Module Code
ECO00005M

MSc Degree Examinations 2021-22

Department:
Economics

Title of Exam:
Applied Microeconometrics

Time Allowed: 24 Hours
PLEASE NOTE: Submissions up to 30 minutes late will receive a 5 mark penalty deduction. Submissions late by more than 30 minutes will not be marked.

Time Recommended:
TWO hours

Allocation of Marks:
ALL questions have the same weight. The weights of the parts of each question vary and are stated in percentage points.

Word Limit:
Each question has a word limit of 1,200 words. You have the flexibility on how to split the word limit across the different sub-questions.

The word count does not include: i) figures and graphs; ii) equations; iii) a bibliography at the end of the question. A bibliography is **NOT** required. In-text references (Author, date) are sufficient.

Markers will not read or mark any material beyond the word limits.

Instructions for Candidates:
Answer TWO out of the 4 questions. If you answer more than two questions, only the first two answers in the order they appear in the script will be marked. Any answers you do not wish to be included in the marking must be clearly crossed out.

Materials Required:
Calculator

A note on Academic Integrity

We are treating this online examination as a time-limited open assessment. You are therefore permitted to refer to written, and online materials, to aid you in your answers.

However, you must ensure that the work you submit is entirely your own, and for the whole time the assessment is live you must not:

- communicate with departmental staff on the topic of the assessment
- communicate with other students on the topic of this assessment
- seek assistance with the assignment from the academic and/or disability support services, such as the Writing and Language Skills Centre, Maths Skills Centre and/or Disability Services. (The only exception to this will be for those students who have been recommended an exam support worker in a Student Support Plan. If this applies to you, you are advised to contact Disability Services as soon as possible to discuss the necessary arrangements)
- seek advice or contribution from any third party, including proofreaders, friends, or family members.

We expect, and trust, that all our students will seek to maintain the integrity of the assessment, and of their award, by ensuring that these instructions are strictly followed.

Failure to adhere to these requirements will be considered a breach of the Academic Misconduct regulations, where the offences of plagiarism, breach/cheating, collusion and commissioning are relevant - see AM.1.2.1" *(Note this supersedes section 7.3 of the Guide to Assessment).*

Question 1

1,200 words limit

(a) [20%]
A researcher estimates using a random sample of workers the following log wage regression

$$log(wage_i) = \beta_0 + \beta_1 S_i + \beta_2 sibs_i + \beta_3 wexp_i + \beta_4 (wexp_i)^2 + u_i, \quad (1.1)$$

where i is the subscript for individuals and goes from 1 to n, $log(wage_i)$ is the natural logarithm of hourly wage, S_i is schooling in number of years, $sibs_i$ is the number of siblings and $wexp_i$ is the number of years of work experience. Provide details on the interpretation of the coefficients. Explain why schooling S_i could be endogenous and describe a potential cause of this endogeneity. Explain why the ordinary least squares estimator of model (1.1) is biased and inconsistent. More points will be given for a formal explanation. Make sure to adapt formulas and discussion to model (1.1).

(b) [15%]
Define two different instruments that could be used for schooling S_i to solve the issue of endogeneity in model (1.1). Define the formula for the two-stage least squares estimation for model (1.1) using the two instruments you proposed and explain how the first and second stages' estimations are computed. Make sure to adapt formulas and discussion to model (1.1).

(c) [20%]
Explain what assumptions your instrumental variables must satisfy to produce a consistent estimation of model (1.1). Explain why you think the two instruments proposed in (1.b) are satisfying these assumptions. Show that the instrumental variable estimation defined in (1.b) is consistent under these assumptions.

(d) [20%]
Explain how you would test for the validity of your instrumental variables. Explain also how you would test for whether there is an endogeneity issue in your model. Provide details on how you would perform these tests for model (1.1).

(e) [25%] The researcher decided to estimate model (1.1) using two-stage least squares estimation and instrumenting schooling with the birth order (*brthord*), which is a variable taking value 1 if the individual is a first-born child, 2 if he/she is a second-born child, and so on. Do you think it is a valid instrument? Notice that the model is conditional on *sibs* so variation in *brthord* is not explained by the number of siblings. Explain in detail your answer first under the assumption that *brthord* is moderately correlated with *sibs* and then under the assumption that the correlation between *brthord* and *sibs* is almost *1.*

Question 2
1,200 words limit

(a) [30%]
Discuss an empirical example of a panel data model where you would use a fixed effect estimation rather than a random effect estimation. The example must consider a panel data of children but you are free to consider any dependent variable. You could consider e.g. school test scores, cognitive skills, socio-emotional skills, health outcomes, level of competitiveness, measure of social skills, average wage for the aspired occupation, parental time investment in the child, expenditure in child private tuition, child health expenditure, hours of physical activity per week, time spent in school, calories intake per day, height, birth weight, BMI, days of absence in school, etc. Write the regression equation and provide details on the dependent and explanatory variables, on the error term and on the interpretation of the coefficients. Explain how you would compute the fixed effect estimation for your defined model.

(b) [35%]
What does the unobserved individual effect in the model defined in (2.a) capture? Explain the differences in the assumptions needed for the consistency of the fixed effect estimation and of the random effect estimation for the model you discussed in (2.a). Why is the fixed effect estimation more appropriate than the random effect estimation in the empirical example you discussed in (2.a)? Explain how you would perform a test to decide whether to adopt a random effect or a fixed effect estimation.

(c) [35%]
Discuss an empirical example of a panel data model where one of the explanatory variables is endogenous because it is correlated with unobserved variables that are relevant to explain both the dependent variable and the endogenous variable. The example must be different from those in lectures, seminars and past exams. Explain under which conditions the fixed effect estimation can solve such an issue of endogeneity. Explain which type of estimation you would adopt to solve the endogeneity issue if the conditions for the consistency of the fixed effect estimation were not satisfied.

Question 3
1,200 words limit

(a) [20%]
Black people are much more likely to have high blood pressure and a researcher wants to understand whether this is caused by genetic differences between black and white people or by differences in diet between black and white people which can lead to overweight and higher level of triglycerides in the blood. The researcher can observe for a sample of black and white people the following variables:
- *highbp*, a dummy variable taking value 1 if an individual has high blood pressure and 0 otherwise,
- *black*, a dummy variable taking value 1 if an individual is black and 0 if white,
- *female*, a dummy variable taking value 1 for women and 0 for men,

- *bmi*, the body mass index which is weight in kilograms divided by height in meters squared,
- *age*, age in years,
- *tgresult*, level of triglycerides (mg/dL) in the blood,
- *tgresult2*, *tgresult* squared.

Provide an interpretation of the results that are reported in Table 3.1 below. Write down the models that the researcher has estimated. Explain what estimator the researcher has used and provide details on the optimization procedure the researcher adopted. Discuss the interpretation of the coefficients and of the tests reported in Table 3.1. Based on these results can you conclude that the higher propensity of black people for high blood pressure is caused by diet? Explain in detail your answer using any test that might be useful.

Table 3.1

```
. reg highbp black female age

      Source |       SS           df       MS      Number of obs   =    10,351
-------------+----------------------------------   F(3, 10347)     =    597.93
       Model |  373.211904         3  124.403968   Prob > F        =    0.0000
    Residual |  2152.78558    10,347  .208058914   R-squared       =    0.1477
-------------+----------------------------------   Adj R-squared   =    0.1475
       Total |  2525.99749    10,350  .244057728   Root MSE        =    .45613

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |    .105039   .0146385     7.18   0.000     .0763447    .1337333
      female |   -.091399   .0089789   -10.18   0.000    -.1089993   -.0737987
         age |   .0106311   .0002606    40.80   0.000     .0101203     .011142
       _cons |  -.0460855   .0140881    -3.27   0.001    -.0737008   -.0184702
------------------------------------------------------------------------------

. reg highbp black female bmi tgresult tgresult2

      Source |       SS           df       MS      Number of obs   =     5,050
-------------+----------------------------------   F(5, 5044)      =    146.98
       Model |   153.92635         5  30.7852701   Prob > F        =    0.0000
    Residual |  1056.46098     5,044  .209449044   R-squared       =    0.1272
-------------+----------------------------------   Adj R-squared   =    0.1263
       Total |  1210.38733     5,049   .23972813   Root MSE        =    .45766

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |    .110363   .0216769     5.09   0.000     .0678669    .1528591
      female |  -.0695884   .0129744    -5.36   0.000    -.0950238   -.0441531
         bmi |   .0272994   .0013635    20.02   0.000     .0246265    .0299724
    tgresult |   .0010101   .0001101     9.17   0.000     .0007942     .001226
   tgresult2 |  -5.66e-07   1.10e-07    -5.16   0.000    -7.81e-07   -3.51e-07
       _cons |  -.4018466   .0348891   -11.52   0.000    -.4702444   -.3334488
------------------------------------------------------------------------------

. test bmi  tgresult tgresult2

 ( 1)  bmi = 0
 ( 2)  tgresult = 0
 ( 3)  tgresult2 = 0

       F(  3,  5044) =  220.53
            Prob > F =    0.0000
```

(b) [25%]

The researcher also produced the results in Table 3.2. Write down the models that the researcher has estimated. Explain what type of estimator the researcher has used and provide details on the optimization procedure the researcher adopted. Discuss the interpretation of the coefficients and the tests reported in Table 3.2. Consider a test procedure to decide which of the two probit models considered in Table 3.2 is to be preferred. Explain in detail your answer.

Table 3.2

```
. probit highbp black female age

Iteration 0:   log likelihood = -7050.7655
Iteration 1:   log likelihood = -6241.0373
Iteration 2:   log likelihood = -6238.8309
Iteration 3:   log likelihood = -6238.8309

Probit regression                                 Number of obs =  10,351
                                                  LR chi2(3)    = 1623.87
                                                  Prob > chi2   =  0.0000
Log likelihood = -6238.8309                       Pseudo R2     =  0.1152

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |    .302711   .042485     7.13   0.000     .2194419    .3859802
      female |  -.2743514   .0261651   -10.49   0.000    -.3256341   -.2230687
         age |      .0297   .0007895    37.62   0.000     .0281526    .0312475
       _cons |  -1.529171   .0427501   -35.77   0.000     -1.61296   -1.445383
------------------------------------------------------------------------------

. probit highbp black female bmi age tgresult tgresult2

Iteration 0:   log likelihood = -3395.4388
Iteration 1:   log likelihood =  -2804.804
Iteration 2:   log likelihood =  -2801.251
Iteration 3:   log likelihood = -2801.2489
Iteration 4:   log likelihood = -2801.2489

Probit regression                                 Number of obs =   5,050
                                                  LR chi2(6)    = 1188.38
                                                  Prob > chi2   =  0.0000
Log likelihood = -2801.2489                       Pseudo R2     =  0.1750

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |   .3658577   .0643065     5.69   0.000     .2398193    .4918961
      female |  -.2607344   .0391055    -6.67   0.000    -.3373798   -.1840891
         bmi |   .0751805   .0043272    17.37   0.000     .0666995    .0836616
         age |   .0264713   .0011995    22.07   0.000     .0241203    .0288224
    tgresult |   .0018013   .0003208     5.62   0.000     .0011726      .00243
   tgresult2 |  -9.67e-07   3.02e-07    -3.21   0.001    -1.56e-06   -3.76e-07
       _cons |  -3.642458   .1256051   -29.00   0.000    -3.888639   -3.396276
------------------------------------------------------------------------------
```

(c) [30%]

Provide an explanation line by line of the Stata code that the researcher has used to produce the results in Tables 3.3 and 3.4. Explain in detail what the reported 'scaleprobit' and 'scaleprobit2' are. Use the results in Tables 3.3. and 3.4 to provide comments on the average marginal effect for each explanatory variable including *tgresult*. Based on these results can you conclude that the higher probability of black

people to have high blood pressure is caused by diet? How would you assess the fitness of the model in Table 3.4.

Table 3.3

```
. probit highbp black female age

Iteration 0:   log likelihood = -7050.7655
Iteration 1:   log likelihood = -6241.0373
Iteration 2:   log likelihood = -6238.8309
Iteration 3:   log likelihood = -6238.8309

Probit regression                                 Number of obs =  10,351
                                                  LR chi2(3)    = 1623.87
                                                  Prob > chi2   =  0.0000
Log likelihood = -6238.8309                       Pseudo R2     =  0.1152

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |    .302711   .042485     7.13   0.000     .2194419    .3859802
      female |  -.2743514   .0261651   -10.49   0.000    -.3256341   -.2230687
         age |      .0297   .0007895    37.62   0.000     .0281526    .0312475
       _cons |  -1.529171   .0427501   -35.77   0.000     -1.61296   -1.445383
------------------------------------------------------------------------------

. predict xbprobit, xb

. gen scaleprobit=normalden(xbprobit)

. sum scaleprobit

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
 scaleprobit |     10,351    .3426472    .0568408   .1919711   .3989289
```

Table 3.4

```
. probit highbp black female bmi age tgresult tgresult2

Iteration 0:   log likelihood = -3395.4388
Iteration 1:   log likelihood =  -2804.804
Iteration 2:   log likelihood =  -2801.251
Iteration 3:   log likelihood = -2801.2489
Iteration 4:   log likelihood = -2801.2489

Probit regression                                 Number of obs =   5,050
                                                  LR chi2(6)    = 1188.38
                                                  Prob > chi2   =  0.0000
Log likelihood = -2801.2489                       Pseudo R2     =  0.1750

------------------------------------------------------------------------------
      highbp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       black |   .3658577   .0643065     5.69   0.000     .2398193    .4918961
      female |  -.2607344   .0391055    -6.67   0.000    -.3373798   -.1840891
         bmi |   .0751805   .0043272    17.37   0.000     .0666995    .0836616
         age |   .0264713   .0011995    22.07   0.000     .0241203    .0288224
    tgresult |   .0018013   .0003208     5.62   0.000     .0011726      .00243
   tgresult2 |  -9.67e-07   3.02e-07    -3.21   0.001    -1.56e-06   -3.76e-07
       _cons |  -3.642458   .1256051   -29.00   0.000    -3.888639   -3.396276
------------------------------------------------------------------------------

. predict xbprobit2, xb
(5301 missing values generated)

. gen scaleprobit2=normalden(xbprobit2)
(5,301 missing values generated)

. sum scaleprobit2

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
scaleprobit2 |      5,050    .3139517    .0892273   .0280051   .3989423

. predict highbphat, p
(5,301 missing values generated)

. gen phighbp=highbphat>0.5

. tab phighbp highbp

           | High blood pressure
   phighbp |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,399        869 |     3,268
         1 |     3,576      3,507 |     7,083
-----------+----------------------+----------
     Total |     5,975      4,376 |    10,351
```

(d) [25%]

By using the sample and explanatory variables described above discuss how you would estimate a model to predict the blood pressure if the dependent variable was taking four ordered values: 0 for normal pressure, 1 for elevated blood pressure, 2 for high blood pressure and 3 for very high blood pressure. Write the likelihood for the model and provide an interpretation for all the parameters included in the likelihood. Make sure to adapt formula and discussion to this specific empirical example.

Question 4
1,200 words limit

(a) [25%]
For a sample of individuals involved in car accidents you observe their insurance reimbursement for health expenditure. Each individual got a reimbursement for health expenditure up to a maximum of £100,000. You cannot observe the individual health expenditure, but you can observe the reimbursement. Explain what type of model you would use to explain the individual health expenditure in pounds using, as explanatory variables, the value of the car before the accident in pounds, the age of the individual in years, and the number of days of hospitalization of the individual. Write down the model and explain how you would interpret the coefficients in this model. Explain why the ordinary least squares estimator would be biased and inconsistent.

(b) [20%]
By considering the sample in (4.a) and dropping all observations with health expenditure over £100,000 consider a truncated regression. Define the truncated model by adapting any formula to the specific example in (4.a) and explain why the maximum likelihood estimator of this truncated model is consistent but inefficient.

(c) [30%]
Write down the likelihoods for the models discussed in (4.a) and (4.b) and define each of the parameters and variables. Make sure to adapt formulas and discussion to the specific example.

(d) [25%]
Now assume that you observe a sample of individuals involved in car accidents whose insurance company has a fixed excess at £200 but no maximum threshold for reimbursements for health expenditure. This implies that all individuals with a damage for less than £200 do not get any payment from the insurance company. For this reason, all people with a health expenditure for less than £200 did not file any claim and are not included in the sample. Explain what type of model you would use to explain the amount of health expenditure in pounds using, as explanatory variables, the value of the car before the accident in pounds, the age of the individual in years, and the number of days of hospitalization of the individual. Write down the model and explain how you would interpret the coefficient in such a model.

END OF EXAMINATION PAPER