AERC CREA

AFRICAN ECONOMIC RESEARCH CONSORTIUM

Collaborative PhD Programme in Economics for Sub-Saharan Africa

COMPREHENSIVE EXAMINATIONS IN CORE AND ELECTIVE FIELDS

FEBRUARY 14 – MARCH 6, 2018

ECONOMETRICS

Time: 08:00 – 11:00 GMT **Date: Friday, February 23, 2018**

INSTRUCTIONS:

1. Answer a total of FOUR questions: ONE question from Section A, ONE question from Section B, and TWO questions from Section C; One of which MUST be Either Question 5 or Question 6.
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. Some relevant formulas for Section B are presented in the Appendix.

SECTION A: (15%)

Answer only ONE Question from this Section

Question 1

Consider the following simple linear regression model with two variables Y and X:

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

where: Y_i is the i*th* observation on the dependent variable, Y; X_i is the i*th* observation on the explanatory variable, X; e_i is the random error term; and 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 US$ and X is the household income in US$ and the estimated coefficients are $\hat{\beta}_1$=5.23, $\hat{\beta}_2$ = 0.65 and $R^2 = 0.85$. Write the estimated regression model and interpret the values of $\hat{\beta}_1$, $\hat{\beta}_2$ and R^2. **[4 marks]**

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

Question 2

(a) The table below reports regression results for the following model

$$q_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varepsilon_t$$

Note that the standard errors are reported in brackets below the estimated coefficients.

	Dependent variable:
	q
x_2	-0.863
	(0.067)
x_3	0.951
	(0.067)
x_4	-22.380
	(4.270)
Constant	83.162
	(6.774)
No. of Observations	120
R^2	0.676
Adjusted R^2	0.667
Residual Std. Error	20.067 (df = 116)
F Statistic	80.529 (df = 3; 116)

Breusch Pagan test: BP = 1.9826, df = 3, p-value = 0.576

(i) Compute the t-statistics for the three slope coefficients and the intercept. **[3 marks]**

(ii) Comment on the statistical significance of individual parameter estimates at 5% level. **[3 marks]**

(iii) Test the overall significance of the regression using 5% level. **[3 marks]**

(iv) Use the information provided to test for heteroscedasticity. **[3 marks]**

(b) What is multicollinearity? Describe one way to detect the presence of multicollinearity. **[3 marks]**

SECTION B: (25%)

Answer only ONE Question from this Section

Question 3

Suppose a researcher is interested in explaining why people in a certain city buy health insurance or not. A sample of observations was drawn to undertake the study.

Let $y_i^* = x_i'\beta + \varepsilon_i \qquad i = 1, 2, 3, \ldots, N \qquad (3.1)$

$y_i = 1$ if $y_i^* > 0$

$y_i = 0$ if $y_i^* \leq 0$

where y_i^* is the net utility to own health insurance and is not observed

$y_i = 1$ if a person has health insurance; $y_i = 0$, otherwise

x_i = vector of explanatory variables

ε_i = unobservable random errors and assumed to have logistic distribution

(a) Write the complete expression for $f(x_i'\beta)$ and $F(x_i'\beta)$ where $f(.)$ and F(.) are the probability density function (pdf) and cumulative distribution function (cdf) of the logistic distribution, respectively. **[5 marks]**

(b) Derive the probability that a person buys health insurance, and the probability that a person does not buy health insurance. **[5 marks]**

(c) Derive the log-likelihood function for the logit model, and the first-order conditions for maximizing the function. **[6 marks]**

(d) Write the expression for marginal effects for the logit model and provide its interpretation. **[5 marks]**

(e) Why is the logit model preferred to the linear probability model? **[4 marks]**

Question 4

The econometric results reported below are obtained from estimation of the following model for an unspecified Sub-Saharan African country:

$$c_t = \beta_1 + \beta_2 y_t + \upsilon_t$$

where c_t is logarithm of consumption, y_t is logarithm of income, and υ_t is the usual error term.

c_t and y_t are both I(1)

Figure 4.1 Income and Consumption Time Series Plot

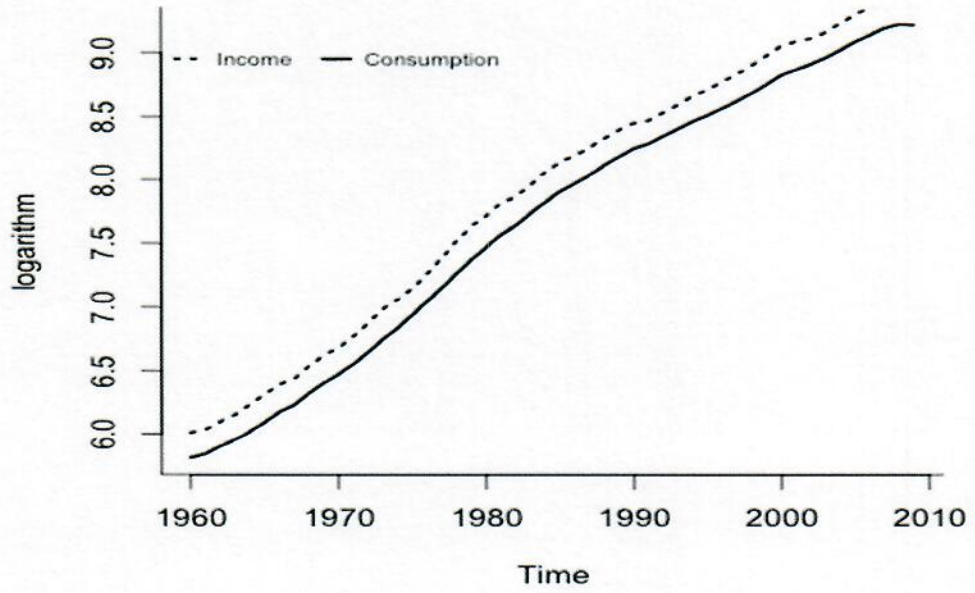

***Table 4.1**: Regression results*

	Dependent variable: Consumption		
	Static regression	Dynamic regression	
		(Unadjusted std. errors)	*(Robust std. errors)*
	(1)	(2)	(3)
Income$_t$	0.819***	0.438***	0.438***
	(0.003)	(0.014)	(0.011)
Consumption$_{t-1}$		0.484***	0.484***
		(0.018)	(0.014)
Constant	-31.888*	10.939**	10.939**
	(18.227)	(4.915)	(4.989)
Observations	50	49	
R^2	0.999	0.999	
Adjusted R^2	0.999	0.999	
Residual Std. Error	83.867 (df = 48)	20.863 (df = 46)	
F Statistic	65,939.630*** (df = 1; 48)	521,204.800*** (df = 2; 46)	
Note:			*p<0.1 **p<0.05 ***p<0.01

Breusch-Godfrey test for dynamic regression model: LM test = 12.149, df = 1, p-value = 0.0004912

ADF unit root test for residuals from the static regression model
Value of test-statistic is: -3.1012
Critical values for test statistics:
1% 5% 10%
tau1 -2.62 -1.95 -1.61

(a) Explain the meaning of an I(1) series. **[2 marks]**

(b) Based on Figure 4.1 above comment on the behavior of the two variables. **[3 marks]**

(c) What is spurious regression? **[2 marks]**

(d) Comment on the Breusch-Godfrey test results. **[4 marks]**

(e) Test for unit roots in the residuals of the static regression model using 1% level, and comment on the results. **[4 marks]**

(f) What are the short-run and long-run marginal propensities to consume? **[3 marks]**

(g) Use your preferred notation to derive the ADF unit root test regression from the random walk process. **[4 marks]**

(h) What are the limitations of the ADF test for unit root? Explain. **[3 marks]**

SECTION C: (60%)

Answer TWO Questions from this Section, ONE of which MUST be Question 5 or 6

Question 5

Treatment evaluation is concerned with measuring the impact of an intervention on outcomes of interest. The policy relevance of this subject is direct since successful treatments can be linked to desirable social programs, or improvements in existing programs to attain objectives of a social policy.

(a) Give an example of an impact evaluation research problem and explain how the treatment effect model may be applied. **[4 marks]**

(b) Define the following concepts:

(i) Counterfactual

(ii) Average treatment effects (ATE)

(iii) Average treatment effects on the treated (ATET)

(iv) Average treatment effects on the untreated (ATEU)

(v) Selection bias

(vi) Control group **[6 marks]**

(c) In the treatment effect model, the average treatment effect on the treated (ATET) is given as:

$$ATET = E[(Y_1 \mid D=1) - (Y_0 \mid D=1)] \quad (5.1)$$

where Y_1 is the outcome for the treated individual
Y_0 is the outcome for the untreated individual
D is an indicator of treatment, D = 1 if treatment is present; D = 0 otherwise.

Show that under some conditionality, the ATET may be estimated using:

$$ATET = E[(Y_1 \mid D=1) - (Y_0 \mid D=0)] \quad (5.2)$$ **[7 marks]**

(d) Describe the basic steps in using the propensity score method (PSM). **[6 marks]**

(e) A researcher investigated whether there is a significant difference in maize yields between those farmers who are participants and non-participants of One-Acre Fund (OAF) program in the Western Province in Kenya using PSM. Aside from participation in the program, maize yield is also a function of various socio-economic characteristics. Estimation results from STATA are provided below. Comment and interpret the results. **[7 marks]**


```
. pscore oafmembership sex age agesq marriage hhsize residence educ , pscore(mypscore)

The treatment is oafmembership
      farmer's |
 membership in |
           OAF |      Freq.     Percent        Cum.
---------------+-----------------------------------
OAF non-member |        126       47.73       47.73
    OAF member |        138       52.27      100.00
---------------+-----------------------------------
         Total |        264      100.00

Estimation of the propensity score
Probit regression                                 Number of obs   =        264
                                                  LR chi2(7)      =      30.81
                                                  Prob > chi2     =     0.0001
Log likelihood = -167.31438                       Pseudo R2       =     0.0843
------------------------------------------------------------------------------
oafmembers~p |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.2688997   .1629024    -1.65   0.099    -.5881825     .050383
         age |   .1242697   .0486731     2.55   0.011     .0288722    .2196672
       agesq |  -.0012759   .0005262    -2.42   0.015    -.0023073   -.0002445
     marriage|  -.1948612   .0998065    -1.99   0.047    -.3904784     .000756
      hhsize |  -.0582892   .0310746    -1.88   0.061    -.1191944     .002616
    residence|  -.6664695   .1733583    -3.84   0.000    -1.006246   -.3266935
        educ |    .049894   .1002478     0.50   0.619    -.1465881    .2463761
       _cons |  -1.664632   1.089726    -1.53   0.127    -3.800455    .4711922
------------------------------------------------------------------------------
Description of the estimated propensity score
                Estimated propensity score
-------------------------------------------------------------

      Percentiles      Smallest
 1%     .1415462       .0978625
 5%     .2382715       .1052622
10%     .3149542       .1415462       Obs                 264
25%      .405464       .1913692       Sum of Wgt.         264

50%     .5122674                      Mean           .5226751
                       Largest        Std. Dev.      .1672656
75%     .6704275       .8286752
90%     .7740445       .8305884       Variance       .0279778
95%     .7944076       .8328853       Skewness       .0075745
99%     .8305884       .8530969       Kurtosis       2.349435


. attnd yield oafmembership sex marriage hhsize residence educ landsize soiltype
agesq, comsup boot reps(100) dots logit

The program is searching the nearest neighbor of each treated unit.
```

This operation may take a while.

ATT estimation with Nearest Neighbor Matching method

(random draw version)

Analytical standard errors

n. treat.	n. contr.	ATT	Std. Err.	t
137	51	3.084	1.512	2.040

Note: the numbers of treated and controls refer to actual

nearest neighbour matches

/* Descriptions of variables used in the estimation

age is the age of maize farm head at the time of survey in years

agesq is the square of age of a farm head in years

educ is the highest level of formal schooling attained by a farm head and measured as 0 = none, 1 = primary, 2 = secondary, 3 = post-secondary.

hhsize is the number of household members

landsize is the area covered by maize plantation measured in acres

marriage is a dummy variable, 1 if married, 0 otherwise.

oafmembership indicates membership to the One-Acre fund program, 1 if a member, 0 otherwise

residence is sub-county a maize farmer's residence and is a dummy: 1 if in Nambale Sub-county, 0 otherwise.

Soiltype is the type of soil measured as dummy; 1 if loam soil, 0 otherwise

sex of farm head, measured as dummy; 1 = male, 0 = female.

yield is the quantity of maize produced for a unit area of land. It was measured in 90kg bags/acre

Question 6

Panel data refers to the pooling of cross-section observations over several periods of time. Using panel data in econometric modeling is said to have several advantages over the use of pure cross section or pure time series data.

(a) State three advantages and one limitation of panel data and explain. **[4 marks]**

(b) Various types of panel models may be used depending on the assumptions made on the intercept and slope coefficients. Enumerate three types of these models. For each model, state the assumptions on the coefficients and how each may be estimated. **[6 marks]**

(c) When lags of the dependent variable are added as additional regressors in a panel data model, none of the static model estimators is appropriate to use in a limited sample. Describe any two alternative estimators to consistently estimate the parameters of a dynamic panel. **[4 marks]**

(d) Testing for unit roots has become standard and has been extended to panel data.

(i) Differentiate between the Levin, Lin and Chu (LLC) panel unit root test and the Im, Pesaran and Shin (IPS) panel unit root test. **[4 marks]**

(ii) Consider a set of macroeconomic variables (all in million of US dollars) for 14 African countries from 1985-2003 which were tested for unit roots, the results of which are reported in Table 6.1. Comment on the order of integration of each variable using 5% level of significance. **[4 marks]**

Table 6.1 LLC and IPS unit roots test results

Variables	**At levels**		**First Differences**	
	LLC - Stat	*IPS W-Stat*	*LLC - Stat*	*IPS W-Stat*
Direct Tax	-1.200 (0.115)	-0.006 (0.498)	-9.389 (0.000)	-9.045 (0.000)
Output	0.139 (0.556)	1.302 (0.904)	-10.879 (0.000)	-7.912 (0.000)
FDI	-6.049 (0.000)	-4.714 (0.000)	-11.251 (0.000)	-9.789 (0.000)
ODA	-3.858 (0.0001)	-3.118 (0.0009)	-10.977 (0.000)	-9.574 (0.000)

Notes: Numbers in parentheses are computed *p-values*.
All the *test equations* include individual intercept and trend.
FDI and ODA denote foreign direct investment, and official development assistance (i.e. foreign aid), respectively.

(e) Consider the following static panel data model explaining the variations in tax revenues for the set of data given in (d) above:

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it} \qquad i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T \qquad (6.1)$$

where y_{it} = observation on direct tax of *i*th country in *t*th period

x_{it} = vector of regressors composed of output, FDI and ODA for the *i*th country in *t*th period.

Three versions of model in (6.1) were estimated, the results of which are presented in Table 6.2. The dynamic version was also estimated, where the lag of the dependent variable was added as a regressor. The estimated results are shown in Table 6.2.

Table 6.2: Panel data estimates of the tax equation

Dep variable: Direct Tax	*Pooled* - Coefficients	*FE* - Coefficients	*RE* - Coefficients	*GMM – Arellano & Bond*
Output	0.145 (14.15)	0.134 (5.54)	0.137 (7.48)	0.071 (1.71)
FDI	1.054 (10.36)	0.656 (6.54)	0.708 (7.24)	0.194 (2.12)
ODA	-0.454 (-6.27)	-0.084 (-0.82)	-0.174 (-1.83)	0.047 (0.19)
Lag of Direct Tax	-	-	-	0.594 (8.35)
Constant	-34.347 (0.76)	-101.678 (-1.09)	-86.339 (-0.88)	-102.038 (-0.66)

Notes: Figures in parentheses are *t*-values using robust variance estimates.

Critical values of t = 2.58, t = 1.96, t =1.645 respectively for 1%, 5%, and 10% significance level.

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_N$; $F_{calc} = 14.45$; *p-value* = 0.000

H_0: α_i uncorrelated with x_i ; $\chi^2(3) = 23.30$; *p-value* = 0.000

H_0: zero autocorrelation in first differenced errors; $Z = -1.51$; *p-value* = 0.132

H_0: overidentifying restrictions are valid; $\chi^2(159) = 9.54$; *p-value* = 0.999

Based on the estimated results and data characteristics, which of the models is the most appropriate to the problem at hand. Justify your answer and interpret the results from your most appropriate model. **[8 marks]**

Question 7

(a) What is the advantage of non-linear time series models over the linear models? **[3 marks]**

(b) What are the main two categories of non-linear time series models? For each category, give two examples. **[4 marks]**

(c) Given the following stochastic processes

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

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

Determine which process is ergodic stationary in the mean. Explain. **[8 marks]**

(d) Given an AR(1) model

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

Show that if $\phi = 1$

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

(e) What are the main fundamental changes in the nature of asymptotic distribution theory as we move from stationarity to non-stationarity (stochastic trends)? **[7 marks]**

Question 8

(a) A PhD student carried out a research which involves among other things estimation of demand for cloth function. The demand function was specified as follows:

$$q_t = \beta_1 + \beta_2 p_t + \beta_3 i_t + \beta_4 D_t + \varepsilon_t \qquad (8.1)$$

where q_t is the logarithm of the quantity demanded, p_t is the logarithm of price, i_t is the logarithm of income, and D_t is a dummy variable representing regular demand shifts. p_t is suspected to be endogenous, and, x_{1t} and x_{2t} have been identified as instruments for p_t. The student decided to estimate Equation (8.1) using OLS, two-stage-least-squares (TSLS) and generalized method of moments (GMM), using both x_{1t} and x_{2t} as instruments (Results of which are presented in Table 8.1). Endogeneity and heteroskedasticity tests were carried out, and the results of which are also reported in Table 8.1 Furthermore, the student estimated the reduced form for p_t, the results of which are reported in Table 8.2.

(**Note**: *All continuous variables are stationary*)

Table 8.1. Regression results

	OLS	**TSLS**	**GMM**
Intercept	88.1274***	109.7915***	109.8727***
	(8.3435)	(11.8916)	(10.3408)
Price	-0.8417***	-1.5323***	-1.5311***
	(0.0866)	(0.1658)	(0.1822)
Income	0.8301***	1.3623***	1.3605***
	(0.1200)	(0.1861)	(0.1678)
Dummy	-23.8083***	-31.2367***	-31.2045***
	(5.4024)	(7.4303)	(7.1928)
R^2	0.5799	0.2282	
Adj. R^2	0.5633	0.1977	
Num. obs.	80	80	80
RMSE	20.5632	27.8720	
Criterion function			105.9127

*p<0.1 **p<0.05 ***p<0.01

Test of endogeneity of p

Durbin-Wu-Hausman (DWH) Chi-square test	96.021
p-value	0.0001

Pagan-Hall general test statistic for heteroskedasticity: 23.612 Chi-sq(4) P-value = 0.0141

Table 8.2 Results for reduced form equation for p_t

	Reduced form regression	
Income	0.565***	(0.050)
Dummy	-12.43**	(3.956)
Instrument1	-0.436***	(0.062)
Instrument2	-0.213***	(0.063)
Intercept	81.14***	(7.217)
N	80	
R^2	0.723	
Standard errors in parentheses		
***p < 0.001, **p < 0.01, *p < 0.05		
H0: $x_{4t}=x_{5t}=0$		
F-test = 68.92, p-value = 0.0003		

Sargan 's test statistic for over-identification: 0.00077341, p-value = 0.97781350

(i) Briefly explain the steps involved in the empirical implementation of the DWH endogeneity test in this case. **[5 marks]**

(ii) What are the implications of the DWH endogeneity test results? Use 5% level of significance in the interpretation **[3 marks]**

(iii) What are the implications of the results of Sargan's test for over identification? Use 5% level of significance in the interpretation. **[5 marks]**

(iv) In light of the reported DWH endogeneity test results, which estimator would you recommend? Explain. **[5 marks]**

(b) Briefly describe Sims' (1980) modeling philosophy. **[4 marks]**

(c) Explain how Johansen's approach to cointegration test addresses the limitations inherent in the Engle-Granger approach to cointegration analysis. **[4 marks]**

(d) Consider the following results of performing Johansen's test for cointegration on a VAR system with 3 variables (money supply, interest rates and output). These results are obtained using real data for a certain Sub-Saharan African country.

Rank	Eigenvalue	Trace test (p-value)	Maximum eigenvalue test (p-value)
0	0.23449	35.168 (0.0101)	27.256 (0.0046)
1	0.066593	7.9116 (0.4821)	7.0292(0.4942)
2	0.0086135	0.88238 (0.3476)	0.88238 (0.3476)

Use the reported results to test for cointegration at 5% level of significance. **[4 marks]**

APPENDIX

Relevant Probability Distributions

Normal Distribution

If x is a normal random variable with mean $=\mu$, and variance $=\sigma^2$, then

p.d.f. $$\phi(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

Standard Normal Distribution

p.d.f. $$\phi(x)=\frac{1}{\sqrt{2\pi}}\exp\{-x^2/2\}$$

c.d.f. $$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{t^2}{2}\}dt$$

Logistic Distribution

c.d.f. $$F(z)=\Lambda(z)=\frac{e^z}{1+e^z}$$

p.d.f. $$f(z)=\Lambda(z)[1-\Lambda(z)]$$