UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Postponed exam: ECON3150/4150 – Introductory Econometrics

Date of exam: Wednesday, June 26, 2019

Time for exam: 09:00 a.m. – 12:00 noon

The problem set covers 5 pages (Incl. cover sheet)

Resources allowed:

- Open book examination where all printed and written resources, in addition to two alternative calculators are allowed.

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

This is an open book examination where all printed and written resources, in addition to a calculator, are allowed. If you are asked to derive something, give all intermediate steps. Do not answer questions with a "yes" or "no" only, but carefully motivate your answer. The percentage at the start of each question indicates how much it weigh in the grading.

1. (10%) Suppose we know that income y is distributed uniformly over an interval $\left[\underline{y}, \overline{y}\right]$ in a population. To estimate the mean income, μ, in the population we draw a random sample of n individuals and calculate the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

   (a) Why is $\overline{y}$ a random variable?

   (b) How is $\overline{y}$ distributed if n is very large?

   (c) Would your answer to b) change if y had a different (than uniform) distribution in the population.

2. (20%) There are two populations of individuals called *Samie* and *Varie* respectively. The height of individuals in the *Samie* population is given by: $X_i = \mu + \varepsilon_i$. The height of individuals in *Varie* is given by: $Y_i = \mu + \theta_i$. Where $\varepsilon_i \sim iid(0, \sigma_2)$ and $\theta_i \sim iid(0, 4\sigma_2)$.

   (a) Suppose we draw a random sample from the population of *Samie,* is the sample mean an unbiased estimator of μ?

   (b) Suppose we draw a random sample from the population of *Varie,* is the sample mean an unbiased estimator of μ?

   (c) Which of the two previous estimators is most efficient, explain your answer?

   (d) You have a sample of N individuals from each population. Is the average of the two sample means an unbiased estimator of the population mean?

   (e) Explain why the average of the two sample means is not the most efficient estimator of the population mean.

   (f) Construct the BLUE estimator.

3. (30%) Suppose you want to estimate the effect of physical exercise on educational attainment for university students. Let $edu_{i,t}$ be a measure of educational attainment (a combination of grades and courses completed) at the end of the academic year t of student i, and $exer_{i,t}$ is the average hours of physical exercise per week for person i in the academic year t.

   (a) Suppose you get a grant to run a controlled experiment to find the effect of physical exercise on educational attainment. Explain how you would design such an experiment to find the causal effect of $exer$ on edu.

   (b) Suppose, more realistically, you need to use data from a survey conducted on a random sample of students in year t. You estimate this equation

   $$edu_{i,t} = \beta_0 + \beta_1 exer_{i,t} + u_{i,t}.$$

   Give at least two factors that are likely to be contained in u and correlated with $exer$ and explain why this correlation would bias the estimate of how physical exercise affects educational attainment.

   (c) You get hold of a dataset that contains the same information for the same students also the year before (in $t-1$). Discuss how combining the data from period $t-1$ and t might, or might not, help to get a better estimate of how physical exercise affects educational attainment.

4. (40%) A researcher wants to investigate the effect of family size on the labour market participation of the children. She collects data from a random sample of persons who are 30 years old. She performs a regression of hours worked per week $Workhours$ on the variable $More2kids$ which equals 1 if a person grew up with at least two siblings (at least three children in the family) and zero if the person grew up with only one brother or sister, or with none at all (grew up as a only child). . The researcher estimates the following regression by OLS

   $$Workhours_i = \beta_0 + \beta_1 More2kids_i + u_i$$

   and obtains the following OLS estimates

```
regress Workhours More2kids, r

Linear regression                               Number of obs     =      2,000
                                                F (1, 1998)       =     426.06
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1760
                                                Root MSE          =      1.018

------------------------------------------------------------------------------
             |               Robust
   Workhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   More2kids |  -.9507726   .0460618
       _cons |   35.68208   .0299857  1189.97   0.000     35.62327    35.74088
------------------------------------------------------------------------------
```

a) Give an interpretation, in words, of the two estimated coefficients.

b) Test the null hypothesis that the coefficient on $More2kids_i$ is equal to zero using a 1 percent significance level.

c) Describe one potential threat to the internal validity of the current regression results.

d) The researcher thinks she can obtain a consistent estimate of the effect of family size on labour supply of children by using 2SLS. She decides to use the incidence of twins at the second birth as an instrument for family size. She estimates the following first stage regression

$$More2kids_i = \pi_0 + \pi_1 Twins2_i + v_i$$

where $Twins2$ equals one if the mother had twins at second birth and zero otherwise. She obtains the following OLS estimates.

```
regress More2kids Twins2, r

Linear regression                               Number of obs     =      2,000
                                                F(0, 1998)        =          .
                                                Prob > F          =          .
                                                R-squared         =     0.0700
                                                Root MSE          =     .47718

------------------------------------------------------------------------------
             |               Robust
   More2kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Twins2 |   .6033666   .0112256    53.75   0.000     .5813514    .6253819
       _cons |   .3966334   .0112256    35.33   0.000     .3746181    .4186486
------------------------------------------------------------------------------
```

Do you think that the instrument relevance condition holds? Is $Twins2_i$ a weak instrument?

e) Do you think that $Twins2_i$ satisfies the instrument exogeneity condition? Explain why or why not.

f) The following table shows the averages of $Workhours_i$ and $More2kids_i$ for the individuals whose mother had twins at second birth ($Twins2_i = 1$) and for the individuals whose mother did not have twins at second birth ($Twins2_i = 0$). Use the results in the table below to obtain the instrumental variable estimate of the effect of $More2kids_i$ on years of hours worked per week ($Workhours_i$).

	$Twins2_i = 1$	$Twins2_i = 0$
$\widehat{E}\left[Workhours_i \mid Twins2_i = x\right]$	35.00	35.30
$\widehat{E}\left[More2kids_i \mid Twins2_i = x\right]$	1.00	0.40