UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Exam: **ECON4137 – Applied Micro Econometrics**

Date of exam: Thursday, May 31, 2018 **Grades are given:** June 15, 2018

Time for exam: 09.00 to 12.00

The problem set covers 6 pages (incl. cover sheet)

Resources allowed:

- Open book examination, where all written and printed resources, in addition to some calculators, are allowed. Calculators allowed for examination:
  - Aurora HC106
  - Casio FX-85EX

The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

Exam 4137 – Spring 2018

1. The attached Stata log shows estimation results from a dataset on 453 people who have 4 commuting options: take the bus, rail, car alone, or use carpool. The dataset has for each individual one observation per option denoted by the variable '`option`'. The variable '`cost`' is the dollar cost of a trip, the variable '`time`' is the time cost of a trip. The binary variable '`choice`' equals 1 if an option was chosen, and is 0 otherwise.

   We start out by only using the information on the observed choice.

   (a) Interpret and discuss the coefficients in the logit model.

   (b) What is the partial effect at the sample average of increasing cost by 0.5$ on the probability of commuting alone by car? [Hint: $\Pr(y_i = 1|x = \bar{x}) \approx \Pr(y_i = 1)$]

   The plan is to introduce a 50% tax on driving alone. You are asked to estimate the effect of the tax on commuting choices. Let U_{ij}^* be the utility from choosing alternative j, and assume individual i chooses the alternative j that maximizes utility. Someone suggests you estimate a model with the following specification

   $$U_{ij}^* = \beta_j + \gamma cost_{ij} + \delta time_{ij} + \varepsilon_{ij} \quad j = bus, car, carpool, rail$$

   where ε_{ij} are unobserved taste shifters that follow a Type I extreme value distribution.

   (c) Is this a good choice given the data? what is the role of β_j?

   (d) How would you estimate this model in Stata and compute the policy impact?

2. A recent paper investigates the relationship between pollution and crime. The paper uses a daily panel dataset on wards (neighborhoods) in London and estimates the following equation

   $$\log crime_{it} = \beta AQI_{it} + f(Weather_{it}) + C_{it}\Pi + \mu_t + \gamma_i + \varepsilon_{it} \tag{1}$$

   where, $\log crime_{it}$ is the log of crime in ward i on day t and AQI_{it} is the corresponding air quality index with the following summary statistics

	Mean	Std.Dev.
$\log crime_{it}$	1.4	0.7
$crime_{it}$ (# per 100,000)	34.1	38.4
AQI_{it}	30.1	9.2

   Pollution is affected by weather on the one hand and human/economic activity on the other. The regression therefore also controls for weather conditions through mean temperature, relative humidity, as well as local measures of wind speed and rain. C_{it} is a vector of local area controls – metro (tube) activity, unemployment, and police deployment – accounting for time-varying conditions potentially related to

pollution and crime. μ_t and γ_i are time and ward fixed effects respectively. Finally, ε_{it} is an idiosyncratic error term.

Below you find Table 2 from the paper with OLS and fixed effects estimates.

(a) What is the main threat to validity when estimating (1) in your opinion? Motivate your answer.

(b) Interpret the OLS effect estimates of β in columns (1) and (2). Discuss economic and statistical significance, and the difference between these estimates.

(c) Column (3) adds ward fixed effects. Discuss the validity of the strict exogeneity assumption here, and interpret the estimate.

(d) The dependent variable in column (6) is the number of crimes per 100,000 inhabitants. Is the size of the effect here in line with the log specification in Column (5)? What can explain the difference?

3. A recent paper is interested in the causal effect of Dutch proficiency on the integration of migrants in the Netherlands.

They estimate the following outcome equation

$$y_i = \delta D_i + x_{1i}'\beta_1 + x_{2i}'\beta_2 + \varepsilon_i \tag{2}$$

where y_i is a binary variable which equals 1 if respondents report that they "feel Dutch" and is 0 otherwise. D_i is a self-reported binary measure of Dutch proficiency, and x_{1i}, x_{2i} are variable vectors that control for year, country of origin, reason for migration, homesick, naturalized, same-origin partner, and income.

(a) Give reasons why estimation of equation (2) with OLS will result in an inconsistent estimate of the causal effect of language proficiency on integration, and discuss the direction of the bias.

They implement an instrumental variables approach with the following first stage

$$D_i = z_i'\gamma + x_{1i}'\pi_1 + u_i \tag{3}$$

where z_i is a dummy variable that equals 1 if they have a satellite antenna at home, and is 0 otherwise.

(b) The paper uses z_i as an instrument because it "is associated with watching TV channels from homeland in mother tongue and points to a high intensity of mother tongue usage at home, possibly at the expense of host country language. This variable largely captures a high degree of home country orientation of immigrants and little inclination to learn host country language, assuming that media preferences reflect language preferences." Discuss instrument validity.

(c) Discuss whether the following controls strengthen the validity of the empirical strategy: country of origin, same-origin partner, and income.

(d) What is the Local Average Treatment Interpretation (LATE) interpretation of δ?

(e) Consider the simple case when there is one binary instrument z_i, $x_{1i} = 0$ and $x_{2i} = x_i$, that is:

$$y_i = \delta D_i + \beta x_i + \varepsilon_i \tag{4}$$

$$D_i = \gamma z_i + \pi x_i + u_i \tag{5}$$

and discuss what happens to the reduced form and first-stage estimates (and consequently the IV estimate of δ) when x_i is excluded from (5). What does this mean for the specification in the paper above?

```
Contains data from commute_long.dta
  obs:         1,812
 vars:             5
 size:        36,240
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              int     %9.0g
option          byte    %8.0g      option
choice          byte    %9.0g
cost            double  %10.0g
time            double  %10.0g
-------------------------------------------------------------------------------
Sorted by: id option



. ta option

     option |      Freq.     Percent        Cum.
------------+-----------------------------------
        bus |        453       25.00       25.00
        car |        453       25.00       50.00
    carpool |        453       25.00       75.00
       rail |        453       25.00      100.00
------------+-----------------------------------
      Total |      1,812      100.00


. ta option , s(choice)

            |          Summary of choice
     option |        Mean   Std. Dev.       Freq.
------------+------------------------------------
        bus |   .17880795   .38361507         453
        car |    .4812362    .5002002         453
    carpool |   .07064018   .25650611         453
       rail |   .26931567    .4440947         453
------------+------------------------------------
      Total |         .25   .43313224       1,812


. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |      1,812         227    130.8056          1        453
      option |      1,812         2.5    1.118343          1          4
      choice |      1,812         .25    .4331322          0          1
        cost |      1,812    2.701797    1.632214   .1293384   8.855542
        time |      1,812    39.06076    16.24341   1.968711   75.68108


. g car = option==2


. l in 1/9, sep(4) noobs

  +-----------------------------------------------------+
  | id    option   choice        cost        time   car |
  |-----------------------------------------------------|
  |  1       bus        0   1.8005123   20.867794     0 |
  |  1       car        1   1.5070097     18.5032     1 |
  |  1   carpool        0   2.3356118   26.338233     0 |
  |  1      rail        0   2.3589198   30.033469     0 |
  |-----------------------------------------------------|
  |  2       bus        0   2.2371285   67.181889     0 |
  |  2       car        0   6.0569985   31.311107     1 |
  |  2   carpool        0   2.8969191   34.256956     0 |
  |  2      rail        1   1.8554505   60.293126     0 |
  |-----------------------------------------------------|
  |  3       bus        0   2.5763845   63.309057     0 |
  +-----------------------------------------------------+


. logit car cost time if choice

Logistic regression                             Number of obs     =        453
                                                LR chi2(2)        =     328.26
                                                Prob > chi2       =     0.0000
Log likelihood =  -149.5456                     Pseudo R2         =     0.5232

------------------------------------------------------------------------------
         car |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cost |   2.472324   .2640377     9.36   0.000     1.954819    2.989828
        time |  -.1030401   .0146957    -7.01   0.000    -.1318432    -.074237
       _cons |  -3.420068   .4733536    -7.23   0.000    -4.347824   -2.492312
------------------------------------------------------------------------------
```

Table 1: Pooled OLS and Fixed Effect Models of Air Pollution's Impact on Crime

	Polled OLS		Fixed Effects			
	(1)	(2)	(3)	(4)	(5)	(6)
AQI (10 units)	0.064***	0.029***	0.020***	0.009***	0.009***	0.457***
	(0.0124)	(0.0101)	(0.0041)	(0.0025)	(0.0025)	(0.1134)
Controls	N	Y	Y	Y	Y	Y
Ward FE	N	N	Y	Y	Y	Y
DOW FE	N	N	N	Y	Y	Y
Year-Month FE	N	N	N	N	Y	Y
R-squared	0.007	0.060	0.372	0.383	0.385	0.688
Observations	419,210	398,437	398,437	398,437	398,437	433,277

Notes: Each column in the table represents a separate regression. In column (1)-(5), the dependent variable is the (log) number of criminal offences per day and ward and in column (6) the dependent variable is the crime rate per 100,000 people. AQI is based on air pollution readings from the three closest AURN monitoring stations (weighted by inverse squared distance). Control variables include weather characteristics (temperature, relative humidity and wind speed), ward-level police deployment and unemployment levels. Standard errors are cluster-robust in two dimensions, over wards and dates. * p<0.1, ** p<0.05, *** p<0.01.