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		Supplementary Summative Assessment Opportunity January/ February 2021	
ASSESSMENT DATE	TBC	**SESSION**	TBC
ASSESSOR(S)	Dr Kwame P. Osei-Assibey		
MODERATOR(S)	Dr T. Mugadza (External); Sutene Mwambi (Internal)		
DURATION	3 hours (180 min)	**TOTAL MARKS**	100
NUMBER OF PAGES OF QUESTION PAPER (Including cover page)			5

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.

QUESTION ONE [40 MARKS]

1.1.1 Explain what we mean by natural experiment. In your explanation, clearly distinguish how the use of control and treatment groups help us to examine effects of a policy. **(5)**

1.1.2 The following regressional framework was designed to examine the effect of a policy change on a group of people;

$$y_{it} = \alpha + \beta D_i + \gamma Post + \delta(D_i \times Post) + \varepsilon_{it}$$

Where

- y_{it} is the outcome variable of interest
- D_i is a dummy variable which equals 1 for respondents in the treatment group and 0 otherwise.
- $Post$ is also a dummy variable which equals 1 for the post-policy period and 0 otherwise

Use the given regression above to fill in the missing spaces. **(6)**

	Before	**After**	**After-Before**
Control	α		
Treatment			$\gamma + \delta$
Treatment-Control	β		

1.2.1 Write down a hypothetical linear unobserved (heterogeneity) effects regression model for panel that has data measured across a three year period. Define all the components of your model. **(3)**

1.2.2. Write out the first difference regression model for the three period linear unobserved regression model above. Define all the components of your model. **(3)**

1.3 Consider the following regression model;
$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots \beta_k x_{itk} + a_i + u_{it}$
Where a_i is the fixed effects component and d_t is the time effect component

1.3.1 What is the reason for inclusion of the fixed effects component? **(2)**

1.3.2 What is the reason for inclusion of the time effect component? **(2)**

1.3.3 You are thinking about using either fixed effect or random effect estimators to analyze a panel data based on the regression model above: State the two key assumptions about a_i that can help you choose between the two estimators. **(1)**

1.4 Consider the following simple regression model
$y_i = \alpha + \beta x_i + u_i$: Where $Cov(x_i, u_i) = \rho \; and \; \rho > 0$.
Suppose there exist an instrument z such that $Cov(z_i, u_i) = 0$

1.4.1 State the probability limit $(plim)$ for both the OLS and IV estimator of β **(2)**

1.4.2 If z is a poor instrument, what is the direction of the asymptotic bias if $corr(z, x) > 0; corr(x, u) > 0$? **(4)**

1.5 Explain the following concepts, supporting your responses with the relevant examples:

1.5.1. Corner solution responses **(3)**

1.5.2. Censoring problem **(3)**

1.6 Consider a Tobit model with censoring from below at 0, where

$y_i^* = \beta_0 + \beta_1 x_i + u_i \quad u_i / x_i \sim N(0, \sigma^2)$; And

$$\begin{cases} y_i = 0 \; if \;\; y_i^* \leq 0 \\ y_i = y, \; if \;\; y_i^* > 0 \end{cases}$$

1.6.1 What is the conditional expectation of positive outcomes (i.e. $E(y_i / y_i > 0)$ **(3)**

1.6.2. What is the conditional expectation for all outcomes (i.e. $E(y_i / x_i)$ **(3)**

QUESTION 2 [35 MARKS]

This exercise uses the MROZ dataset. It is a cross-sectional data consisting of 753 observations on 22 variables. One can use a number of characteristics contained in this data to predict the labour force participation of women.
For this exercise, the dependent variable is the annual number of hours supplied by women respondents (***hours***). The explanatory variables are wages (***lwage***); education (***educ*** is years of education); having kids who are less than 6 years (***kidslt6***); age of respondents (***age***) and income not related to wages (***nwifeinc***)

2.1 Write down the population regression model to be estimated. What are your expectations of the impact of each of the independent variables on labour participation of women (hours)? **(4)**

2.2 Estimate the model using OLS and present your estimated model with t-statistics in brackets right below the estimated coefficient. Discuss the statistical significance of the estimated coefficients and explain whether your obtained results are consistent with your expectations. **(6)**

2.3 Indicate three reasons why lwage might be endogenous. **(3)**

2.4 Consider **exper** and **expersq** as potential instruments for **lwage**; list two reasons why you think these variables can serve as good instruments. Perform the necessary test(s) to ascertain if these instruments are relevant and strong for lwage. Present all the statistics you used to arrive at your conclusion. **(7)**

2.5 Using the suggested instruments perform a 2SLS regression. Present your estimated model with z-statistics in brackets right under the estimated coefficient. Discuss the statistical significance of the estimated wage coefficient. Did the instruments regression improve your result? **(7)**

2.6 Test for endogeneity. Based on the evidence from the endogeneity test results, would you say it was wise after all to perform the IV regression? Explain? **(4)**

2.7 Test to ascertain if there is an over-identification problem. **(4)**

QUESTION 3 [25 MARKS]

This exercise uses the affairs dataset. The data is collected from surveys filled out by readers for a magazine. One can use a number of characteristics contained in this data to predict the probability of people in having an affair. For our exercise, the dependent variable is the binary variable affair (***affair***=1 if the person had at least one affair, 0 otherwise). The explanatory variables for this exercise are religiosity (***relig*** is a categorical variable from 1 to 5, where 5 is very religious); happiness in marriage (***ratemarr*** is a categorical variable from 1 to 5, where 5 is very happy); gender of respondent (=1 if ***male***, 0 otherwise); age of respondents (***age***); number of years married (***yrsmarr***); kids (***kids***=1 if respondents have kids, 0 otherwise) and years of education (***educ***).

3.1 Write down the population regression model to be estimated. What are your expectations of the impact of each of the independent variables on a respondent having an affair? **(5)**

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. **(5)**

3.3 From your estimated coefficients, explain the impact of religiosity, happiness in marriage and having kids on respondents' probability of having an affair. Are these consistent with your expectations? **(5)**

3.4 Give two reasons why a logit regression may produce a better fit over the LPM**. (2)**

3.5 Use STATA to perform a logit regression and present your estimated models with t-statistics in brackets right below the estimated coefficients. Discuss the statistical significance of the estimated coefficients.

3.6 Compare the expected probability under both the LPM and the logistic model of having an affair for a very religious (relig=5) 25-year old female, without kids, who is very happily married (ratemarr=5) for a year and has 9 years of education. What do you observe and how does this inform your choice between the two models? **(4)**

3.7 Estimate the odds ratios for all explanatory variables. Present your estimated models with t-statistics in brackets right below the estimated odds ratios. Using the result for education as an example, explain what they represent. **(4)**