FACULTY/COLLEGE		College of Business and Economics	
SCHOOL		School of Economics	
CAMPUS(ES)		APK	
MODULE NAME		Econometrics 4B	
MODULE CODE		ECM8X02	
SEMESTER		Second	
ASSESSMENT OPPORTUNITY, MONTH AND YEAR		Final Summative Assessment Opportunity October/December 2022	
ASSESSMENT DATE	October/December 2022	**SESSION**	08:30 – 11:30
ASSESSOR(S)	Dr Kwame P. Osei-Assibey		
MODERATOR(S)	Dr T. Mugadza (External); Dr Josine Uwilingiye (Internal)		
DURATION	3 hours (180 min)	**TOTAL MARKS**	100
NUMBER OF PAGES OF QUESTION PAPER (Including cover page)			6

INFORMATION/INSTRUCTIONS:

- This is a closed-book assessment.
- Answer all questions
- Answer each question in a separate book as instructed.
- Data and statistical table required for analyses are saved on your assigned computer.
- Candidate audited STATA command sheets (maximum of 5 commands allowed) will be distributed.
- Number your answers clearly and correctly as per the question paper.
- Write neatly and legibly on both sides of the paper in the answer book, starting on the first page.

Section A. (Theory and Concepts)

QUESTION ONE **[45 MARKS]**

1.1. In their paper *The Effect of Fast-Food Restaurants on Obesity and Weight Gain*, Currie et al (2010) examine how locating a new fast-food restaurant close to a high school affects the body weight of 9th graders at the school. Let's suppose that you gather some of the same type of data as the authors of this paper for Gauteng Province. First, you determine which high schools in Gauteng did not have a fast-food restaurant within 0.1 miles of the school in 2004. This will be your sample of high schools. Next, among these schools in your sample, you identify those schools where a new fast-food restaurant opened up within 0.1 miles of the school in 2005. Finally, you gather information on the percent of obese 9th graders (as measured in the late spring of 9th grade year) at each school in your sample for the years 2004 and 2006. The average obesity rates for the schools that **have** a new nearby fast-food restaurant estimated in 2006 are: 0.29 (in 2004) and 0.40 (in 2006). The average obesity rates for the schools that **do not have** a new nearby fast-food restaurant estimated in 2006 are: 0.30 (in 2004) and 0.33 (in 2006).

1.1.1. Can this analysis be classified as a natural experiment? Explain. **(3)**

1.1.2 Clearly identify the treatment and the control groups for this experiment. **(2)**

1.1.3 Use the difference-in-differences estimator to determine whether the opening of a new fast-food restaurant within 0.1 mile of a school affected the obesity rates of 9^{th} graders. **(5)**

1.1.4 There are two main advantages of pooling the information across time (data on both the control group and the treatment group) and running a single regression framework to examine the effect of the policy (siting fast-food restaurant within 0.1 mile of a school). Mention these two advantages. **(2)**

1.1.5 Write down the pooled regression model that allows you to directly estimate the impact of the policy in a single pooled OLS regression framework. Be sure to define all of the variables. Indicate the slope coefficient that compares to the one you obtained from the difference-in-differences procedure. **(5)**

1.2 In this question, we consider the returns to education study by Card (1993). Assuming the equation of interest is as follows:

$$Log\ (Wage_i) = \beta_0 + \beta_1 Education_i + u_i$$

1.2.1. State the major assumption (about u_i) under which OLS will consistently estimate the effect of education on wages. Do you think this assumption always hold? Explain. **(3)**

1.2.2 Assuming education is endogenous because of unobserved ability. Give an example for a proxy variable that could be used in place of ability. Does including it into the model make the OLS estimator of the parameter on education consistent for the returns to education? Provide reasons. **(3)**

1.2.3 Card recommended proximity (**prox**) to a university (college) as well as the number of siblings (**sib**) a respondent have as instruments for education. List the characteristics of each of the variables that qualifies them to be good instruments candidates. **(3)**

1.2.4 Define the terms: over identified, exactly identified and under identified coefficient in 2SLS regression. How will you classify the coefficient of education? **(4)**

1.2.5 Write down the first stage reduced form equation for the 2SLS procedure. **(1)**

1.2.6 Explain (using the relevant equations and test hypothesis) how the first stage reduced form equation is used test whether education is truly endogenous. **(4)**

1.2.7 Explain (using the relevant equations and test hypotheses) how the first stage reduced form equation is used test whether the instruments are relevant and strong. **(5)**

1.3 Suppose we have data on $Y_1, Y_2 \ldots, Y_n$ and each Y_i is a count variable drawn from a Poisson distribution and takes the values $y_i = 0,1,2,3 \ldots, n$. Take λ as the mean rate of occurrence (average number of successes within an interval) determined within a Poisson regression as follows:

$E(Y_i) = \lambda = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots, + \beta_k x_{ki})$ or in compact form

$$E(Y_i) = \lambda = exp(X_i'\beta)$$

1.3.1 What is the probability distribution, $P_r(Y_i = y_i/x_i)$ of each observation? **(2)**

1.3.2 Write down the mathematical expression of the marginal effect of x_{2i} on $E(Y_i)$ **(1)**

1.3.3 With reference to the assumption of equidispersion property of Poisson models, explain what is meant by overdispersion and its implications. How do we correct for overdispersion problem? **(2)**

Section B (Practice and Applications Using STATA)

QUESTION 2 **[27 MARKS]**

In the paper *COVID Induced Uncertainty*, Baker et al (2020) examine the impact of the COVID-19 pandemic on policy uncertainty across the globe. In the task that follows, we attempt a similar but simple study using quarterly data for twenty randomly selected countries across the globe in the year 2020. The dataset for this task is the **pandemic** dataset and the variables for the assessment are:

polunc (dependent variable)- Is the economic policy uncertainty variable.

panunc - Is the pandemic uncertainty variable. ***panunc is in levels. You have to log transformed it before using it in your regression.***

2.1 At the first stage, you are requested to estimate the relationship using Pooled OLS. The econometric model to be estimated by OLS is as follows:

$$polunc_{it} = \beta_0 + \beta_1 \log(panunc)_{it} + \varepsilon_{it}$$

i. Describe the expected effect of the independent variable on the dependent variable. **(1)**

ii. Estimate the model using pooled OLS and present your estimated model with t-statistics in brackets right below the estimated coefficient. Provide a discussion on the overall fitness of the estimated model and the statistical significance of the estimated coefficients. **(5)**

iii. Mention any two potential sources of bias with the application of OLS to panel data. **(2)**

2.2 You want to deal with the potential biases in the estimated coefficients. You are contemplating applying the fixed effects (FE) method or the random effect (RE) methods.

i. Present the functional form of the linear unobserved heterogeneity model for your analysis **(1)**

ii. State the assumptions about the unobserved heterogeneity component that should inform you on your choice between the FE and RE estimation techniques. **(2)**

iii. What is the cross-sectional identifier of the **pandemic** dataset? **(1)**

iv. Estimate the model using both the FE and RE techniques. In your presentation ensure that you include analyses on the statistical significance of the estimated coefficients and compare results to those from the pooled OLS estimation. **(6)**

v. Use the appropriate tests to choose between your estimated Pooled OLS, RE and FE models. Present all the statistics you used to arrive at your conclusion. **(4)**

2.3 Modify your preferred model obtained from Question 2.2 (part v) by including time dummies and re-estimate (Note that you need to generate the time dummies).

i. Present your estimated model with t-statistics in brackets right below the estimated coefficient. Discuss the results **(2)**

ii. Can you identify any trend effect? Test the joint significance of the period dummy variables in your model. **(3)**

QUESTION 3 **[28 MARKS]**

The final exercise uses the apple dataset. The data is provided by Blend and Ravenswaay (1999) who used it to measure consumer demand for ecologically friendly apples in their paper titled *Measuring Consumer Demand for Ecolabeled Apples* . In this exercise, you are going to analyse the determinants of demand for ecologically friendly apples (ecolabeled apples) using probability regressions. The regression model to estimate is:

$$ecobuy_i = \beta_0 + \beta_1 ecoprc_i + \beta_2 regprc_i + \beta_3 faminc_i + \beta_4 hhsize_i + \beta_5 educ_i + \beta_6 age_i + u_i$$

Where:

ecoprc **–**Price of ecolabeled apples.

regprc – Price of regular apples.

faminc – Family income.
hhsize – Household size.

educ - Years of schooling of respondent

age – Age of respondent.

ecolbs – The quantity of ecolabeled apples purchased at any given price. It is not part of the dependent variable but will be used to generate the binary dependent variable ***ecobuy***

ecobuy is a binary variable defined as:

$$\boldsymbol{ecobuy} = \begin{cases} 1 \; if \; ecolbs > 0 \\ 0 \quad if \; ecolbs = 0 \end{cases}$$

i.e., ***ecobuy*** assumes a value of 1 if a respondent buys an ecolabeled apple at any given price and 0 otherwise.

To generate data on ecobuy binary variable as defined above in STATA, type the command: **gen ecobuy = (ecolbs>0)**

3.1 What are your expectations of the impact of regprc, educ and faminc on a respondent demand for ecolabeled apple? **(3)**

3.2 Use STATA to perform an appropriate linear probability model (LPM) regression and present your estimated models with t-statistics in brackets right below the estimated coefficients. Discuss the statistical significance of the estimated coefficients. **(4)**

3.3 From your estimated coefficients, explain the impact of regprc, educ and faminc on respondents' probability of buying an ecolabeled apple: Are these consistent with your expectations? **(4)**

3.4 Formulate the hypothesis to test whether the nonprice variables are jointly significant. Perform the test using STATA and state your decision and the statistics you used to arrive at it. **(4)**

3.5 Obtain the fitted valued for ***ecobuy*** based on your LPM estimation. Are any of the fitted values negative or greater than one? What is the implication on the robustness of your LPM model? **(3)**

3.6 Use STATA to perform a probit regression and present your estimated models with z-statistics in brackets right below the estimated coefficients. Discuss the statistical significance of the estimated coefficients. **(4)**

3.7 Interpret the PCP (percent correctly predicted) of the estimated probit model. Explain the sensitivity and specificity statistics. **(3)**

3.8 What is the PEA and APE of ***hhsize***? What is the difference between the two measurements in terms of their definition? **(3)**