PhD/MA Econometrics Examination
Part A
Tuesday, January 11th, 2022

Total Time: 3 hours

Answer any TWO of the following three questions

Q1. Statistics

a. Let the pdf of random variable x be $f(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, \quad x < 0 \end{cases}$. Define a new random variable $y = x^2$. Find the pdf of y.

Suppose $X_1, \ldots, X_n$ form a random sample of iid normal distribution with mean 0 and variance σ^2.

b. Find the expected value and variance of X_i^2 .

c. Determine the asymptotic distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Consider two random variables X_1 and X_2, whose joint density function is

$$f(x_1, x_2) = \begin{cases} 21x_1^2 x_2^3 & , 0 < x_1 < x_2 < 1 \\ 0, & \text{elsewhere} \end{cases}$$

d. Find the conditional mean of X_1, given $X_2=x_2$ and $0 < x_2 <1$.

e. Find the conditional variance of X_1, given $X_2=x_2$ and $0 < x_2 <1$.

f. Find the distribution of the new variable $Y = \mathrm{E}(X_1 \mid X_2)$.

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 b.

b. Explain the Frisch-Waugh theorem and its implications.

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

d. If a restriction $R\beta = q$, is imposed on the regression coefficients, derive the formula of restricted OLS regression parameters.

e. Suppose a variable is left out from the regression equation, what type of problem does it cause? Illustrate using the example of $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, when X_2 is left out and $\beta_2 \neq 0$.

Q3. Regression application

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

Regression Statistics	
R Square	…
Adjusted R Square	0.86
Standard Error	…
Observations	36

ANOVA

	df	*Sum of Squared*	*Mean of Sum Squared*	*F-stat*	*Significance F*
Regression	2	3614.02	1807.01	…	0.00
Residual	33	557.17	16.88		
Total	35	…			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	99.15	0.68	144.78	0.00	97.76	100.54
X_1	8.53	0.70	…	0.00	7.10	9.96
X_2	3.78	0.35	10.74	0.00	3.06	…

SSR = 3614.02, SSE = 557.17.

$X'X =$

36	0	0
0	36	-17
0	-17	144

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

0.0278	0.0000	0.0000
0.0000	0.0293	0.0034
0.0000	0.0034	0.0073

$X'Y =$

3569
245
404

a. Explain what is the R^2 and calculate the missing R^2 and SST estimates in the output table.

b. Calculate the unbiased sample variance estimate of σ^2 and fill the missing Standard Error estimate in the table. Calculate the covariance matrix var(b) of the estimated parameter vector.

c. Evaluate the significance level of these parameters. Fill the missing t-stat and the missing upper 95% CI for X_2.

d. Test the hypothesis that parameter $b_1=b_2$. (b_1 and b_2 are parameters for X_1 and X_2)

e. Test the hypothesis that parameter b_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.

f. Fill the missing F-stat in the table and explain its implications.

PhD/MA Econometrics Examination
Part B
Wednesday, January 12th, 2022

Total Time: 3 hours

PART B
(Answer TWO of the following three questions)

Q1: OLS Statistics, Hypothesis Testing, and Interpretation

You are examining the results from an experiment intending to improve the productivity of sales for people in a firm. There are two programs being tested, sales training *(training)* and a higher sales commission rate *(commission)*.

For sales training, workers are randomly assigned into different number of hours each week (*training* is a variable for number of hours). For the new higher commission rate, workers are randomly placed into two groups, those who receive a higher commission and those who continue to receive the old commission rate (*commission* = 1 for people receiving the higher rate and zero otherwise). The variable *sales* reflect annual sales revenue of this individual and *experience* reflects years worked at the firm.

There are 204 sales peoples participating in this experiment. You find a sum of squared residuals of 400.

Table 1: OLS Regression Results

	Coefficients	
	Model 1	**Model 3**
Intercept	44.929149	44.929149
	(0.554151)	(0.554151)
Training	0.005000	0.010000
	(0.00100)	(0.001000)
Commission	0.020000	0.010000
	(0.00300)	(0.003000)
Experience	-0.001000	0.008000
	(0.002000)	(0.002000)
Experience X Commission		0.002000
		(0.000020)
Experience X Training		-0.002000
		(0.000020)
Degrees of Freedom	200	200
R-squared	0.1800	0.1900
Adjusted R-Squared	0.1795	0.1805

Standard Errors in Parentheses

Model 1:

$$\log(sales) = \beta_0 + \beta_1 training + \beta_2 commission + \beta_3 experience + \varepsilon$$

You run the OLS regression and get the following results reported in the first column of **Table 1**.

a. What does the R-squared tell you about the regression?
b. Calculate the 95% confidence interval on $\hat{\beta}_1$. The critical value is 1.96
c. Interpret the coefficient estimate on training (*take for granted that the unbiasedness assumptions hold)*.
d. Interpret the coefficient estimate $\hat{\beta}_2$ from the model estimated (*take for granted that the unbiasedness assumptions hold)*.
e. Suppose you wanted to test the hypothesis that neither intervention had any effect on sales, clearly state the null hypothesis you would test.

f. Now you estimate the following model (i.e., the same as Model 1 except that the two intervention variables are excluded from the regression)

Model 2:

$$\log(sales) = \beta_0 + \beta_1 experience + \varepsilon$$

and find a sum of squatted residuals of 420. Test the hypothesis that neither intervention affects sales (the critical value at the 1% significance level for this problem is 4.506). Clearly walk through the steps to test this hypothesis, and state the conclusion of the test *(Hint: write the F-stat formula in terms of sum of squared error rather than R-squared)*.

g. Suppose the F-statistic you calculated in part (f) has a p-value of 0.015. What does this p-value tell you about the null hypothesis?

Now, a co-worker suggests that the following estimated model better explains sales.

Model 3

$$\log(sales) = \beta_0 + \beta_1 training + \beta_2 commission + \beta_3 experience + \beta_4(experience \times commission) + \beta_5(experience \times training) + \varepsilon$$

You run the OLS regression of Model 3 and get the results reported in Column 2 of **Table 1.**

h. Derive and interpret the marginal effect of an additional hour of training. (Take for granted the OLS assumptions hold, and that our coefficients' estimates are significant).
i. Explain how experience and job training interact in this sales department.
j. Derive and interpret the marginal effect of how receiving the higher commission rate affects sales.
k. Do more experienced or less experienced sales reps respond better to the new commission incentive? (Which coefficient and what about it answers this question).
l. Suppose a coworker points out that training likely helps sales a lot at low levels, but the returns to training decrease (and perhaps even turn negative) as reps receive more and more of it. What variable would you add to Model 1 to estimate this effect and test this hypothesis?

Q2: OLS Regression and the Normal Equations

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.

Q3: Maximum Likelihood Estimation

Suppose the joint distribution of the two= random variables x and y have the following pdf: $f(x_i, y_i|\beta, \theta) = \frac{\theta e^{-(\beta+\theta)y_i}(\beta y_i)^{x_i}}{x_i!}$, where $\beta, \theta > 0, y, x \geq 0, and\ i = \{1,2, \dots, n\}$.

a. What is the likelihood of observing your data (i.e., what is the likelihood function for your sample)?
b. Derive the log likelihood.
c. Derive the Score functions for β and θ.
d. Derive the Maximum Likelihood Estimators for β and θ.
e. Derive the asymptotic variances for β and θ.

PhD/MA Econometrics Examination
Part C
Thursday, January 13th, 2022

Total Time: 3 hours

(Part C: Answer Any Two)

Q7. A household survey collected the data on various health-related incidents and socio-demographic variabes: **AnnualHouseHoldIncome, Age, NoChildren, Education, Annual Doctor's Visit Distance, Distance to Clinic in KM, Number of Illness (morbidity), MicroHealthInsurance** etc.

Turning to the number of doctors visit –demand for health care access let's postulate the following relationship:

DocVisit =f(age, income, distance, NoOfChildren, Insurance etc.)

Part 1

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?

Part 2

d. There will be obviously many people with a 0 entry (with no visit recorded over the last 12 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.

Q8. a. Estimation of the above model will require maximization of the log likelihood function. Present the step-by-step numerical optimization method (e.g., Newton Raphson).

b. Present the marginal effect formula/expression for the effect of *distance*.

c. Provide the formula for the CI (in details/steps) for the marginal effect expression in **b.**

Q9. You have a choice to do either the system question (Part 1) or the selection question (Part 2)

Part 1: Given the following 2-equation system,

$$ChildLabor_t = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * ChildAge_t + \alpha_3 * FatherRemittance_t + u_t$$

$$FatherRemittance_t = \beta_0 + \beta_1 * Poverty_t + \beta_2 * FatherAge_t + v_t$$

FatherRemittance variable captures the absence of fathers working overseas sending money back home to the family.

a. Identify endogenous variables and pre-determined variables.
b. How would you estimate the *child labor* equation using a 2-sls method? Briefly discuss the two steps.
c. Likewise, how would you estimate the *Remittance* equation? Be brief and to-the-point.
d. Set up this system in a grand matrix notation

$$Y = Z * \beta + U \quad (1)$$

e. Derive the Var-Cov(U). Hint: You may use the generic notation for a typical two-equation simultaneous model as we had done in the class.
f. Discuss the iterative 3-SLS estimation method, complete with the necessary steps and derivations.
g. Why is 2sls called a limited information method, whereas the 3sls is considered a full information method?
h. Under what condition would you be able to estimate the two-equation system above as a SUR system? Suggest changes in equations and/or variables above to justify your answers.

OR *Part 2*

Outcome equstion:

$$ChildLabor_t = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * ChildAge_t + \alpha_3 * FatherRemittance_t + u_t$$

Assume the selection here is the school attendance by the children. That is, you observe child labor only if the children are not in school.

a. Write out a reasonable selection equation by choosing the appropriate independent variables keeping in mind the identification issue.
b. Derive in details (Inverse Mills Ratio etc.) 1. Two-step Heckman procedure **OR** 2. The full information maximum likelihood method.