PhD Econometrics Examination

Part A

Wednesday, March 17th, 2021

Total Time: 3 hours

PART A

(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.

Let *X₁, X₂, ..., Xₙ* represent a random sample following $\chi^2(1)$. $\bar{X}$ is sample mean.

b. Find the limiting distribution of $\sqrt{\bar{X}}$. (note: $\chi^2(k)$ distribution has mean k and variance $2k$)

c. Find the limiting distribution of $\sqrt{n}(\sqrt{\bar{X}} - 1)/\bar{X}$.

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

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

d. Find the marginal density function of X_1 and X_2.

e. Find the conditional density function of X_1 given 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. Write out the matrix representation of linear regression, detail the dimensions of each matrix/vector.

b. Derive the OLS solution of parameter vector estimate b.

c. Interpret the goodness of fit measurement R^2 and derive its expression as a function of y, $\bar{y}$ and e.

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

e. If a restriction Rβ = q, is imposed on the regression coefficients, demonstrate what happens to the sum of squared errors e'e.

Q3. Regression application

Consider the following time series regression output. Someone is regressing per capita annual gasoline consumption on a set of explanatory variables including linear trend, income and price of each year.

```
log(gas/pop) ~ log(income) + log(price) + ltrend

Residuals:
     Min       1Q   Median       3Q      Max
-0.06302 -0.01648  0.00579  0.01840  0.04726

Coefficients:
              Estimate         Std. Error     t value        Pr(>|t|)
Intercept   -16.634946         1.002185       -16.599        < 2e-16 ***
log(income)   1.870306         0.114454       16.341         < 2e-16 ***
log(price)   -0.114410         0.022667       -5.048         1.73e-05 ***
ltrend       -0.017939         0.002599       ???            ???
---

Residual standard error: 0.02956 on 32 degrees of freedom
Multiple R-squared:  0.9653,   Adjusted R-squared:  0.962
F-statistic: 296.7 on 3 and 32 DF,  p-value: < 2.2e-16
```

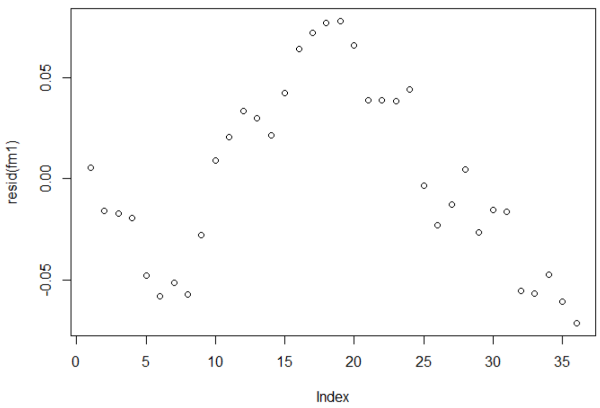
a. Using $t_{0.025} = 2$, calculate the 95% confidence interval for variable log(income). Round to the 2nd decimal place.

b. Calculate the t-value of ltrend coefficient. Is it significant at 1% level?

c. Test the hypothesis that the income elasticity equals to 1.

d. To perform an F test on the hypothesis that the coefficient of log(price) = 0, what is the test statistics value and why?

e. Suppose below is the residual plot of this time series regression, what type of problem is likely to exist in the error terms? What kind of problem does it introduce to the estimates?


resid(fm1)
0.05
0.00
-0.05
0
5
10
15
20
25
30
35
Index

PhD Econometrics Examination

Part B

Thursday, March 18th, 2021

Total Time: 3 hours

PART B

(Answer any TWO of the following three questions)

Q1. Causal Analysis

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:

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

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.1) 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. A large body of evidence indicates that *in utero* and early life health affects adult economic well-being. Suppose that you are interested in estimating the effect of being born at a low birth weight (an indicator of poor *in utero* health and health at birth) on adult economic well-being for a sample of individuals in Ethiopia. *Note: in utero refers to the period during which a child was in his/her mother's womb (i.e., after conception but before birth).* So, you estimate the following model:

$$Y_i = \alpha + \beta LBW_i + \varepsilon_i \qquad \text{(B.2)}$$

where Y_i is the adult earnings of individual i and LBW_i is equal to one if individual i was born at a low birth weight and zero otherwise. Ideally, then, β would capture the effect of being born at a low birth weight on adult earnings.

What assumption is required in order for the OLS estimate of β to represent the unbiased, causal effect of low birth weight on earnings? Why is this assumption likely to fail?

d. You decide to instrument for low-birth weight using the instrumental variable, Z. Which two requirements are necessary for this instrument to be valid? Please define both mathematically and with words.

e. Using Z as your instrument, derive the GMM-IV estimator for β.

f. Derive and describe the estimator for β using the control function approach and using Z as your instrument.

g. Would the following variables be plausible instruments for being born at a low birth weight? Explain why or why not and be sure to address each of the requirements for a valid instrument in your explanation.

   i. Mother's and/or father's birthweight

   ii. Mother's and/or father's highest grade attainment

   iii. The prevalence of infectious disease in the area where individual i was born during his/her *in utero* period.

   iv. Rainfall in during the most recent growing season prior to individual $i's$ birth.

Q2. Maximum Likelihood

Let $\tilde{y}$ be some unobserved latent variable such that

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

You observe y_i and x_i, $i = 1, \dots N$, such that

$$y_i = \begin{cases} 1 \; if \; \tilde{y}_i > 0 \\ 0 \; if \; \tilde{y}_i \leq 0 \end{cases}$$

Define $\phi(\theta)$ as the pdf for a standard normal and $\Phi(\theta)$ as the cdf for the standard normal. Note:

$$\frac{\partial \Phi(z)}{\partial \theta} = \phi(z)\frac{\partial x}{\partial \theta}$$

a. What is θ, the identifiable parameter of interest in this problem?

b. Derive the probabilities that $y_i = 1$ and $y_i = 0$ for individual i.

c. Derive the contribution of each individual in your sample to the overall likelihood function (i.e., derive $\mathcal{L}_i(\theta)$) and the individual log-likelihood function.

d. Derive the Score Function needed to identify $\hat{\theta}_{MLE}$.

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

Q3. Partitioned Regression and Frisch-Waugh-Lovell Theorem

Consider the model $y = Xb + e$, where X is a $n \times k$ matrix. Let the data matrix X be partitioned into two matrices, $X = [X_1 : X_2]$, 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

$$y = X_1 b_1 + X_2 b_2 + e. \qquad \text{(B.3)}$$

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.3. 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.

PhD Econometrics Examination

Part C

Friday, March 19th, 2021

Total Time: 3 hours

PART C

(Answer any TWO of the following three questions)

Q7. Consider the following study/survey scenario. In order to find out people's preference to buy micro health insurance, a choice experiment study was set up. Three attributes were considered with respective levels:

Price:	10, 200, 500 (in NRP, Nepalese Rupees / month).
Major Surgery:	No, Yes
Doctor's Visits (/ month):	1, 2, unlimited
Lab Work:	No, Yes
Immunization:	No, Yes

In addition, other socio economic variables were collected: Age, Gender, Income, Current Insurance (Yes/No); Education Level, and Distance (in minutes) to Nearest Clinic, number of children under 18. The objective was to calculate the marginal willingness to pay value.

a. Set up a RUM structure for this model using the indirect utility functions etc.

b. Present two examples of choice sets.

c. Set up the log-likelihood function for this model.

d. Why is it called a conditional logit model?

e. Present the formula for the marginal willingness to pay for each attribute, and interpret them.

f. What would be the total WTP value?

Q8. In the same survey, new mothers were asked the following health outcome questions: 1) Number of times the women visited the clinic for antenatal care; 2) The BMI of the child at birth; 3) Mode of delivery (At-home by family members, by community midwife; or at Clinic), and 4) Self-rated health status (4= Feeling very well …. 1 = Not feeling well at all). That is, there were four different data generating processes to describe the outcome variables (Antenatal visits, BMI, Mode of delivery, and the self-reported ranked health status.

a. For each health outcome measure, choose an appropriate modeling/estimation method and describe as to why you chose that estimation/modeling method.

b. For each health outcome case, write in steps all regression equations; and the log-likelihood functions.

c. Also, describe the expected sign for each of the independent variables you chose to include in your model.

Q9. Define and describe the difference between the Tobit and the Heckman Selectivity model. Give a read-world example for each of the cases with the corresponding loglikelihood functions. Show your work.