AFRICAN ECONOMIC RESEARCH CONSORTIUM

Collaborative PhD Programme in Economics for Sub-Saharan Africa

COMPREHENSIVE EXAMINATIONS IN CORE AND ELECTIVE FIELDS

FEBRUARY 19 – MARCH 10, 2021

ECONOMETRICS

Time: 08:00 – 11:00 GMT **Date: Monday, March 1, 2021**

INSTRUCTIONS:

1. 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**.
2. The sections are weighted as indicated on the paper.
3. The null hypothesis and the alternative hypothesis for all the statistical tests in this examination should be indicated.
4. Statistical Tables are provided.

SECTION A: (15%)

Answer only ONE Question from this Section

Question 1: [15 Marks]

Consider the following regression model

$$y = \beta_0 + \beta_1 x + u \qquad (1.1)$$

where *y* is wage in US dollars, *x* is number of years of experience and u is an error term with the property $u \sim N(0, \sigma^2)$.

(a) Suppose the assumption of normality of the error term is violated. Will the OLS estimates for b_0 and b_1 be biased? **[1 Mark]**

(b) What is the importance of the normality assumption for statistical inference? **[4 Marks]**

(c) What is multicollinearity? How does it affect the estimated regression results?
[5 Marks]

(d) Suppose the true regression model is

$$y = a + b_1 x_1 + u$$

and you incorrectly estimate the following regression model

$$y = a + b_1 x_1 + b_2 x_2 + u,\ u \sim N(0, s^2)$$

What will be the consequence(s) on the estimated regression results? **[5 Marks]**

Question 2: [15 Marks]

Consider the following simple linear regression model:

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1, 2, \ldots, N \qquad (2.1)$$

where: Y_i is the ith observation on the dependent variable, Y; X_i is the ith observation on the explanatory variable, X; e_i is the random error term; β_1 and β_2 are the unknown parameters to be estimated.

(a) Derive the least squares estimator of β_2. [**3 Marks**]

(b) Suppose Y is household expenditure on food in USD and X is the household income in USD and the estimated coefficients are $\hat{b}_1$ = 5.23, $\hat{b}_2$ = 0.65 and R^2 = 0.85. Write the estimated regression model and interpret the values of $\hat{b}_1$, $\hat{b}_2$ and R^2. [**4 Marks**]

(c) The number of observations used in the estimation in part (b) is 100, the standard error of $\hat{b}_1$ is 0.2 and the standard error of $\hat{b}_2$ is 0.13. Use this information together with the information in part (b) to test for the significance of $\hat{b}_2$. Use 5% level of significance.
[**2 Marks**]

(d) State the assumptions underlying the classical linear regression model. [**3 Marks**]

(e) Choose one case where an assumption of the classical linear regression model is violated. Explain the effects on the estimated model and what solution(s) may be taken to deal with the problem. [**3 Marks**]

SECTION B: (25%)

Answer only ONE Question from this Section

Question 3: [25 Marks]

Economic theory suggests that similar goods should be close substitutes for each other. The information below shows results from a time series analysis on the average monthly prices of regular oranges and organic oranges (grown without chemical pesticides and fertilizers) in a certain market. These are two closely related products, but many consumers are willing to pay somewhat more for organic oranges, perceiving them to be healthier. As a researcher, analyze the relationship between the two prices given the following information:

Figure 3.1. Plots of Prices of Regular and Organic Oranges

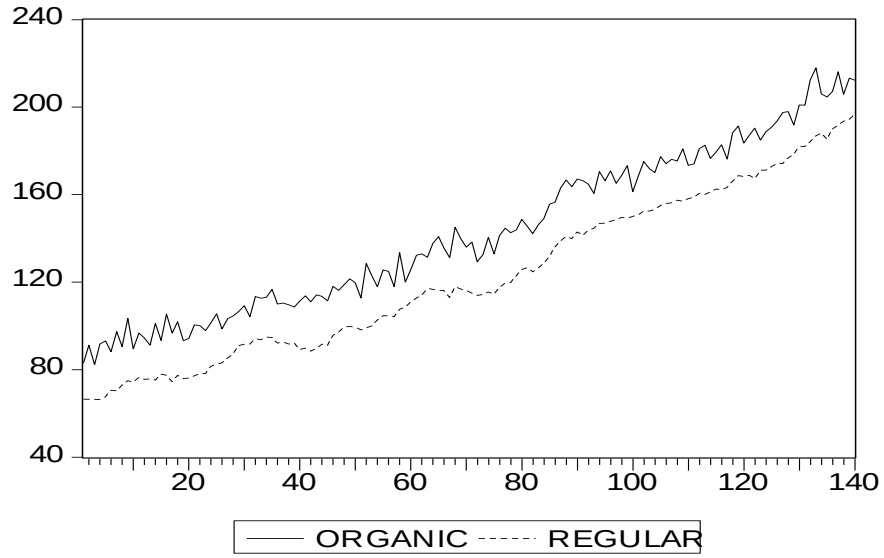

Table 3.1.a: ADF Tests for Prices of Organic Oranges

No. of Lags	t-ADF		t-ADF	
	Intercept	AIC	Trend and Intercept	AIC
Nlag = 0	-0.7283	6.4806	-6.7606	6.2087
Nlag = 1	0.1329	6.1621	-3.9656	6.0624
Nlag = 2	0.4125	5.9980	-2.5337	5.9620
Nlag = 3	0.6228	6.0087	-2.4942	5.9726
Nlag = 4	0.7670	6.0254	-2.5321	5.9865

Notes: ACV = -2.89 at 5% for model with Intercept; and
ACV = -3.45 at 5% for model with Intercept and Trend.

Table 3.1.b: ADF Tests for Prices of Regular Oranges

No. of Lags	t-ADF		t-ADF	
	Intercept	AIC	Trend and Intercept	AIC
Nlag = 0	1.4308	3.8130	-1.5588	3.8047
Nlag = 1	1.5764	3.8175	-1.2661	3.8156
Nlag = 2	1.2281	3.8216	-1.5206	3.8145
Nlag = 3	1.0248	3.8376	-1.6728	3.8270
Nlag = 4	1.0734	3.8590	-1.7035	3.8471

Notes: ACV = -2.89 at 5% for model with Intercept; and
ACV = -3.45 at 5% for model with Intercept and Trend.

ADF Test of Residuals (From Regression of Organic on Regular Oranges)

Null Hypothesis: EHAT has a unit root

Exogenous: None

Lag Length: 0 (Automatic - based on SIC, maxlag=13)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.75907	0.0000
Test critical values:	1% level	-2.581705	
	5% level	-1.943140	
	10% level	-1.615189	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(EHAT)
Method: Least Squares
Sample (adjusted): 2000M02 2011M08
Included observations: 139 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
EHAT(-1)	-1.081559	0.084768	-12.75907	0.0000

R-squared	0.541214	Mean dependent var	0.003435
Adjusted R-squared	0.541214	S.D. dependent var	5.771903
S.E. of regression	3.909526	Akaike info criterion	5.571878
Sum squared resid	2109.246	Schwarz criterion	5.592989
Log likelihood	-386.2455	Hannan-Quinn criter.	5.580457
Durbin-Watson stat	1.959746		

Note: COINTEGRATION ASYMPTOTIC CRITICAL VALUE (ACV) FOR m = 2 (2 variables) is -3.90 at 1%; -3.34 at 5%; -3.04 at 10% (source: Davidson and MacKinnon 1993).

(a) Analyze the behavior of the two series as shown on Figure 3.1. **[3 Marks]**

(b) Differentiate between stationary and non-stationary variables. **[4 Marks]**

(c) Do the prices exhibit stationary behavior? Support your answer. **[6 Marks]**

(d) What is cointegration? **[3 Marks]**

(e) Describe the Engle and Granger residual-based test for cointegration. **[4 Marks]**

(f) Is there a presence of cointegration between the prices of regular and organic oranges? Support your answer. **[5 Marks]**

Question 4: [25 Marks]

Consider the following random utility models:

$$U_a = w'\beta_a + z_a'\gamma_a + \varepsilon_a \quad (4.1)$$

$$U_b = w'\beta_b + z_b'\gamma_b + \varepsilon_b \quad (4.2)$$

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). **[4 Marks]**

(b) What is the probability that an individual will choose home ownership? **[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. **[6 Marks]**

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

SECTION C: (60%)

Answer ONE Question from Questions (5) and (6);

and ONE Question from Questions (7) and (8)

Question 5: [30 Marks]

(a) A stronger condition than convergence in probability is the *mean square convergence.* A random sequence $\{X_t\}$ is said to converge in mean square to a constant b, denoted as

$$X_t \xrightarrow{m.s.} b$$

if for every $\delta > 0$ there exists a value N such that, for all $T \geq N$, $E(X_t - b)^2 < \delta$.

(i) Use the above result to derive the condition(s) needed for *serially dependent variables* to obey the Weak Law of Large Numbers (WLLN). **[7 Marks]**

(ii) Use intuition [proof is not needed] to explain that mean square convergence is a stronger condition than convergence in probability. **[3 Marks]**

(iii) Explain the statistical implication of the condition(s) you have derived in (i) above. **[3 Marks]**

(iv) Can we use the WLLN to establish the consistency of estimators when time series variables are non-stationary? Why or why not? **[3 Marks]**

(b) Given the following stochastic processes

$y_t = c + \theta u_t + u_{t-1}$, u_t is a white noise process; and

$y_t = \alpha + \beta y_{t-1} + u_t$, $|\beta| < 1$, and u_t is a white noise process.

Determine which process is ergodic stationary in the mean. Explain. **[6 Marks]**

(c) Given an AR(1) model

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

Show that if $\phi = 1$

$$\sum_{t=1}^{T} y_{t-1}\varepsilon_t \xrightarrow{d} \tfrac{1}{2}\left[c^2(1) - 1\right]$$ **[8 Marks]**

Question 6: [30 Marks]

(a) Distinguish between a pure random walk process and a random walk with drift process. **[4 Marks]**

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

$$y_t = y_{t-1} + e_t \qquad (6.1)$$

where e_t is a white noise process.

Find the mean, variance, and first and second autocovariances of (6.1). Comment on its stationarity properties. **[10 Marks]**

(c) Suppose you are given the following random walk with drift

$$y_t = b + y_{t-1} + e_t \qquad (6.2)$$

where b is a drift parameter, and e_t is a white noise process.

Show that its first difference is stationary. **[6 Marks]**

(d) Given an AR(1) model without a drift, show how the series can be tested using the Dickey-Fuller unit root test. **[4 Marks]**

(e) In the multivariate setting

$$\mathbf{y_t} = \mathbf{A}_1\mathbf{y_{t-1}} + \boldsymbol{\varepsilon_t}, \qquad (6.3)$$

where $\mathbf{y}_t = (y_{1t},.....y_{Kt})'$ is a (K×1) vector of variables, $\mathbf{A}_1$ is a (K×K) matrix of fixed coefficients, and $\boldsymbol{\varepsilon_t} = (\varepsilon_{1t},....\varepsilon_{Kt})$ is a K-dimensional *white noise* or *innovation process*, that is $E(\boldsymbol{\varepsilon}_t) = 0$, $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \Sigma_\varepsilon$, and $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s') = 0$ for $s \neq t$.

Show how the variables can be tested for cointegration using the Johansen test. **[6 Marks]**

Question 7: [30 Marks]

Let y_i be an observed binary variable (coded 0 or 1) for an individual *i*. You are asked to model the probability of occurrence of the event $y_i = 1$ by the logistic distribution given as follows,

$$\Pr(Y_i = 1) = \Lambda(\alpha) \qquad (7.1)$$

where $\Lambda(\cdot)$ denotes the cumulative density function (cdf) of the logistic distribution and α a real number. The probability density function (pdf) of the logistic distribution will be denoted by $\lambda(\cdot)$ The model has no explanatory variable.

The sample consists of 32 individuals. A total of 21 individuals' response outcomes is $y_i = 0$ and the other 11 individuals had the variable *Y* taking on the value 1.

(a) Show that the log-likelihood function of the logit model above is given as,

$$l(\alpha; y) = 11 \times \ln\left[\Lambda(\alpha)\right] + 21 \times \ln\left[1 - \Lambda(\alpha)\right] \qquad \textbf{[6 Marks]}$$

(b) Let $\hat{\alpha}$ be the maximum likelihood estimator of the parameter α. Establish the first order condition of the log-likelihood maximization. Deduce that $\hat{\alpha} = \ln\left(\frac{11}{21}\right)$. **[10 Marks]**

(c) Show that the second order derivatives known as the Hessian matrix is given by, $H(\alpha) = -32\lambda(\alpha)$. **[4 Marks]**

(d) You are now interested in testing the null hypothesis $H_0 : \alpha = 0$ against the alternative $H_1 : \alpha \neq 0$.

(i) Show that under the null hypothesis H_0, the log-likelihood is $\ell_0 = -32\ln 2$. **[3 Marks]**

(ii) Show that under the alternative, the log-likelihood becomes

$$\ell_1 = 11 \times \ln 11 + 21 \times \ln 21 - 32 \times \ln 32. \qquad \textbf{[3 Marks]}$$

(iii) Deduce the likelihood ratio test statistic $LR = -2\left[\ell_0 - \ell_1\right]$ which follows a chi-square distribution with one as degree of freedom. Conclude the test at 5% significance level given that the related critical value is 3.8415. **[4 Marks]**

Question 8: [30 Marks]

(a) Consider the following two-way panel data model:

$$y_{it} = \gamma + X_{it}'\beta + v_{it}, \qquad i = 1, \ldots, N; t = 1, \ldots, T \tag{8.1}$$

with

$$v_{it} = \alpha_i + \lambda_t + \varepsilon_{it} \tag{8.2}$$

where *i* and *t* denote the cross-section and time-series indices, respectively; γ is a scalar; β is a $K \times 1$ vector of unknown parameters; and X_{it} is a $K \times 1$ vector of explanatory variables.

(i) Why do we call equation (8.2) as a two-way error structure? Give an economic example that may require this error structure modeling. **[2 Marks]**

(ii) Show that the overall disturbance in matrix form becomes,

$$v = Z_\alpha \alpha + Z_\lambda \lambda + \varepsilon$$

where all vectors and matrices are conventionally defined. **[7 Marks]**

(iii) Show that the variance-covariance matrix of the overall error *v* is given by,

$$\Omega = \sigma_\alpha^2 \left(I_N \otimes J_T \right) + \sigma_\lambda^2 \left(J_N \otimes I_T \right) + \sigma_\varepsilon^2 \left(I_N \otimes I_T \right) \qquad \textbf{[7 Marks]}$$

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

$$C_{it} = \alpha_i + \beta_1 P_{it} + \beta_2 F_{it} + \beta_3 Q_{it} + u_{it}, \; i = 1, 2, \ldots, N; \; t = 1, 2, \ldots, T \tag{8.3}$$

where $u_{it} \sim iid\ (0, \sigma^2)$ and uncorrelated with *P*, *F* and *Q*, *C* = oil consumption in USD; *P* = oil price in USD; *F* = foreign capital inflows in USD; and *Q* = real output in USD.

(i) Describe the limitations of pooled OLS in estimating the intercept and slope parameters of model (8.3). **[4 Marks]**

(ii) The intercept in model (8.3) 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]**

(iii) The model in (8.3) was estimated using the Fixed Effects within estimator (see Table 8.1 below). Interpret the estimated coefficients of the model. Is the estimation of fixed effects model justified? **[5 Marks]**

Table 8.1: Panel data regression of oil consumption in Africa

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

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%.