AFRICAN ECONOMIC RESEARCH CONSORTIUM

Collaborative PhD Programme in Economics for Sub-Saharan Africa

COMPREHENSIVE EXAMINATIONS IN CORE AND ELECTIVE FIELDS

FEBRUARY 13 – MARCH 3, 2017

ECONOMETRICS

Time: 08:00 – 11:00 GMT **Date: Friday, March 3, 2017**

INSTRUCTIONS:

Answer a total of FOUR questions: ONE question from Section A; ONE question from Section B. In Section C, you MUST answer ONE question from Questions (5) and (6); and ONE question from Questions (7) and (8).

The sections are weighted as indicated on the paper.

Statistical tables are provided.

Section A: (15%)

Answer only ONE Question from this Section

Question 1

Suppose that a researcher estimated a regression model explaining wages, using wage data on 250 randomly selected male and female workers in Kenya, and obtained the following results:

$$\widehat{W}_i = \underset{(0.23)}{12.65} + \underset{(0.36)}{2.15}\, MALE_i \qquad R^2 = 0.32 \qquad \sigma^2 = 4.2$$

(s.e.)

where wage (W) is measured in USD/week, MALE is a binary variable that is equal to 1 if the worker is male and 0 otherwise; the figures in the parentheses are the standard errors (s.e.); and σ^2 is the variance of the regression.

The gender wage gap is defined as the difference in the mean earnings between men and women.

(a) What is the estimated gender wage gap? **[3 Marks]**

(b) Is the estimated gender wage gap significantly different from zero at 5% level? **[3 Marks]**

(c) Construct a 95% confidence interval for the gender wage gap. **[3 Marks]**

(d) Based on the sample of 250 workers, what is the mean wage of male workers? What is the mean wage of female workers? **[3 Marks]**

(e) Interpret the reported R^2. **[3 Marks]**

Question 2

Consider the following linear regression model:

$$Y_t = \alpha + \beta X_t + u_t \qquad t = 1,2,\ldots,T \qquad (2.1)$$

(a) State the assumptions of the classical linear regression model in order to estimate the unknown parameters in (2.1) by OLS. **[3 Marks]**

(b) Derive the least squares estimators for α and β. **[3 Marks]**

Suppose a researcher used model (2.1) to estimate the relationship between food expenditure (Y) and income (X), using the observations from 150 households in a particular country (both quantities are in USD). The estimated results are as follows (standard errors (s.e.) are shown in parentheses):

$$\hat{Y}_t = 83.42 + 0.1021\, X_t \qquad R^2 = 0.385$$

(s.e.) (43.41) (0.0209)

(c) Test the hypothesis that $\beta = 0$ at 5% significance level. **[3 Marks]**

(d) Interpret the estimated intercept and slope coefficients. **[3 Marks]**

(e) Would heteroscedasticity be a potential concern in this case? Explain. **[3 Marks]**

Section B: (25%)

Answer only ONE Question from this Section

Question 3

Consider the following random utility models:

$$U_a = w'\beta_a + z_a'\gamma_a + \varepsilon_a \qquad (3.1)$$

$$U_b = w'\beta_b + z_b'\gamma_b + \varepsilon_b \qquad (3.2)$$

representing the individual utility (U) derived from home ownership (a), and from rental housing (b), respectively; w is a vector of observable socio-economic characteristics; z_a and z_b are vectors of specific attributes of home ownership and rental housing, respectively; β_a, β_b, γ_a, and γ_b are the unknown parameters of the models; ε_a and ε_b are random errors with zero means and constant variances with the same symmetric distributions.

(a) Since utilities are unobservable, use the random utility framework above to define an observable binary variable y, where an individual chooses either alternative (a) or alternative (b). **[5 Marks]**

(b) What is the probability that an individual will choose alternative (a)? **[5 Marks]**

(c) The linear probability model (LPM) could be used to estimate the binary choice model. Show that the variance of the LPM error term is not constant. Briefly describe any other limitations of the LPM. **[5 Marks]**

(d) Describe two other alternatives to the LPM. **[5 Marks]**

(e) Derive the log-likelihood function for one of the two alternative functional forms you identified in part (d) above. **[5 Marks]**

Question 4

(a) Explain the difference between a pure random walk process and a random walk with drift process. **[2 Marks]**

(b) Suppose you are given the following simple random walk model:

$$y_t = y_{t-1} + \varepsilon_t$$

where ε_t is a white noise process.

Show that the above equation is an integrated series and the variance and autocovariances are functions of time. **[8 Marks]**

(c) Suppose you are given the following AR(1) process

$$y_t = c + \phi y_{t-1} + u_t$$

where ε_t is white noise and both ϕ and c are constants.

Suppose ϕ=1 what specific AR(1) process is this? Show that the first difference of the process is stationary.

[4 Marks]

(d) To what extent is it true that a cointegration test is a unit root test within the multivariate time series framework?

[6 Marks]

(e) Suppose $|\phi|<1$ in part (c) derive the corresponding infinite moving average representation

[5 Marks]

Section C: (60%)

Answer TWO Questions from this Section,

one of which must be Question 5 or 6 AND Question 7 or 8

Question 5

Consider the following static panel data model for oil consumption in Africa:

$$C_{it} = \alpha_i + \beta_1 P_{it} + \beta_2 F_{it} + \beta_3 Q_{it} + u_{it} \qquad (5.1)$$

Where $u_{it} \sim iid\,(0, \sigma^2)$ and uncorrelated with P, F and Q, $(i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T)$; C = oil consumption (in USD); P = oil price (in USD); F = foreign capital inflows (in USD); and Q = real output in USD (2000 = 100).

(a) Describe the limitations of pooled OLS in estimating the intercept and slope parameters of model (5.1)

[5 Marks]

(b) The intercept in model (5.1) varies with i, which suggests two alternative model specifications. State the important assumptions underlying each of the two alternative specifications. Briefly explain how each of the alternative specifications may be estimated.

[5 Marks]

(c) The model in (5.1) was estimated using the Fixed Effects within estimator. Interpret the estimated coefficients of the model. Is the estimation of fixed effects model justified?

[5 Marks]

$$\mathbf{L\_OILCONS_{it} = -9.719 + 0.590\ LP\_OIL_{it} - 0.0002\ LF\_INFLOW_{it} + 0.929\ LQ_{it}}$$
$$(-1.14)\quad (4.69)\qquad\qquad (-0.008)\qquad\qquad (2.37)$$

Total panel obs. = 195; Countries = 13
Hausman Test = 18.70***
White Heteroskedasticity Test = 4.87
L_OILCONS = log of oil imports in real values; LP_OIL = log of the price of oil; LF_INFLOW = log of foreign inflow in real values; and LQ = log of real output.

*Note: The values in the parentheses are t-values; *** significant at 1%.*

(d) Suppose the lag of the oil consumption variable is added as one of the regressors in the model. Explain the reasons for including lagged oil consumption. **[5 Marks]**

(e) Explain why the estimators for the static panel data model may not be used when the lag of oil consumption is one of the regressors. **[5 Marks]**

(f) Explain how the Arellano & Bond GMM approach can rectify the problem that may arise in part (e) above. What are the limitations of this approach? **[5 Marks]**

(g) If the time series dimension of the panel is from 1996 to 2010, what panel unit root test would you think is appropriate in this particular sample? Justify your answer.

[5 Marks]

Question 6

Let y_t be a K dimensional multiple time series $y_1, \ldots\ldots, y_T$ where $y_t = (y_{1t}, \ldots\ldots, y_{Kt})$ is generated by a stationary, stable reduced form $VAR(p)$ process:

$$y_t = c + A_1 y_{t-1} + \ldots. + A_p y_{t-p} + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6.1)$$

where $c = (c_1, \ldots\ldots, c_K)'$ is a $K \times 1$ vector of intercept terms. A_i are coefficient matrices. u_t is white noise process with non-singular covariance matrix Ω. $A_1, \ldots., A_p$ are unknown, and are supposed to be estimated.

Following an appropriate column stacking, equation (6.1) can be compactly written as

$$Y = BZ + U \ldots\ldots\ldots\ldots\ldots\ldots(6.2)$$

and the OLS estimator of B is given by

$$\hat{B} = YZ'(Z'Z)^{-1}$$

(a) Show that the OLS estimator of B is consistent. **[5 Marks]**

(b) Is it appropriate to use OLS to estimate the parameters of the K - equations in (6.1)? Why? **[2 Marks]**

(c) Suppose that the K variables in (6.1) are non-stationary; and therefore we are interested in examining whether the variables in the system are cointegrated using the following vector error correction model (VECM)

$$\boldsymbol{\Delta y}_t = \pi \boldsymbol{y}_{t-p} + \sum \pi_i \boldsymbol{\Delta y}_{t-i} + \boldsymbol{u}_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6.3)$$

Show that the VECM in equation (6.3) can be derived from the VAR(p) process in equation (6.1) **[8 Marks]**

(d) Explain how equation (6.3) could be used to examine whether or not the variables are cointegrated. **[5 Marks]**

(e) A vector moving average (VMA) representation of the reduced form VAR(p) process in (6.1) can be derived and then be used to understand the dynamic inter-relationship between the variables in the system. What are the main challenges in carrying out this kind of analysis? How can you address these challenges? **[5 Marks]**

(f) Explain the difference between the sequential and information based approaches to the determination of the optimal number of lags for a VAR model. **[5 Marks]**

Question 7

Consider the standard Tobit model with the given specifications and assumptions:

$$y_i^* = x_i' \beta + \varepsilon_i \qquad i = 1, 2, \ldots, N \qquad (7.1)$$
$$y_i = y_i^* \qquad \text{if } y_i^* > 0$$
$$y_i = 0 \qquad \text{if } y_i^* \leq 0$$

where $\varepsilon_i \sim NID(0, \sigma^2)$ and independent of x_i. This model is also called a censored regression model since some observations are censored (in this case from below).

(a) Using the specification (7.1), find the probability that $y_i = 0$ given x_i, that is $P\{y_i = 0 |x_i\}$, and the conditional expectation of y_i given a positive outcome, $E\{y_i|y_i > 0\}$. **[8 Marks]**

(b) Explain why zero values should be included in the estimation of the model. **[4 Marks]**

(c) The Tobit model may also be written in the general form

$$y_i^* = x_i' \beta + \varepsilon_i \qquad i = 1, 2, \ldots, N \qquad (7.2)$$

$$y_i = y_i^* \qquad \text{if } y_i^* > L$$

$$y_i = L \qquad \text{if } y_i^* \leq L$$

where L = lower bound.

Since the censored variable y is a transformation of y^*, $y = g(y^*)$, its probability density function (pdf) may be obtained based on conditional distribution of the latent variable y^* given x. Then, the model parameters may be consistently and efficiently estimated using maximum likelihood (ML).

(i) What is the pdf of the censored variable y? **[5 Marks]**

(ii) Derive the first-order conditions for ML estimation of the unknown parameters β and σ^2. Show your work/derivation. **[10 Marks]**

(d) What are the main weaknesses of ML estimation of the Tobit model? **[3 Marks]**

Question 8

In the context of multinomial models, individuals may face more than two alternatives to choose from. Depending on the type of observations on the explanatory variables (choice attributes, z_{ij}, or individual characteristics, x_i, where index i refers to ith individual and j stands for a particular choice), the probability of the individual choosing j, $j = 1, 2, \ldots, M$ may be expressed as:

$$P_{ij} = \mathrm{P}\{y_i = j|z_{ij}\} = \frac{\exp\{z_{ij}'\gamma\}}{\sum_{l=1}^{M} \exp\{z_{il}'\gamma\}} \qquad (8.1)$$

for conditional logit, or

$$P_{ij} = \mathrm{P}\{y_i = j|x_i\} = \frac{\exp\{x_i'\beta_j\}}{\sum_{l=1}^{M} \exp\{x_i'\beta_l\}} \qquad (8.2)$$

for multinomial logit.

(a) Explain the difference between the conditional logit and the multinomial logit models. **[4 Marks]**

(b) For the conditional logit model, derive the marginal effects, that is the change in the probability of $y_i = j$, with respect to a given change in z_{ik}. **[6 Marks]**

(c) For the multinomial logit model, derive the marginal effects, that is the change in the probability of $y_i = j$, with respect to a given change in in x_{ik}. **[6 Marks]**

(d) The two models given in (8.1) and (8.2) can be combined into what some authors called a "mixed logit model". For this model, write the probability that $y_i = j$ given x_i and z_{ij}. **[3 Marks]**

(e) In order to use multinomial logit, what restriction is implicitly imposed on the odd-ratios in relation to M different alternatives? If this assumption fails, what alternative models may be estimated? Briefly describe each model. **[8 Marks]**

(f) Given a set of multi-response data, explain when an ordered logit model is appropriate to use. **[3 Marks]**