PhD/MA Econometrics Examination

January 2018

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. Distributions: joint, marginal, conditional, and more

Suppose *Y1* and *Y2* denote the proportion of time that employees *Bob Smith* and *Samantha Ford* spend working on their assigned tasks during a workshop. The joint probability density function of *Y1* and *Y2* is given by:

$f(y1,y2) = y1+y2 \qquad 0 \le y1 \le 1, \; 0 \le y2 \le 1$

$0 \qquad$ Otherwise

Ms. Ford has a higher productivity rating than Mr. Smith and a measure of the total productivity of the pair of employees is $15Y1 + 20Y2$.

a. Find the marginal density, *f(y1).*
b. Find the conditional density, *f(y2|y1)*
c. Are Y1 and Y2 independent?
d. Derive the conditional expectation formula *E(y2|y1).*
e. Find the expected value of this measure of productivity.
f. Calculate P(y2 < 0.5|y1=0.3)
g. Calculate P(y2<0.5|y1>0.3)
h. Now, $f(y1,y2, y3) = y1+y2 + y3 \qquad 0 \le y1 \le 1, \; 0 \le y2 \le 1, \; 0 \le y3 \le 1,$ derive the pdf for *f(y1,y2).*

Q2. OLS

Use one the following two models to answer questions: a, b, c etc..

$$y(t) = \beta + u(t)$$
$$Or$$
$$y(t) = \beta x(t) + u(t)$$

$[Note: These\ models\ are\ in\ scalar form; i.e., x(t) is\ one\ independent\ variable.$

a. State all the OLS assumptions.
b. Derive the least squares estimator of the parameter β.
c. Show E($\hat{\beta}$) = β
d. Derive the variance of the OLS estimator.
e. What is the residual variance formula/expression?
f. Show that the OLS estimator $\hat{\beta}$ is efficient.

Q3. Consider the attached STATA regression output tables (Model 1 and Model 2) for a sample of 300 children, many of whom are engaged in the labor force (also known as child labor). Father's remittance variable in the model is 1 and 0 binary variable (1 if the father is overseas working, and sending money back home, 0 otherwise.)

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

MODEL 1

```
      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  3,   296) = 2719.98
       Model |  24887.4637     3  8295.82122           Prob > F      =  0.0000
    Residual |   902.78634   296  3.04995385           R-squared     =  0.9650
-------------+------------------------------           Adj R-squared =  0.9646
       Total |    25790.25   299  86.2550167           Root MSE      =

------------------------------------------------------------------------------
ChildLaborHour~k |      Coef.   Std. Err.      t    P>|t|
-----------------+------------------------------------------------------------
 MotherEducation |  -1.402188   .0221106   -63.42   0.000
        ChildAge |   1.994355     .02915    68.42   0.000
FatherRemittance |  -1.302394   .2184927    -5.96   0.000
           _cons |   3.290908   .3770693     8.73   0.000
```

A. Do the signs of the estimated coefficients make sense? Interpret the slope coefficients.
B. Calculate the 95% CI for the Remittance coefficient, α_3. Interpret this result.
C. Calculate the root mean square error (SEE, or standard error of the regression), $\hat{\sigma}$.
D. Perform the joint significance test for the slope coefficients (show all the work: set up null, alternate, F formula, critical value etc.)

MODEL 2

$$Log(ChildLaborHoursPerWk_t + 1) = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * ChildAge_t + \alpha_3 * FatherRemittance_t + u_t$$

Model 2 was estimated by taking the log of the dependent variable. In order to avoid taking the log of 0, a value of 1 was added to the Child Labor variable.

```
      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  3,   296) =  485.52
       Model |  228.191411     3  76.0638037           Prob > F      =  0.0000
    Residual |  46.3727294   296  .156664626           R-squared     =  0.8311
-------------+------------------------------           Adj R-squared =  0.8294
       Total |   274.56414   299  .918274717           Root MSE      =  .39581

------------------------------------------------------------------------------
lnChildLaborHo~k |      Coef.   Std. Err.      t    P>|t|
-----------------+------------------------------------------------------------
 MotherEducation |  -.1378222   .0050112   -27.50   0.000
        ChildAge |   .1855499   .0066066    28.09   0.000
FatherRemittance |  -.1618096   .0495194    -3.27   0.001
           _cons |   1.449095   .0854594    16.96   0.000
```

E. Interpret the remittance coefficient.
F. Can you use the R-Squared and/or the adjusted R-Squared value to compare the goodness of fit between these two models? If so, which model is better?

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

$$\ln W = \beta_1 + \beta_2 S + \beta_3 A + \beta_4 A^2 + \beta_5 T + \beta_6 (S \times T) + \beta_7 F + \beta_8 (F \times S) + \beta_9 (F \times A) + \beta_{10} (F \times A^2) + \beta_{11} (F \times T) + \beta_{12} (F \times S \times T) + u \quad (1)$$

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
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:

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

a. Use the estimation results to compute the estimated marginal effects of each S, A, and T for the 274 males in the sample. Now do the same for the 252 females in the sample.
b. Write the expression (or formula) for the marginal effect of A on $\ln W$ for male employees in terms of the unknown parameters (i.e., in terms of the slop coefficients prior to estimation), implied by the Equation (1). Do the same for female employees. Suppose you want to test that the marginal effect of A on $\ln W$ for males is equal to that for females. Write out the null and alternative hypotheses. Give the equation for the restricted regression implied by the null hypothesis. Suppose an OLS regression of the restricted model returns an SSE=81.7242. Using this information, together with the information provided in the problem formulation, calculate the test statistic, indicating which is the its approximate distribution under the null hypothesis. Establish a decision rule, using a 5% significance level. What is the conclusion on the inference provided by the test?
c. Write out the restriction in terms of the parameters in Equation (1) that the marginal effects of S on $\ln W$ and of T on $\ln W$ are equal for male and female employees. Provide an expression of Equation (1) imposing these two restrictions. Suppose an OLS estimation of the restricted regression returns an SSE=80.8747. Using this information, together with the information provided in the problem formation, calculate the appropriate test statistic, indicating its approximate distribution under the null hypothesis. Establish a decision rule, using a 5% significance level. What is the conclusion on the inference provided by the test?
d. Suppose you now want to test that males and females have identical wage equations. Write out the appropriate restriction in terms of the parameters in Equation (1) that the regression equations are identical for male and female employees. In other words, what is the appropriate restriction indicating that the mean log-wage of female employees with any given values of S, A, and T equals the mean log-wage of male employees with the same values of S, A, and T. Provide an expression for the null hypothesis to test this restriction and for the alternative hypothesis. The restricted OLS regression yields an SSE=93.1805. Using this information, together with the information provided in the problem formation, calculate the appropriate test statistic, indicating its approximate distribution under the null hypothesis. Establish a decision rule, using a 5% significance level. What is the conclusion on the inference provided by the test?

Q5. A sample of data consists of n observations on two variables, y and x. The true model is

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad (2)$$

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

$$y_i = b_1 + b_2 x_i + e_i \qquad (3)$$

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 (2), we can state

$$\bar{y} = b_1 + b_2 \bar{x} + \bar{e} \qquad (4)$$

where $\bar{e}$ is the sample mean of e. Subtracting (4) from (3), we obtain

$$y^* = b_2 x_i^* + e^* \qquad (5)$$

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:

$$y_i^* = b_1^* + b_2^* x_i^* + e_i^* \qquad (6)$$

d. Comparing the regressions in (3) and (6), 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.

e. Comparing regressions in (3) and (6), demonstrate that $y_i^* = y_i - \bar{y}$.

f. Demonstrate that the residuals in (6) are identical to the residuals in (3).

g. Explain why the specification in (6) is incorrect and the second researcher should have estimated a model excluding the constant? If the second researcher had excluded the constant from his estimation, how would that have affected his estimate of β_2? Explain why.

Q6. Let $X_1,\ldots,X_n$ be iid with pdf $f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0, \ \theta > 0$

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.

PART C: Answer any Two

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

Q7. A sample of 600 households was randomly collected to study the viability of a micro health insurance program in a rural clinic. The sampling was performed by using the proportional sampling design representing all the nine wards in the village. The households were presented with options to enroll in one of the four micro health insurance plans:

Basic (clinic visits), General (clinic + plus pharmacy), Comprehensive (clinic visits, pharmacy + minor surgery), and Premium (clinic visits, pharmacy + minor surgery + long-term chronic care). The information was collected on the following individual specific variables: age, income, and clinic distance.

a. In this scenario and with the information provided above (individual specific only), could you argue that an ordered logit may be fine to use? Explain.
b. Now assume that you also have information about the **cost** associated with these choices. Assuming that the age, income, distance and cost have the same marginal impact regardless of your choice, set up the conditional logit model in a RUM framework. Show all the steps. Write out the log-likelihood function.
c. Now, allow cost to have different impact across the choices. Repeat Q7 b (i.e., set up the RUM model and the log-likelihood).
d. Present the step-by-step method involved in Newton-Raphson algorithm to optimize the LnL function. [Remember, this is not a (non-linear) least-squares problem.]
e. Explain the method of obtaining the standard-errors of the coefficients involved in this log-likelihood function.

Q8. Consider the following 2-equation simultaneous **linear** system,

$$MotherAntenatalVisit_t = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * DistanceToClinic_t + \alpha_3 * FatherEducation_t + u_t$$

$$ChildBMI_t = \beta_0 + \beta_1 * MotherBMI_t + \beta_2 * MotherAntenatalVisit_t + \beta_2 * HouseholdPovertyLevel_t + v_t$$

a. Identify endogenous variables and pre-determined variables.
b. How would you estimate the *ChildBMI* equation using a 2-sls method? Briefly discuss the two steps.
c. Likewise, how would you estimate the Antenatal 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.

Q9. Now let's visit the Antenatal Equation in the above question. This is a count variable, which ranges from 0 to 15.

$$MotherAntenatalVisit_t = \alpha_0 + \alpha_1 * MotherEduc_t + \alpha_2 * DistanceToClinic_t + \alpha_3 * FatherEduc_t + u_t$$

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 nine months of pregnancy), leading to 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

Choose one of the three options and fully develop the derivation, including the log likelihood function.