PhD/MA Econometrics Examination
August, 2022

PART A
Answer any TWO of the following three questions

Time: 3 hours

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 $y = x^{\frac{1}{2}}$. Find the pdf of y.

Suppose *X*$_1$, ..., *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

$$f(x_1, x_2) = \begin{cases} 12x_1x_2^3 & ,0 < x_1 < x_2 < 1 \\ 0, & \text{elsewhere} \end{cases}$$

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* = E(*X*$_1$ | *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 multicollinearity issue, its implication and what to do with it.

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 an irrelevant variable is included to the regression equation, would there be any issue? Illustrate using the example of $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, when X_2 is the irrelevant variable and $\beta_2=0$.

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	...

Regression Statistics	
R Square	...
Adjusted R Square	0.26
Standard Error	...
Observations	50

ANOVA

	df	*Sum of Squared*	*Mean of Sum Squared*	*F*	*Significance F*
Regression	2	963.08	481.54	...	0.00
Residual	47	2387.10	50.79		
Total	49	...			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.11	3.71	1.11	0.27	-3.35	11.57
x1	0.86	0.70	...	0.22	-0.54	2.26
x2	0.65	0.15	4.35	0.00	0.35	...

SSR = 963.08, SSE = 2387.1.

X’X =

50	222	361
222	1099	1436
361	1436	5126

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

0.2711	-0.0468	-0.0060
-0.0468	0.0095	0.0006
-0.0060	0.0006	0.0004

X’Y =

631
2787
6052

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
August, 2022

PART B
Answer any TWO of the following three questions

Time: 3 hours

Q4. Consider the model $y = Xb + e$, where X is a $n \times k$ matrix. Let the data matrix X be partitioned into two matrices, $X = [X_1 \colon X_2]$, where X_1 and X_2 have the dimensions $n \times k_1$ and $n \times k_2$, respectively, and $k_1 + k_2 = k$. Thus, we can rewrite the model as

$$y = X_1 b_1 + X_2 b_2 + e. \qquad \text{(B.1)}$$

a. Perform an OLS regression of X_1 on X_2. Derive the matrix of residuals from this regression and denote it e_{12}. *(Hint: use the residual matrix for X_2).*

b. Perform and OLS regression of y on X_2. Derive the matrix of residuals from this regression and denote it e_{y2}. *(Hint: use the residual matrix for X_2).*

c. Perform and OLS regression of e_{y2} on e_{12}. Derive the OLS coefficient from this regression and denote it $\tilde{b}_1$. (*You may use the normal equations to do this*).

d. Show that $\tilde{b}_1 = \hat{b}_1$, where $\hat{b}_1$ is the OLS coefficient on X_1 obtained from a regression of y on both X_1 and X_2. (*Hint: use the answer derived in part c and substitute it into the full model for y represented by equation B.1. The residual e in the regression of y on X is orthogonal to both X_1 and X_2.)*

e. Denote the residuals from the regression of e_{y2} on e_{12} as $\tilde{e}$. Show that these residuals, based on the model, $e_{y2} = e_{12}\tilde{b} + \tilde{e}$, are the same as the residuals obtained from the regression of y on both X_1 and X_2. *(Hint: decompose e_{y2} and e_{12} into their original parts. You will also need to use the results from part d).*

f. Suppose that $X_1 X_2 = 0$, meaning that the two sets of variables are orthogonal. Show that, in this case, $\widetilde{b_1} = b_1^*$, where b_1^* is the OLS coefficient on X_1 obtained from a regression of y on X_1 alone.

g. Define the Frisch-Waugh Theorem and describe its intuition.

Q5. Suppose $\tilde{y}$ is an unobserved latent variable that measures an individual's economic productivity (which can be proxied by hourly earnings), such that:

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

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.

$$\text{Note: } \frac{\partial \Phi(z)}{\partial \theta} = \phi(z)\frac{\partial z}{\theta}$$

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 $L_i(\theta)$) and the individual log-likelihood function.

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

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

Q7. In econometric analysis, we are often concerned with estimating the causal effect of some treatment variable, D_i, on an outcome variable of interest, Y_i. However, obtaining a causal estimate can be challenging.

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

b. Define Y_{1i} as the outcome of individual i if she/he is treated and Y_{0i} and the outcome of that same individual if she/he were not treated. If treatment were randomly assigned across individuals in the sample, then the treatment effect of D_i on Y_i is as follows:

$$E_i[Y_{1i} - Y_{01}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 1], \qquad \text{(B.3)}$$

where D_i is equal to one if individual i was treated and equal to zero if she/he was not treated (i.e., was in the control group).

Suppose that treatment *was not* randomly assigned. Using the Potential Outcomes Framework, decompose the expectation in equation (B.3) 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.

c. Suppose you are looking to estimate the returns to education using the model

$$W_i = \alpha + \beta E_i + \gamma X_i + \varepsilon_i \qquad \text{(B.4)}$$

Where W_i and E_i indicate individual *i's* monthly earnings and educational attainment, respectively. X_i is a vector of control variables for individual *i.*

d. Which OLS assumption is likely to fail when estimating this model? Why? What does this mean for your estimate on the returns to education in earnings?

e. You decide to for instrument educational attainment using the instrumental variable, Z. Which two assumptions are necessary for an instrument to be valid? Define these assumptions in words and math.

f. Derive the 2SLS-IV estimator for β.

g. Would the following variables be plausible instruments for educational attainment? Please explain why or why not. Please be specific in relating the instrument to the requirements for a valid instrument.

i. The educational attainment of individual *i's* parents.

ii. An index variable indicating school quality in an individual's community of residence.

iii. Quarter of individual *i's* birth (i.e. born between January-March, April-June, July-September, or October-December) given that there is usually a cut-off for school in entry in which children turning 5 in the last quarter must begin school the following year.

PhD Econometrics Examination
August, 2022

PART C
Answer any TWO of the following three questions

Time: 3 hours

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.

Q8. A survey of 750 household was conducted to understand two issues: school missing days for children under 18 and the life satisfaction status of the mothers (1 (poor), 2 (fair), 3(good), 4(excellent). Many control variables were also collected for both groups whenever applicable – age, education level (0 illiterate – 16 years of education), household income, drinking water source (spring versus tap water), father's presence (home versus abroad for work) etc.

A. First, consider modeling the children's missing schooling days which may contain many zeros. Pick out the appropriate important independent variables from the list as given above and spell out a hypothesis or any other variable that you might find important to set up as a hypothesis. You are also welcome to choose other control variables and/or confounding factors. Choose an appropriate modeling approach either poisson or the negative binomial and present all the econometric steps all the way to setting up of the log-likelihood function.

B. Using an appropriate modeling approach, set up the loglikehood function for the mother's life satisfaction status. Show all the steps, including diagram if needed. Give some thought as to what independent variables you might use on the right hand side to explain the mother's well-being.

Q9. A survey was conducted in a village (n=490), where the head of the households were asked if they were willing to pay for an insurance program to cover their basic health needs (doctor's visit, blood tests, and minor surgery). They were shown one of the randomly chosen monthly premium costs – 50, 100, 300, 700, and 1200 in Rs and the answers of yes (1) or no (0) were recorded. Other variables like annual income in 1000 was also collected.

WTP*_hat = 1.2 -1.2 Cost + .2* Income
se: (0.35) (0.51) covariance between cost coefficient and the income coefficient = -.002

Some average values are as follows: mean (Cost) = 365; mean (Income) = 53.

Q. Calculate the Mean WTP value.

Q. BONUS: find the CI for the WTP estimate