PhD/MA Econometrics Examination
Part A
January, 2023

(Answer any TWO of the following three questions)

A1. Statistics

a. Explain the pairwise relationship of the following distributions: normal, t, chi-square and F. Be specific on their degrees of freedom.

b. Suppose a random sample of 100 observations is drawn from a normal distribution with mean μ and variance σ^2. Find the 90% confidence interval for σ^2 as a function of variance s^2. Critical values can be denoted algebraically.

c. For a random variable x following log normal distribution with

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right],\ x>0.$$

Let $y=\ln(x)$, find the pdf of y.

d. For a sample $y_1, y_2, \ldots y_n \sim N(\mu, \sigma^2)$, find the estimates of μ *and* σ^2 using the MLE approach.

A2. OLS estimation

For a dependent variable vector *z* with *n* observations, its corresponding independent variable matrix is T with λ variables, parameter vector is θ and residual vector is μ.

a. Derive the OLS solution of parameter vector estimate θ.

b. Explain what is the heteroskedasticity issue and its problems on estimation results.

c. Using the dataset to run a restricted regression, which is a regression without a constant. The new residual vector is *e* and the goodness of fit is denoted as R_1^2. The original full regression goodness of fit is denoted as R_0^2. Write down the formula for R_1^2 and R_0^2, and compare which is bigger.

d. Derive the distribution of estimated parameter vector, assuming the normality of residual distribution as $N(0, \sigma^2 I_n)$.

e. For parameter θ_3, suppose we know $\widehat{\theta_3} - \theta_3$ is asymptotically N(0, σ^2/n), use the delta method to derive a formula to test the significance level of its transformation $g(\widehat{\theta_3}) = (\widehat{\theta_3})^{\frac{1}{2}} - 1$.

A3. Regression application

A multiple regression of Y on a constant, X_1 and X_2 produces the following results:

Regression Statistics	
R Square	0.92
Adjusted R Square	0.92
Standard Error	...
Observations	...

ANOVA

		Sum of Squared	*F-stat*	*Significance F*
Regression	SSE	2091	286.78	0.00
Residual	SSR	171		
Total	SST	2262		

	Coefficients	*Standard Error*	*t Stat*	*Lower 95%*	*Upper 95%*
Intercept	1.71	...	...	...	...
X_1	1.13	...	...	...	...
X_2	2.00	...	...	...	...

$X'X =$

50	92	127
92	364	324
127	324	680

$(X'X)^{-1} =$

0.0463	-0.0069	-0.0053
-0.0069	0.0058	-0.0015
-0.0053	-0.0015	0.0032

X'Y =

443
1217
1943

a. What's the number of observations and the sample mean of y, x1 and x2?

b. What's the sample variance of y, x1 and x2?

c. What's the correlation between x1 and x2? What's and the unbiased sample variance estimate of σ^2 ? Fill the missing Standard Error estimate in the table.

d. Calculate the covariance matrix var(b) of the estimated parameter vector.

e. Based on d, fill in the standard errors and t-stat values (use an approximate critical value). Evaluate their significance.

PART B

(Answer any TWO of the following three questions)

B1. You are interested in the correlation between witnessing domestic violence among parents and the likelihood of adolescent drug use in the Philippines. Using data on 1,992 Filipino male and female 18 year-olds you estimate the following model:

$$Drugs_i = \beta_0 + \beta_1 PV_i + \beta_2 male_i + \varepsilon_i \quad \text{(B.1)}$$

where

$Drugs_i =$ 1 if individual *i* had tried drugs by the time she/he was 18 years old and is 0 otherwise

$PV_i =$ 1 if individual *i* remembers violence between her/his parents as a child and is 0 otherwise

$male_i =$ 1 if individual *i* is male and 0 if female.

You run an OLS (i.e., linear probability model) regression on equation (B.1). 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	0.0521824	0.0151624
	(0.0149529)	(0.0218199)
Male	0.2014499	0.1690732
	(0.0149453)	(0.0204057)
Witnessed Parental Violence X Male		0.0696621
		(0.0299318)
Constant	0.0105386	0.0272904
	(0.0128062)	(0.014678)
N	1,992	1,992
R-squared	0.0890	0.0910

Standard errors are in parentheses

a. What does the R-squared in column 1 of Table 1 tell you about the regression?
b. Please interpret the values of each of the estimated coefficients other than the constant (i.e., $\widehat{\beta_1}$ and $\widehat{\beta_2}$) reported in column 1 of Table 1.
c. If the dummy variable for male in the regression provided in Column 1 of Table 1 was replaced with a dummy variable for female (=1-male), what would be the new intercept (i.e., the coefficient on the constant).
d. 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* $\sigma/\sqrt{N}$.

Now suppose you wanted to test that there is a differential effect across sexes of parental violence on adolescent drug use. To check for this, you add an interaction between male and witnessing parental violence and estimate the following restricted model:

$$Drugs_i = \alpha_0 + \alpha_1 PV_i + \alpha_2 male_i + \alpha_3 PV_i \times male_i + \omega_i \qquad \text{(B.2)}$$

e. Why is equation (B.1) referred to as the unrestricted model and equation (B.2) the restricted model?
f. Please interpret the values of each of the estimated coefficients other than the constant (i.e., $\widehat{\alpha_1}$, $\widehat{\alpha_2}$ and $\widehat{\alpha_3}$) reported in column 2 of Table 1
g. Derive the marginal effect of PV_i on $Drugs_i$. What is the marginal effect for males and what is it for females? What does this tell you about the correlation between parental violence and adolescent drug use in the Philippines?
h. Based on the regression in column 2, what is the predicted likelihood of trying drugs for a male who witnessed parental violence?
i. 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.
j. The p-value of the F-statistic you calculated in part (i) has a p-value of 0.020. What does this p-value tell you about your null hypothesis?
k. Finally, suppose that you believe that the effect of parental violence on adolescent drug use depends on your wealth status. For example, you might think that the influence of parental violence is stronger for households below the poverty line. Describe a regression model that you could estimate to test the null hypothesis that the effect of parental violence on drug use is independent of poverty status. 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.

B2. Causal Analysis

Head Start is a federally funded preschool program in the U.S. that promotes school readiness and provides education, health, nutrition and parent involvement services to low-income children and their families. To be eligible to participate in Head Start, a child must be between the ages of 0 to 5 years and be from a low-income family as defined by the U.S. Federal Poverty Guidelines. To understand whether or not this program improves later school outcomes you are interested in estimating the impact of Head Start participation on later test scores. Therefore, you estimate the following OLS Model:

$$Y_i = \beta_0 + \beta_1 HS_i + \varepsilon_i \quad \text{(B.3)}$$

where

Y_i	=the 3rd grade achievement test score of child *i*
HS_i	=1 if child *i* previously participated in Head Start and 0 otherwise

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 three potential sources of endogeneity (generally, not specific to the model in B.3). Explain each of them.
c. If you were to regress (B.3) using OLS, which of the OLS assumptions is likely to fail? Explain why? What does this mean for your estimate for the treatment effect of Head Start participation on later test scores?
d. What is the "Fundamental Problem of Causal Inference". Please define it and explain what it means for empirical analysis.
e. Define Y_{1i} as the 3rd grade test scores of child i if she/he participated in Head Start and Y_{0i} and the test score of that same child if she/he did not participate in Head Start. If Head Start participation were randomly assigned across children in the sample, then the treatment effect of HS_i on Y_i is as follows:

$$E_i[Y_{1i} - Y_{01}] = E[Y_i|HS_i = 1] - E[Y_i|HS_i = 0], \quad \text{(B.4)}$$

However, Head Start participation *is not* randomly assigned. Using the Potential Outcomes Framework, decompose the expectation in equation (B.4) into the *"average treatment effect"* and *"selection bias"*. Say in words what is captured by the average treatment effect term and the selection bias terms that you derive.

f. You decide to instrument Head Start Participation using the instrumental variable, Z. Which two requirements are necessary for this instrument to be valid? Please define both mathematically and with words.
g. Using Z as your instrument, derive the GMM-IV estimator for β_1.
h. Now describe how you would estimate β_1 using the two stage least squares (2SLS) and using Z as your instrument. Be sure to specify the equations you would use.
i. What is the weak instrument problem and its consequences?

j. Would the following variables be plausible instruments for Head Start participation? ***Explain why or why not and be sure to address each of the requirements for a valid instrument in your explanation***.

i. The percent of households in a county that are below the federal poverty line.
ii. The child's parents' educational level.
iii. The number of Head Start centers in the county.
iv. The per capita income of the child's household.

B3. Maximum Likelihood

Consider the following classic linear model:

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

where y and x are continuous and fully observed.

Therefore, the error term has the following normal pdf.

$$f(\varepsilon_i | x_i, \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(^{-\varepsilon_i^2}/_{2\sigma^2})$$

a. What is the likelihood of observing your data (i.e., what is the likelihood function for your sample)?
b. Derive the log likelihood function.
c. Derive the score function for estimating the parameters β and σ^2.
d. Derive the Maximum Likelihood Estimates for β and σ^2.
e. Derive the asymptotic variances of β and σ^2 using the information matrix method.

PART C

(Answer any TWO of the following three questions)

C1. Consider a study where a sample of 400 households was selected to find out their preference to buy a micro health insurance package. They were shown a randomly drawn price from a set (Rs. 10, 40, 150, 300, 700, 1500) and were asked to record a response of Yes or No to the offered price. The package they had been offered was to include doctor's visit and lab tests. In addition, their socio-economic demographic information was collected: age, income, comorbidity, and education. The purpose of this study was to calculate their willingness to pay value.

a. In this study, a condition called monotonicity is expected to hold. What is it and how would you demonstrate it graphically?
b. A linear WTP function is assumed to be: WTP* = x'b + u, where WTP* is the willingness to pay, which is when compared against the offered price results in an estimable logit model. Show all the steps to derive the loglikelihood function.

   After the derivation of the log-likelihood (either logit or probit) that is usually used in STATA, explain as to how you can extract and or estimate the parameters for the WTP equation.

   What is the mean WTP expression? What is the median formula for the WTP? Explain how you would go about getting the CI for the Mean-WTP (delta or simulation methods).
c. Now, consider the log-linear form for the WTP function: WTP = exp(x'b + u). Derive the log-likelihood function. What are the formulas for the mean and the median WTP? Under what scenario will the median WTP be preferred over the mean value?

C2. Consider the groups of these models:

i. Binomial Logit, Multinomial Logit, Ordered Logit, and Nested Logit.
ii. Selectivity, Censored, and Truncated.
iii. Poisson, Negative Binomial, Hurdle Poisson, and Zero-inflated Poisson.

a. Describe the data generating processes for these models and explain with examples when you will use these models. Briefly contrast and compare models within each group.
b. Pick one model of your choice from each group and write out the loglikelihood function.

C3. This part has two options:

Option 1: Box-cox model (non-linear)

We can formulate a more flexible model in the following way to incorporate both types of transformations –log or linear:

$$(y(t)^{\lambda} - 1) / \lambda = a0 + a1 * x(t) + u(t) \quad \text{where } u(t) \sim N(0, \sigma^2)$$

where the value of λ determines the form of transformation. If $\lambda = 1$ => linear. If $\lambda = 0$ => log. The question arises if there in any other possible transformation between two extreme (linear versus log-linear). This is known as the box-cox transformation or BC model. The only way we can find out is to estimate the lambda parameter and test if it is 0 or 1 using the t-test. This model can be estimated in two ways – non-linear least squares or maximum likelihood. The mle version follows.

Qa: Derive the log-likelihood function. Again, pay attention to the change of variable issue while going from u to y in deriving the likelihood function. Show the complete work.

Qb: Write out the formula or expression for the marginal effect: $\frac{\partial y}{\partial x} =$? Discuss as to how you would go about getting the confidence interval (delta method, e.g.). Show the work (formula etc).

Or

Option 2: multiplicative heteroscedasticity (heteroscedastic model estimation using Mle)

Consider the following model:

$$y(t) = a0 + a1 * x(t) + u(t) \quad \text{where } u(t) \sim N(0, \sigma(t)^2)$$

where the variance of $u(t)$ is non-constant. That is, we will make it a function of some variables:

$$V(u(t)) = \sigma(t)^2 = \exp(\Upsilon 0 + \Upsilon 1 * \mathrm{z(t)})$$

Qa: Derive the log-likelihood function for this heteroscedastic model.
Qb: How can you tell (test?) if the data has heteroscedasticity problem or not?