PhD Econometrics Examination
January 2024

Legibility requirement: exam-takers are responsible to write the answers clearly, and marked with appropriate order corresponding to the questions. For answers that are out of order and extremely difficult to read, graders shall deduct half of the points off.

Part A
Answer any TWO of the following three questions

Q1. Statistics

a. Let Y_1, Y_2, …, Y_n be n pairwise uncorrelated random variables, with common mean m and common variance σ^2. Let $\mathrm{W} = \mathrm{a}_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$, where a_i are constant coefficients. Find the expected value E(W) and variance Var(W) for W.

Suppose we need to study whether having a pingpong table reduces pepsi consumption in the department. We obtained a random survey sample on the difference of pepsi consumption from before and after the pingpong table was installed for each surveyed individual. Let d_i stand for the difference for individual i.

b. For this study, what should be the null and alternative hypothesis?

c. Suppose the random sample has 100 observations. The sample mean is -1, sample variance is 121. What conclusion can we get from this hypothesis test? Please derive the analysis and state your rule-of-thumb decision criterion.

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

e. For positive random variables X and Y, suppose the expected value of Y given X is $\mathrm{E}(\mathrm{Y}|\mathrm{X}) = \theta\mathrm{X}$. The constant parameter θ shows how the expected value of Y changes with X. define a random variable Z = Y/X. Find E(Z), using the law of iterated expectations.

Q2. OLS estimation

For a dependent variable vector *y* with *n* observations, its corresponding independent variable matrix is X with k variables, parameter vector is β and residual vector is *e*.

a. Derive the OLS solution of parameter vector estimate for β.

b. Let $\mathrm{P} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ and M=I_n-P. Derive the estimated residual vector $\hat{e}$ using either P or M.

c. Find the covariance, cov(X, $\hat{e}$).

d. Find the expected value, E($\hat{e}\hat{e}'$|X).

e. Suppose for 3 parameters in the model, describe how to jointly test the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$.

Q3. Regression application

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

Regression Statistics	
R Square	0.93
Adjusted R Square	0.92
Standard Error	…
Observations	…

ANOVA

		Sum of Squared	*F-stat*	*Significance F*
Regression	SSE	189	80.8	0.00
Residual	SSR	14		
Total	SST	203		

	Coefficients	*Standard Error*	*t Stat*
Intercept	3.74	…	…
X_1	2.01	…	…
X_2	0.22	…	…

$$X'X = \begin{matrix} 15 & 34 & 32 \\ 34 & 124 & 82 \\ 32 & \dots & 86 \end{matrix}$$

$$(X'X)^{-1} = \begin{matrix} 0.3878 & -0.0314 & -0.1156 \\ -0.0314 & 0.0241 & -0.0111 \\ -0.1156 & \dots & 0.0654 \end{matrix}$$

$$X'Y = \begin{matrix} 132 \\ 402 \\ 311 \end{matrix}$$

a. What’s the number of observations and the sample mean of y, x1 and x2? Fill the missing elements in the (X’X) and $(X’X)^{-1}$ matrix.

b. From the sample size, what problem do you think exist in this regression, and what’s the consequence of this problem?

c. What’s the correlation between x1 and x2? Is there a potential problem for the regression?

d. What’s the unbiased sample residual variance estimate of σ^2 ? Fill the missing Standard Error estimate.

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

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

PART B: Answer any Two

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

Q4. You are conducting an econometric investigation into the hourly wage rates of male and female employees. In particular, you are interested in understanding the determinants of male and female wages and how they might differ across sexes. Your data sample consists of a random sample of 526 paid employees; 252 observations in your sample are female and 274 are male. You estimate the following model:

$$\ln W = \beta_1 + \beta_2 S + \beta_3 A + \beta_4 A^2 + \beta_5 T + \beta_6 (S \times T) + \beta_7 F + \beta_8 (F \times S) +$$

(B.1)

$$\beta_9 (F \times A) + \beta_{10} (F \times A^2) + \beta_{11} (F \times T) + \beta_{12} (F \times S \times T) + u$$

where W = hourly wage rate, measured in dollars per hour
S = number of years of formal education, in years
A = age, in years
T = length of tenure in firm, in years;
F = 1 if employee is female and =0 if employee is male

Assume all the classical assumptions hold. Estimating (1) via Ordinary Least Squares (OLS) yields the following results:

	Coefficient	Standard Error ($\sigma/\sqrt{n}$)
Schooling	0.5937	0.01104
Age	0.0798	0.01216
Age Squared	-0.00093	0.00015
Tenure	-0.01057	0.01128
Schooling x Tenure	0.00227	0.00087
Female	0.03593	0.3373
Female x Schooling	0.01684	0.01684
Female x Age	-0.03847	0.01715
Female x Age Squared	0.000422	0.000219
Female x Tenure	0.0185	0.02652
Female x Schooling x Tenure	-0.002107	0.002187
Constant	-0.5667	0.2385

The corresponding sum of squared error (SSE) and the total sum of squares (SST) of the n=526 observations are:

$SSE = \sum_{i=1}^{n} \hat{u}_i^2 = 80.57$ and $SST = \sum_{i=1}^{n}\left(\ln W_i - \overline{\ln W}\right)^2 = 148.33$

A. Write out the expression (or formula) for the marginal effect of years of schooling on log wages in terms of unknown parameters and variables.

B. Interpret the values of each of the estimated coefficients $\hat{\beta}_2$ and $\hat{\beta}_8$, reported in the Table above.

C. Interpret the values of each of the estimated coefficients $\hat{\beta}_3$ and $\hat{\beta}_9$, reported in the Table above.

D. Based on your answers in parts b and c (and ignoring the age squared term for now), what do you learn from these results about the differential impact of schooling and age on wages across males and females at entry-level employment (i.e., for those for whom T=0)?

E. Construct a 95% confidence interval for $\hat{\beta}_2$. The critical value is 1.96. (*Please note that the standard errors in the Table are already normalized by the square root of the sample size. In other words, the reported standard errors for each coefficient are equal to* $^{\sigma}/_{\sqrt{N}}$.) Is this coefficient statistically different from zero at this confidence level? What was your decision rule (i.e., what was the rule you used to either reject of fail to reject the null hypothesis that the coefficient is zero)?

F. Please do the following:

i. Write the expression (or formula) for the marginal effect of Age on the natural log of wages for male employees in terms of unknow parameters (i.e., in terms of the slope coefficients prior to estimation), implied by equation B.1. Do the same for female employees.

ii. Suppose you want to test that the marginal effect of age on $\ln W$ for males is equal to that for females. Write out the null and alternative hypotheses implied by this test.

iii. Give the equation for the restricted regression implied by the null hypothesis.

iv. Suppose an OLS regression of the restricted model returns a sum of squared error of $SSE = 81.7242$. Using this information, together with the information provided in the problem set up, calculate the F test statistic to test your null hypothesis. (*Hint: Remember the formula forR*^{2} *is* $R^2 = 1 - {}^{SSE}/_{SST}$. *Use this formula in the F-test statistic formula in terms of* R^2 *we gave in class.*)

v. Set up the decision rule to reject or fail to reject this hypothesis at the 5% significance level. The critical value for the F-distribution with the given degrees of freedom and restrictions numbers is approximately 3. What is the conclusion of this test?

G. Now, instead, you are interested in testing the hypothesis that both the marginal effect of both schooling on $\ln W$ and tenure on $\ln W$ are equal for males and females.

i. Provide the null hypothesis and alternative hypothesis that tests this restriction.

ii. Provide an expression that imposes these restrictions on the original model.

iii. Suppose an OLS estimation of the restricted regression returns an $SSE = 80.8747$. Using this information, together with the information provided in the problem set up, calculate the F-test statistic to test this null hypothesis.

iv. Set up the decision rule to reject or fail to reject this hypothesis at the 5% significance level. The critical value for the F-distribution with the given degrees of freedom for this hypothesis is approximately 2.6. What is the conclusion of this test?

H. Suppose now you want to test that schooling, tenure, and age do not differently impact the wages of males versus females. In other words, you want to test that males and females have identical wage equations.

i. Provide the null hypothesis and alternative hypothesis that tests this restriction.

ii. Write out the restricted model that imposes the null hypothesis on the original model.

iii. Suppose an OLS regression of the restricted model returns a sum of squared error of $SSE = 93.1805$. Using this information, together with the information provided in the problem set up, calculate the F test statistic to test your null hypothesis.

iv. Set up the decision rule to reject or fail to reject this hypothesis at the 5% level. The critical value for the F-distribution with the given degrees of freedom and restriction numbers is approximately 2.09. What is the conclusion of your test?

Q5. Let $x_1, x_2, \dots, x_n$ be iid with the pdf, $f(x|\theta) = \frac{1}{\theta} e^{\frac{-x}{\theta}}, x \geq 0, \theta > 0$.

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 θ.

c. Derive the Maximum Likelihood Estimate for θ.

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

Q6. A sample of data consists of n observations on two variables, y and x. The true model is

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad (1)$$

where β_1 and β_2 are parameters to be estimated and ε is a disturbance term that satisfies the usual regression model assumptions.

Suppose you estimate (2) via OLS resulting in the following fitted relationship

$$y_i = b_1 + b_2 x_i + e_i \qquad (2)$$

a. Show that the least squares normal equations imply $\sum_i e_i = 0$ and $\sum_i x_i e_i$

b. Show that the solution for the constant term is $b_1 = \bar{y} - b_2 \bar{x}$, where $\bar{y}$ and $\bar{x}$ are sample means of y and x.

c. Show that the solution for b_2 is $b_2 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$.

In view of the true model specified in (1), we can state

$$\bar{y} = b_1 + b_2 \bar{x} + \bar{e} \qquad (3)$$

where $\bar{e}$ is the sample mean of e. Subtracting (3) from (2), we obtain

$$y^* = b_2 x_i^* + e^* \qquad (4)$$

where $y_i^* = y_i - \bar{y}$, $x_i^* = x_i - \bar{x}$, and $e_i^* = e_i - \bar{e}$. Note, by construction, the sample means of y^*, x^*, and e^* are all equal to zero.

Suppose a second researcher estimates the following model:

$$y_i^* = b_1^* + b_2^* x_i^* + e_i^* \qquad (5)$$

d. Comparing the regressions in (2) and (5), and making use of the OLS estimators (based on the normal equations) of the intercept and the slope coefficient, demonstrate that $b_2^* = b_2$ and $b_1^* = 0$. Explain the intuition behind this.

PART C: Answer any Two

Q7. A survey was conducted in the village of Bahunipati, which had been affected by the massive earthquake of 2015. You can find this survey posted on Canvas. The primary focus of the survey was to test the hypothesis that a certain type of social capital had an impact on self-reported health and well-being. Social capital is typically measured in terms of the number of friends and family in one's network or participation in organizations. What sets this survey apart is that it doesn't consider just the quantity of "friends" as valuable social capital, particularly during times of crisis. Instead, it emphasizes the quality of this social capital. In this context, the survey explored not only the number and scale of social networks but also whether there had been reciprocity established before the crisis. The self-reported well-being index was measured on a scale from 1 (poor health) to 5 (excellent health) and is referred to as the "**HealthIndex**." The reciprocity variable was an index measuring the number of times respondents engaged in reciprocal interactions, including receiving loans, receiving help during health crises, and providing assistance with food, referred to as the "**ReciprocityIndex**." The number of friends served as another measure to reflect the size of the social network (**NumFriends**)

Additionally, the survey collected data on other variables, including age, gender, number of children, and education.

a. Set up the appropriate model starting with the underlying equation and all the variables of interest.

b. Choose the appropriate modeling strategy and derive the corresponding log-likelihood function.

c. Justify the rationale behind the modeling strategy you chose.

d. Explain what sign do you expect for each variable and why?

Q8. A survey was conducted in three schools within a district of Nepal with the objective of assessing awareness about the HPV vaccine among adolescent female students and their mothers. Specifically, they were asked about their willingness to pay for the 2-course vaccine shots, with one of five different price levels randomly offered to them before seeking their response of "Yes" or "No." This survey was conducted separately for both the mothers and the daughters. Our aim was to determine the willingness to pay estimates for both samples, each considered independently. In addition to the Yes/No questions and the price levels, we collected data on their age, education, understanding of the associated risks, and income.

a. Set up the RUM model for the mother’s response, spelling out all the steps clearly.
b. Derive the log-likelihood function using the logit distribution.

c. What is the WTP formula or expression and how would you calculate it from your logit estimates?
d. Describe how you would go about calculating the confidence level of this WTP estimate.

Q9. (You have two option A or B)

A. In the HPV survey, let's assume you wish to conduct choice experiment by providing them with two alternatives/options: do nothing (status-quo), receive shots. In addition to three levels of prices (100, 300, 1000), we have additional attributes like the number of shots required (1 shot, 2 shots). Income and education variables were also collected.

a. Lay out the indirect utility functions clearly with the entire equation and variables etc.
b. Spell out the log-likelihood function.
c. Present examples of two choice sets.
d. What is the marginal willingness to pay (MWTP) expression for the number of shots attributes?
e. Explain how you would calculate the confidence interval for this MWTP estimate.

OR

B. Now turning to 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, NoOfChildren, Insurance

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
Negative Binomial (Type II)
Hurdle Model

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