**PhD/MA Econometrics Examination**
**Part A**
**August, 2023**
**Answer any TWO of the following three questions**

**Q1. Statistics**

a. Let the pdf of X be $f(x) = 2xe^{-x^2}, 0 < x < \infty$, and 0 otherwise. Find the pdf of $Y = X^2$.

b. Suppose a sample of $Y_1, Y_2, \ldots, Y_{100}$ is drawn from a gamma distribution with E($Y$)=1.5 and Var($Y$)=0.75. What's the 95% confidence interval of $\bar{Y}$?

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

d. 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 $y$.

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

**Q2. OLS estimation**

For a dependent variable vector *f* with *n* observations, its corresponding independent variable matrix is S with $\lambda$ variables, parameter vector is $\theta$ and residual vector is μ.

a. Derive the OLS solution of parameter vector estimate $\theta$.

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

## Q3. Regression application

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

| *Regression Statistics* | |
|---|---|
| R Square | 0.934 |
| Adjusted R Square | 0.929 |
| Standard Error | … |
| Observations | … |

ANOVA

| | | *Sum of Squared* | *F-stat* | *Significance F* |
|---|---|---|---|---|
| Regression | SSE | 157 | 191.5 | 0.00 |
| Residual | SSR | 11 | | |
| Total | SST | 168 | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|
| Intercept | 1.70 | … | … | … | … |
| $X_1$ | 2.47 | … | … | … | … |
| $X_2$ | -1.32 | … | … | … | … |

$$X'X = \begin{matrix} 30 & 14 & 22 \\ 14 & 50 & 51 \\ 22 & 51 & 107 \end{matrix}$$

$$(X'X)^{-1} = \begin{matrix} 0.0398 & -0.0051 & -0.0057 \\ -0.0051 & 0.0397 & -0.0180 \\ -0.0057 & -0.0180 & 0.0191 \end{matrix}$$

$$X'Y = \begin{matrix} 83 \\ 135 \\ 87 \end{matrix}$$

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 the unbiased sample variance estimate of $\sigma^2$ ? 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.

**PhD/MA Econometrics Examination**

**PART B**

**August, 2023**

**Answer any TWO of the following three questions**

4. Consider the following model $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where $X_1$ is a matrix of $k_1$ variables and $X_2$ is a matrix of $k_2$ variables such that

$$X_1 = \begin{bmatrix} x_{11}^1 & x_{11}^2 & \cdots & x_{11}^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n}^1 & x_{1n}^2 & \cdots & x_{1n}^{k_1} \end{bmatrix}, \ X_2 = \begin{bmatrix} x_{21}^1 & x_{21}^2 & \cdots & x_{21}^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n}^1 & x_{2n}^2 & \cdots & x_{2n}^{k_1} \end{bmatrix}.$$

Denote $b_1$ and $b_2$ as the Ordinary Least Squares (OLS) estimates for $\beta_1$ and $\beta_2$, respectively.

a. Derive the expression for the OLS estimator $b_1$ as a function of $Y, X_1, X_2$, and $b_2$ using the partitioned regression model.

b. Suppose you only observe $X_1$ but not $X_2$. Thus, you regress the OLS model, $Y = X_1\beta_1 + \varepsilon$.

   i. Derive the expression for the OLS estimate $b_1$ that you would estimate under these conditions (i.e., what is the usual OLS estimator for $b_1$ when you regress $Y$ on $X_1$).

   ii. If $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ is the true model, give an expression for the bias of $b_1$ in this circumstance (that you estimated above) as a function of $X_1$, $X_2$, and $b_2$.

c. Now suppose you observe both $X_1$ and $X_2$. Derive the OLS estimator for $b_2$ as a function of $Y$, $X_1$, $X_2$, and $M_1$ using the partitioned regression model, where $M_1$ is the residual maker and $M_1 = I - X_1(X_1'X_1)^{-1}X_1{}'$.

d. Define the Frisch-Waugh-Lovell Theorem and describe its intuition.

e. Under what conditions is the bias you solved for in b.ii. equal to zero. What does this mean in the context of the Frisch-Waugh-Lovell Theorem (i.e., what happens when you regress $X_2$ on $X_1$).

5. You are interested in understanding the impact of early motherhood on school completion in Madagascar. Using data on a sample of 466 Malagasy mothers in their early twenties, all of whom gave birth between the ages of 13 and 23 years of age, you run the following OLS regression:

$$YrsSchool_i = \beta_0 + \beta_1 AgeFirstBirth_i + \varepsilon_i$$

where $YrsSchool_i$ is equivalent to the completed years of schooling of mother $i$ and $AgeFirstBirth_i$ is equal to her age in years when she gave birth to her first child.

You find obtain the following results:

| | |
|---|---|
| Age at First Birth | 0.349618 |
| | (0.067482) |
| Constant | -0.20554 |
| | (1.245402) |

a. What is the interpretation of the coefficient on $AgeFirstBirth$?

b. 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 from the test?

c. Construct a 95% confidence interval for $\beta_1$.

d. Which OLS assumption likely fails in the above regression? Explain why? In other words, give some reasons specific to this scenario as to why you believe this assumption may fail.

e. What is the implication of this failure for your interpretation of your estimate of $\beta_1$?

f. What is the "Fundamental Problem of Causal Inference". Please define it and explain what it means for empirical analysis.

6. Suppose $Y_1,\ldots,Y_N$ has a binomial distribution such that

$$Y = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability } 1\text{-p} \end{cases}$$

This gives $Y_i$ the pdf

$$f(Y_i) = p^{Y_i}(1-p)^{1-Y_i}$$

Note: the number of individuals for which $y_i = 0$ is $N - \sum_i y_i$ .

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 $p$ .

c. Derive the Maximum Likelihood Estimate for $p$ .

d. Derive the asymptotic variance for $\hat{p}_{MLE}$ using the information matrix method.

**PhD/MA Econometrics Examination**

**PART C**

**August, 2023**
**Answer any TWO of the following three questions**

**Q7.** 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: WTP* = x'b + u

   Derive the log-likelihood function so you can estimate the parameter b. After the derivation of the log-likelihood (either for logit or probit), explain as to how you can extract and or estimate the parameters for the WTP equation using 1) canned STATA command (e.g., kappa etc.) b) direct estimation of the parameter b.
c. 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).
d. Now, consider the log-linear form for the WTP function: WTP = exp(x'b + u). Derive the log-likelihood function. (No need to discuss canned versus direct stuff.). What are the formulas for the mean and the median WTP? Under what scenario will the median WTP be preferred over the mean value?

**Q8.** Consider the following linear model

**A.** $y(t) = a0 + a1 * x(t) + u(t)$

Where u(t) is distributed as normal with mean 0 and variance Sig^2 (s2): u ~ N (0,s2)
Derive the log-likelihood function for this case. [Don't forget to apply the change of variable rule to go from u to y (e.g., use of Jacobian) in such likelihood derivations; show the work.]

Q: Derive the log-likelihood function for this linear case. This is a warm-up exercise!

B Now, assume the following log-linear model

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

Q: Derive the log likelihood function for this log-normal case. ~

**C.** The part C has two options:

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

In the above two cases (a and b), we went from a linear form to a log-linear transformation. We can formulate a more flexible model in the following way to incorporate both types of transformations:

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

where the value of $\lambda$ 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.

Q: 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. <u>Bonus: write the STATA script.</u>

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

**Or**

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

Consider the following model:

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

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

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

Q: Derive the log-likelihood function for this heteroscedastic model. <u>Bonus: write the STATA script.</u>

**Q9. Q7.** Consider modeling an ante-natal doctor's visit model as a function of the ***distance to the clinic***, ***annual household income***, ***number of children at home***, and ***education*** level of the mothers.

a. Set up a poisson modelling framework, and spell out the log likelihood function. Show all the steps.
b. Does this function look like your typical demand function? Why or why not?
c. In this case, do we need an exposure variable? Why or why not?
d. What are the expected signs on the independent variables?
e. 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
   Negative Binomial (Type II)

Choose one of the two options and present your rationality along with the derivation of its log likelihood function.