UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Postponed exam: **ECON3150/4150 – Introductory Econometrics**

Date of exam: Wednesday, June 7, 2017

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

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

Resources allowed:

- Open-book exam, where all written and printed resources – as well as calculator - are allowed

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

Exam ECON3150/4150: Introductory Econometrics.
7 June 2017; 09.00h-12.00h.

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.

Question 1

The government of a country wants to investigate whether school size affects schooling outcomes. A government official has a data set with test scores of 25 000 students. The variable $passed_i$ equals one if a student passed the exam at the end of secondary education and zero otherwise and the variable $school\ size_i$ equals the number of students in the school of student i.

a) The government official decides to estimate the following regression model by OLS

$$passed_i = \beta_0 + \beta_1 \cdot ln(school\ size_i) + u_i \tag{1}$$

and obtains the following estimation results

```
. regress passed ln_school_size, robust

Linear regression                               Number of obs     =         25,000
                                                F(1, 24998)       =         443.87
                                                Prob > F          =         0.0000
                                                R-squared         =         0.0169
                                                Root MSE          =         .42936
```

passed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_school_size	-.7994753	.037947	-21.07	0.000	-.8738537	-.725097
_cons	5.01307	.2019471	24.82	0.000	4.617242	5.408898

Give an interpretation, in words, of the estimated coefficient on $ln(school\ size_i)$.

b) Compute a 90 percent confidence interval for $\widehat{\beta}_1$.

c) The government official decides to estimate a probit model and includes $school\ size_i$ instead of $ln(school\ size_i)$ as explanatory variable. She obtains the following estimation results

```
. probit passed school_size, robust

Iteration 0:   log pseudolikelihood =   -14058.379
Iteration 1:   log pseudolikelihood =    -13844.51
Iteration 2:   log pseudolikelihood =   -13844.282
Iteration 3:   log pseudolikelihood =   -13844.282

Probit regression                                 Number of obs     =         25,000
                                                  Wald chi2(    1)      =      429.09
                                                  Prob > chi2       =      0.0000
Log pseudolikelihood =  -13844.282                  Pseudo R2         =      0.0152
```

passed	Coef.	Robust Std. Err.	[redacted]	[redacted]	[redacted]	[redacted]
school_size	-.0123639	.0005969	[redacted]	[redacted]	[redacted]	[redacted]
_cons	3.250046	.1249073	26.02	0.000	3.005233	3.49486

Is the coefficient on $school\ size_i$ significantly different from zero at a 5 percent significance level?

d) Using the results from the probit model, what is the predicted change in the probability of passing the exam that is associated with an increase in school size from 200 to 220 students?

e) Instead of looking at whether a student passed the exam the government official decides to use the logarithm of the test score as dependent variable. She estimates the following regression model by OLS

$$ln(testscore_i) = \beta_0 + \beta_1 \cdot school\ size_i + u_i \tag{2}$$

and obtains the following estimation results

```
. regress ln_testscore school_size, robust

Linear regression                               Number of obs     =         25,000
                                                  F(1, 24998)       =      49558.31
                                                  Prob > F          =      0.0000
                                                  R-squared         =      0.6680
                                                  Root MSE          =      .03651
```

ln_testscore	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
school_size	-.003557	.000016	-222.62	0.000	-.0035883	-.0035257
_cons	4.094445	.0032974	1241.72	0.000	4.087982	4.100908

Give an interpretation, in words, of the estimated coefficient on school size.

f) The government official thinks that there might be a quadratic relation between school size and the logarithm of test scores and she estimates the following equation by OLS

$$ln(testscore_i) = \beta_0 + \beta_1 \cdot school\ size_i + \beta_2 \left(school\ size_i\right)^2 + \varepsilon_i \tag{3}$$

and obtains the following estimation results

```
. regress ln_testscore school_size school_size_2, robust

Linear regression                               Number of obs     =         25,000
                                                F(2, 24997)       =      24797.39
                                                Prob > F          =        0.0000
                                                R-squared         =        0.6680
                                                Root MSE          =        .03651
```

ln_testscore	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
school_size	-.0036916	.0000956	-38.60	0.000	-.0038791	-.0035041
school_size_2	3.24e-07	2.28e-07	1.42	0.155	-1.23e-07	7.70e-07
_cons	4.108367	.0102648	400.24	0.000	4.088247	4.128487

Is the model in equation (3) better than the model in equation (2)? Explain why or why not.

g) Name and explain one threat to internal validity that might apply when estimating equation (2) by OLS.

h) The government decides to set up an experiment and randomly assigns municipalities to a treatment group and to a control group. Schools that are located in a municipality that belongs to the treatment group have to merge with another school in that municipality. The variable $treated_i$ equals 1 when a student lives in a municipality that was assigned to the treatment group en 0 if it was assigned to the control group. The government official decides to use the variable $treated_i$ as instrument for $schoolsize_i$. She obtains the following first stage estimation results.

```
. regress school_size treated, robust

Linear regression                               Number of obs     =         25,000
                                                F(1, 24998)       =      28052.43
                                                Prob > F          =        0.0000
                                                R-squared         =        0.5288
                                                Root MSE          =        9.9941
```

school_size	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treated	21.1735	.1264176	167.49	0.000	20.92571	21.42129
_cons	196.8927	.0892192	2206.84	0.000	196.7178	197.0676

Is $treated_i$ a weak instrument?

i) Do you think that, when using $treated_i$ as an instrument to estimate the effect of $school\ size_i$ on $ln(testscore_i)$, the instrument exogeneity condition holds ? Explain why or why not.

j) The researcher estimates the following two regressions by OLS

$$ln(testscore_i) = \delta_0 + \delta_1 treated_i + \epsilon_i$$

$$school\ size_i = \pi_0 + \pi_1 treated_i + \upsilon_i$$

and obtains the following estimation results.

```
1 . regress ln_testscore treated, robust noheader
```

ln_testscore	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treated	-.0737751	.0006519	-113.18	0.000	-.0750528	-.0724974
_cons	3.393329	.0004452	7622.87	0.000	3.392457	3.394202

```
2 . regress school_size treated, robust noheader
```

school_size	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treated	21.1735	.1264176	167.49	0.000	20.92571	21.42129
_cons	196.8927	.0892192	2206.84	0.000	196.7178	197.0676

What is the instrumental variable estimate of the effect of $school\ size_i$ on $ln(testscore_i)$?

Question 2

Discuss whether each of the following statements is correct or not.

a) In case of perfect multicollinearity the OLS estimator is biased.

b) In case of imperfect multicollinearity the OLS estimator is biased.

c) A confidence interval always contains the true value of the population parameter.

d) In a panel data model with entity fixed effects you can't estimate the effect of time-invariant characteristics.

Question 3

Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ with $Cov(X_i, u_i) = 0$. A researcher wants to estimate β_1 using survey data. It turns out that individuals in the survey systematically under-reported X_i by 50 percent. The researcher therefore has a large data set with i.i.d observations on Y_i and X_i^*, with $X_i^* = 0.5X_i$. He estimates the following equation by OLS

$$Y_i = \beta_0 + \beta_1 X_i^* + v_i$$

a) What is $Cov(X_i^*, v_i)$?

b) Is the OLS estimator of β_1 consistent?