PhD/MA Econometrics Examination

January, 2017

Total Time: 8 hours

MA students are required to answer from A and B.

PhD students are required to answer from A, B, and C.

The answers should be presented in terms of equations, statistical details, and with necessary proofs and statistical deduction. Verbal and brief descriptive discussions will not be sufficient.

PART A
(Answer any TWO from Part A)

Q1. Use the table below. "HWSEI is a constructed variable that assigns a Hauser and Warren Socioeconomic Index (SEI) score to each occupation using the modified version of the 1990 occupational classification scheme available in the OCC1990 variable. The HWSEI variable is a measure of occupational status based upon the earnings and educational attainment associated with each category in the 1990 occupational scheme."

Source	SS	Df	MS			
				Number of obs	=	12104
				F(<u>???</u>,<u>???</u>)	=	498.53
Model	1662649.59	13	127896.1	Prob > F	=	0
Residual	3101628.43	12090	256.5449	R-squared	=	<u>???</u>
				Adj R-squared	=	0.3483
Total	4764278.02	12103	393.6444	Root MSE	=	16.017

Hwsei	**Coef.**	**SE.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Age	1.4378	<u>???</u>	20.55	<u>???</u>	<u>???</u>	<u>???</u>
Age^2	-0.0174	0.0009	-20.44	0.0000	-0.0190	-0.0157
Female	-2.2691	<u>???</u>	<u>???</u>	<u>???</u>	-2.8422	-1.6959
Nursery school to grade 4	-1.4976	2.4374	-0.61	0.5390	-6.2754	3.2801
Grade 5, 6, 7, or 8	-2.3946	1.7404	-1.38	0.1690	-5.8061	1.0169
Grade 9	-1.9035	1.7752	-1.07	0.2840	-5.3831	1.5761
Grade 10	-1.0185	1.7148	-0.59	0.5530	-4.3798	2.3429
Grade 11	0.8919	1.6985	0.53	0.6000	-2.4373	4.2211
Grade 12	8.0669	1.5618	5.17	0.0000	5.0056	11.1283
1 year of college	12.1170	1.5784	7.68	0.0000	9.0230	15.2109
2 years of college	17.0264	1.6285	10.46	0.0000	13.8342	20.2186
4 years of college	26.0401	1.5930	16.35	0.0000	22.9176	29.1626
5+ years of college	35.9438	1.6119	22.3	0.0000	32.7843	39.1033
Constant	-8.8311	2.0393	-4.33	0.0000	-12.8284	-4.8337

Omitted groups: "male" and "no school or N/A"

a) Interpret the coefficient (and other relevant information) on "Grade 11." Grade 11 is a dummy variable for people who have completed through 11th grade, i.e. one year short of finishing high school.

b) Now, looking at the coefficient on the variable Grade 12, is there an important lesson that we can learn?

c) Calculate R^2. Explain your answer. What does the number mean?

d) Find the degrees of freedom for the F.

e) What is the F value? What does it tell you? Can you calculate it from the information given above? If you can calculate F, interpret the value.

f) Calculate the missing values in the row labeled "Age."

g) Calculate the missing values in the row labeled "Female."

h) What is the "Dummy Variable Trap?

*** You should have filled in 10 blanks (**???**) in the table. ****

Q2. Fundamentals of OLS

a. Write out the OLS equation in matrix form. Also, write out the matrices and state their dimensions.
b. State the OLS assumptions in mathematical statements and in sentences (words).
c. Show that the OLS estimator is BLUE and define BLUE. Show all parts: B, L, U, and E.
d. What are the properties (hint: there are six) of the OLS estimator? State them in mathematics and words. Also, state any requirements which are necessary for these properties to hold.
e. Given the properties in part d, what can you infer about the disturbances from the residuals?
f. Write out a simple OLS model. Define your variables and describe how your model might meet or not meet all of the assumptions you stated above.

Q2. Probability Theory, Distributions, and More

a. State Bayes' Theorem
b. In words, describe what Bayes' Theorem means
c. What distribution is below:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

d. Find the first and second moments of this distribution directly from the distribution itself.
e. Find the first and second moments using the moment generating function. Also, discuss if these moments are the same or different from the moments you found in **part d** and why or why not.
f. This distribution has a property called "memoryless." Prove that this distribution is memoryless.
g. Name a practical application of this distribution, i.e., where does it naturally occur?
h. Name the third and fourth central moments – *name them do not calculate them.*
i. Define and draw the PDF and CDF of the distribution.

Part B: Answer any two of the following three questions

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

Q4. Maximum Likelihood

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

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

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. *(Hint: Solve $\sum_i y_i$ in terms of n and $\hat{p}$. Recall that $E[y_i] = \bar{y}$. Then replace $\sum_i y_i$ with what you solved for above in the information matrix. Remember that $E[p] = \hat{p}$).*

Q5. Generalized Least Squares

Consider the model

$$y = x\beta + \varepsilon$$

where $\text{var}(\varepsilon) = \sum \neq \sigma^2 I$

a. Which assumption for OLS fails? What is the implication for the OLS estimator?
b. Show that the OLS estimator for β is still unbiased.
c. Suppose $\text{var}(\varepsilon_i) = \delta x_i$, $i = 1, \ldots, N$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. What does $\sum$ look like?
d. What do the P and Ω matrices look like?
e. By transforming the data, derive the GLS estimator for β. Do this using linear algebra. You do not have to do it with the full matrices written out.
f. Derive $\text{var}(\hat{\beta}_{GLS})$.
g. Show that if $\sum = \sigma^2 I$ then $\hat{\beta}_{GLS}$ reduces to the OLS estimator for β.

Q6. Frish-Waugh-Lovell Theorem and Partitioned Regression

Consider the model $y = X_1 b_1 + X_2 b_2 + e$, such that the data matrix X is broken up into two matrices X_1 and X_2: $X = [X_1, X_2]$.

a. Perform an OLS regression of X_1 onto X_2. Derive the matrix of residuals from this regression and denote it X_{12} *(Hint: use the residual maker 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 y_2 *(Hint: use the residual maker matrix for* X_2).

c. Perform an OLS regression of y_2 on X_{12}. Derive the OLS coefficient from this regression and denote it $\tilde{b}_1$.

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 in the full model for y. The residual e in the regression of y against X is orthogonal to both* X_1 *and* X_2.)

e. Show that the residuals in the regression of y_2 on X_{12} are the same as the residuals obtained from the regression of y on X_1 and X_2.

f. Suppose that $X_1 X_2 = 0$, meaning the two sets of variables are orthogonal. Show that $\tilde{b}_1 = b_1^*$ in this is the case, where b_1^* is the OLS coefficient on X_1 obtained from a regression of y on X_1 alone.

PART C: Answer any Two

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

Q 7.

Background:
In order to encourage people in rural Nepal to use an improved solar-powered filter system, the two Ministries (Health and Renewal Energy) joined forces to undertake an intervention program. They were interested to know whether or not providing scientific information about the water borne diseases would encourage the adoption rate. They put together a 5-minute video explaining the waterborne diseases, and its health implications.

Data summary:

- A. Two hundred small rural communities (say, n=200) were selected for the pilot project. From each community, 8-10 households were selected proportionally (number of total HH, N = nx8 to 12 = 2000).
- B. HH level Socio--demographic questions: annual HH income; age, education, and gender of the HH head, and the HH water source (piped, stream), HH family size.
- C. 5-minute video Intervention: half of the communities (100 communities or 1000 HH) was shown the 5-minute video (Treatment), and the other half (100 communities) were treated as control.
- D. Outcome of interest: **WTP values:** The survey folks at the ministry collected willingness to pay (WTP) data using the randomly selected prices (200, 500, 1000, 1500, & 2500 Rs.) That is, each HH was presented with a randomly selected price, and were asked to express their WTP for the system as Yes or No answer to the offered bod/price.

Analysis:
Q a. Set up the random utility modeling framework for this willingness to pay analysis using the proper notation and all the appropriate variables on the right hand side, including the treatment variable. Present the log-likelihood function. (Show all the steps.)
Q b. Describe all the steps behind the calculation / derivation of the WTP value with and without the intervention: WTP1 verus WTP0. E.g., 1 => treatment and 0 => control.
Q c. Explain a method for calculating the confidence interval for the WTP estimator. Choose one of the following: delta, boot, or simulation.
Q d. How would you go about showing (with the appropriate statistical test) as to whether or not there is a difference in WTP between the treatment and the control samples?

Q 8.
Consider the following 2-equation system that links children's BMI with the mother's smoking habit during the pregnancy.

$$ChildBMI_t = \alpha_0 + \alpha_1 * ChildBirthOrder_t + \alpha_2 * ChildAge_t + \alpha_3 * MotherSmoked_t + u_t$$

$$MotherSmoked_t = \beta_0 + \beta_1 * MotherEduction_t + \beta_2 * Income_t + v_t$$

MotherSmoked = Number of cigarettes smoked during the pregnancy, which was calculated based on the monthly reported smoking habits.

a) Do the two equations fulfill the identification rule? Explain.
b) Explain the 2 SLS estimation method of estimating the *ChildBMI* equation. Explain the method step-by-step.
c) **OPTION 1** Now, present the 3-SLS GLS method of estimating these two simultaneous equation as a full-information method. You can be flexible in setting up the model is a general notation, if you wish. But do not forget to derive the Var-Cov matrix of the two errors, before outlining the 3 SLS estimation method (iterative GLS).

OR

c) **OPTION 2** Estimate the ChildBMI equation as a selectivity model, where the selection equation is the mother's decision to have a baby (e.g., stopping the contraception use) (yes versus no).

Outcome equation:

$$ChildBMI_t = \alpha_0 + \alpha_1 * ChildBirthOrder_t + \alpha_2 * ChildAge_t + \alpha_3 * MotherSmoked_t + u_t$$

Selection equation:

$$StopFamilyPlanningMethod *_t = \beta_0 + \beta_1 * MotherEduction_t + \beta_2 * Income_t + v_t$$

Where *StopFamilyMethod* = 1 (decides to have a baby) or 0 (continues with the family planning methods). Show all the Heckman selection modeling steps including the derivation of the log-likelihood function (derivation of the Inverse Mills ratio etc…)

Q 9. Now, consider the solar-powered filtering problem as described in Q 7, where you have access to the community level aggregated data organized as follows:

ComID	TotalHH	#HHYesToAdopt	VideoShown	SchoolLevel	NoNGOs
1.	10	8	1	10	5
2.	12	5	0	11	3
3.	9	8	1	8	10
4.	11	7	1	12	14
.					
.					
200	10	6	0	10	3

ComIFD = Community ID. Number of NGOs currently in operation in each community is also available (NONGOs), which may reflect the level of social awareness.

Now consider the following behavioral regression framework:

$$y_t^* = x_t' \beta + u_t$$

where *x* is the row vector of all the independent variables (education level, NoNGOs, Treatment). The left hand side dependent variable (*Y*) can be modeled:

a. as an adoption rate variable defined as AdoptionRate = #HHYesToAdopt/TotalHH, which you can model using, e.g., an exponential distribution: $f(y_t) = \lambda * e^{-y_t \lambda}$ Where λ (mean/location parameter) can be modeled with a link function $\lambda_t = e^{x_t' \beta}$
b. Or as a count variable (# of household willing to adopt: 3rd column: #HHYesToAdopt) -- Poisson or Negative Binomial, with a similar log link function as in the exponential function: $\mu_t = e^{x_t' \beta}$.
c. as a log linear or more generally as a box-cox type model
$$(y_t^\lambda - 1)/\lambda = x_t' \beta + u_t$$
where Y = # of HH willing to adopt the new technology. (Here I am assuming that the Poisson approaches normal as the sample size increases.)

Q: Now pick one of the three Data Generating Processes (DGP: a or b or c) and

a) Derive the log-likelihood function. Show all the details and necessary steps.
b) Present the Newton Raphson numerical optimization method for the model of your choosing from the above options (a or b or c).
c) Discuss: the importance of choosing the right value for the step function; Marquadt adjustment when the inversion of the Hessian matrix fails; difference between BHHH and NR; and the appropriate scaling of the variables.