \qquad (4)$$

Why is equation (3) referred to as the unrestricted model and equation (4) the restricted model?

f. Test the hypothesis that the restricted model is the correct model by constructing an F-statistic, which tests that the two models are actually statistically equivalent. Clearly walk through the steps of this test and state the conclusion of the test.
g. The p-value of the F-statistic you calculated in part (f) has a p-value of 0.0185. What does this p-value tell you about your null hypothesis?

Part C: Answer Any Two

Question 7

Q7. 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 that 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).

Q1.1. 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 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.

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

Q1.2. Note: cogit 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?

Q1.3. 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?

Question 8

Q.8 Using the number of doctors visit –demand for health care access—model, spell out a few modeling options. Let's postulate the following relationship

DocVisit are influenced by age, income, distance, NoChildren

Which can be specified as:

$$Y^* = a+b*age+c*income+d*distance+e*Female + u$$

a. Set up a Poisson modelling framework, and spell out the log likelihood function. Show all the steps.
b. In this case, do we need an exposure variable? Why or why not?
c. What are the expected signs on the independent variables?
d. There will be obviously many people with a 0 entry (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 couple of options to deal with this situation:

   Zero inflated Poisson framework (ZIP)
   Negative Binomial (Type II)
   Although less-desirable, a Tobit option is also available to deal with the zeros with the assumption that the dependent variable is continuous and not a count.

   Set up the log likelihood function (with all the steps spelled out properly in detail) for **ONE** of the three data generation processes ZIP or NB-II or Tobit.

Question 9

Q9. Consider the following 2-equation model for a cross-country inflation transmission:

$$InfUS(t) = C1 + \phi 11 * InfUS(t-1) + \phi 12 * InfMexico(t-1) + u1(t)$$
$$InfMexico(t) = C2 + \phi 21 * InfUS(t-1) + \phi 22 * InfMexico(t-1) + u2(t)$$

a. Why is it called a seemingly unrelated regression (SUR), and not a simultaneous model?
b. Derive the var-cov matrix of the two cross errors u1(t) and u2(t) or the vector U. Show all the steps.
c. Derive the GLS estimator.
d. Show that the GLS estimator reduces to OLS if the right hand side variables are identical or if the cross-error co-variances between u1 and u2 are zero