Exam ECON3150/4150: Introductory Econometrics.
18 May 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

A researcher wants to investigate whether class size affects long term labour market outcomes. He has a data set with labour market outcomes of 100 000 30-year-old individuals that live in Norway. The variable $employed_i$ equals one if an individual is employed and zero otherwise and the variable $class\ size_i$ equals the number of students in the class of the individual when he/she was in school.

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

$$employed_i = \beta_0 + \beta_1 \cdot class\ size_i + u_i \tag{1}$$

and obtains the following estimation results

```
. regress employed class_size, robust

Linear regression                               Number of obs     =        100,000
                                                F(1, 99998)       =        1201.81
                                                Prob > F          =         0.0000
                                                R-squared         =         0.0119
                                                Root MSE          =         .29821
```

employed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.0158525	.0004573	-34.67	0.000	-.0167488	-.0149562
_cons	1.264691	.0102837	122.98	0.000	1.244535	1.284847

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

Solution (10 points): $\widehat{\beta}_1 = -0.016$. *If class size increases with 1 student this is associated with a decrease in the probability of employment at age 30 by on average 0.016 (1.6 percentage points).*

b) Compute a 99 percent confidence interval for $\widehat{\beta}_0$.

Solution (10 points): *99% confidence interval for* $\widehat{\beta}_0$*:*

$$\widehat{\beta}_0 \pm 2.58 \times SE(\widehat{\beta}_0)$$

filling in the numbers from the regression output gives

$$1.265 \pm 2.58 \times 0.010$$

$$(1.239 \quad , \quad 1.291)$$

c) The researcher decides to estimate a logit model and obtains the following estimation results

```
. logit employed class_size, robust

Iteration 0:   log pseudolikelihood =   -32508.297
Iteration 1:   log pseudolikelihood =   -31919.121
Iteration 2:   log pseudolikelihood =   -31906.557
Iteration 3:   log pseudolikelihood =   -31906.551
Iteration 4:   log pseudolikelihood =   -31906.551  (backed up)

Logistic regression                               Number of obs     =        100,000
                                                  Wald chi2(    1)      =     1220.95
                                                  Prob > chi2       =        0.0000
Log pseudolikelihood =  -31906.551                  Pseudo R2         =        0.0185

------------------------------------------------------------------------------
             |               Robust
    employed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -.1794293   .0051351   -34.94   0.000    -.1894939   -.1693647
       _cons |   6.379385    .121634    52.45   0.000     6.140987    6.617783
------------------------------------------------------------------------------
```

Is the coefficient on $class\,size_i$ significantly different from zero at a 10 percent significance level?

Solution (10 points): $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$. *Construct the t-statistic:*

$$t = \frac{-0.1794293 - 0}{0.0051351} = -34.94$$

The absolute value of the t-statistic is bigger than 1.64 so we reject H_0*. The the coefficient on* $class\,size_i$ *is significantly different from zero at a 10 percent significance level.*

d) Using the results from the logit model, what is the predicted change in the probability of being employed at age 30 that is associated with a reduction in class size from 25 to 20 students?

Solution (10 points): *The predicted change in the probability of being employed at age 30 that is associated with a reduction in class size from 25 to 20 students equals:*

$$\triangle Pr(\widehat{employed}_i = 1) = \Pr(\widehat{employed}_i = 1 | class\ size_i = 20) - \Pr(\widehat{employed}_i = 1 | class\ size_i = 25)$$

$$\begin{aligned} \triangle Pr(\widehat{Employed}_i = 1) &= \left(1/\left(1 + e^{-(6.379 - 0.179 \cdot 20)}\right)\right) - \left(1/\left(1 + e^{-(6.379 - 0.179 \cdot 25)}\right)\right) \\ &= 0.943 - 0.870 \\ &= 0.073 \end{aligned}$$

e) The data set also contains information about yearly income and the researcher decides to estimate the following regression model by OLS

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

and obtains the following estimation results

```
. regress ln_income class_size, robust

Linear regression                               Number of obs     =        100,000
                                                F(1, 99998)       =      15530.38
                                                Prob > F          =         0.0000
                                                R-squared         =         0.1236
                                                Root MSE          =         .05719
```

ln_income	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.0103901	.0000834	-124.62	0.000	-.0105535	-.0102267
_cons	3.181756	.0019192	1657.88	0.000	3.177995	3.185518

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

Solution (10 points): *Equation 2 is a log-linear model, we can therefore interpret the coefficient on class size (approximately) as follows: if class size increases by 1 student this is associated with a reduction in yearly income by about 1 percent (0.01*100).*

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

Solution (10 points): *The most obvious one is omitted variable bias. There might be variables that affect income and that are related to class size. Individuals that were in a big class when they were in school might differ in characteristics from students that were in a small class. It might also be that schools in big cities have bigger classes than schools in small towns and that earnings are higher in bigger cities than in small towns.*

g) The researcher decides to include the years of completed education of both the father ($edu\ father_i$) and the mother ($edu\ mother_i$) as explanatory variables. He estimates the following equation by OLS

$$ln(income_i) = \beta_0 + \beta_1 \cdot class\ size_i + \beta_2 \cdot edu\ father_i + \beta_3 \cdot edu\ mother_i + \varepsilon_i \qquad (3)$$

and obtains the following estimation results

```
. regress ln_income class_size edu_father edu_mother, robust

Linear regression                               Number of obs     =        100,000
                                                F(3, 99996)       =      10633.90
                                                [redacted]        =      [redacted]
                                                R-squared         =         0.2417
                                                Root MSE          =         .05319
```

ln_income	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.0007064	.0001123	-6.29	0.000	-.0009265	-.0004863
edu_father	.0098685	.000473	20.87	0.000	.0089415	.0107955
edu_mother	.0153749	.0003337	46.07	0.000	.0147208	.0160289
_cons	3.008313	.0023325	1289.75	0.000	3.003741	3.012884

Test the null hypothesis that the coefficient on class size and the coefficients on the years of completed education of both the father and the mother are equal to zero at a 5 percent significance level.

Solution (10 points): $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ *vs* $H_1 : \beta_1 \neq 0$ *and/or* $\beta_2 \neq 0$ *and/or* $\beta_3 \neq 0$. *The F-statistic is given in the Stata output; F=10633.9. There are 3 restrictions under the null hypothesis and we can use the large sample approximation because n=100000. This gives the following critical value of the F-statistic:* $F^{5\%}_{3,\infty} = 2.6$. *Since 10633.9>2.6 we reject* H_0 *at a 5 percent significance level.*

h) The researcher wants to analyze whether the effect of class size depends on the education of the parents. Describe in detail how you can test the null hypothesis that the effect of class size does not depend on the education of the mother and the father.

Solution (10 points): *The regression should be augmented to include two interaction terms as follows:*

$$\begin{aligned} ln(income_i) = \; & \beta_0 + \beta_1 \cdot class\ size_i + \beta_2 \cdot edu\ father_i + \beta_3 \cdot edu\ mother_i \\ & + \beta_4 \cdot (edu\ father_i \times class\ size_i) \\ & + \beta_5 \cdot (edu\ mother_i \times class\ size_i) + \epsilon_i \end{aligned}$$

The hypothesis can be tested by using an F test testing H_0: $\beta_4 = \beta_5 = 0$.

i) The researcher decides to use an instrumental variable approach. He thinks that class size is on average higher in Oslo than in the rest of Norway and therefore use the dummy variable $Oslo_i$ as instrument for $class\ size_i$. $Oslo_i$ equals 1 when an individual lived in Oslo when he/she was in school and zero otherwise. The researcher obtains the following estimation results.

```
. regress class_size Oslo, robust

Linear regression                               Number of obs     =        100,000
                                                F(1, 99998)       =      97275.24
                                                Prob > F          =         0.0000
                                                R-squared         =         0.4943
                                                Root MSE          =         1.4695

------------------------------------------------------------------------------
             |               Robust
  class_size |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Oslo |    1.006355   .0032266   311.89   0.000     1.000031    1.012679
       _cons |    20.48096   .0093368  2193.57   0.000     20.46266    20.49926
------------------------------------------------------------------------------
```

Is $Oslo_i$ a weak instrument?

Solution (10 points): *The first stage F-statistic equals $F = (t)^2 = (311.89)^2 = 97275$ (can also use overall-regression F-statistic since regression includes only 1 regressor), which is bigger than the rule-of-thumb value of 10. The instrument $Oslo_i$ is therefore not a weak instrument.*

j) Do you think that, when using $Oslo_i$ as an instrument to estimate the effect of $class\ size_i$ on $ln(income_i)$, the instrument exogeneity condition holds ? Explain why or why not.

Solution (10 points): *Instrument exogeneity: $Cov(Oslo_i, u_i) = 0$. A possible reason why the instrument exogeneity condition might be violated is that the labour market in Oslo is different from that of the rest of Norway and that income in Oslo might be therefore higher on average regardless of any differences in class size.*

k) The researcher estimates the following two equations by OLS

$$ln(income_i) = \delta_0 + \delta_1 Oslo_i + \epsilon_i$$

$$class\ size_i = \pi_0 + \pi_1 Oslo_i + \upsilon_i$$

and obtains the following estimation results.

```
1 . regress ln_income Oslo, robust noheader
```

ln_income	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Oslo	-.0002051	.0001339	-1.53	0.126	-.0004675	.0000573
_cons	2.943244	.0003881	7583.53	0.000	2.942483	2.944004

```
2 . regress class_size Oslo, robust noheader
```

class_size	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Oslo	1.006355	.0032266	311.89	0.000	1.000031	1.012679
_cons	20.48096	.0093368	2193.57	0.000	20.46266	20.49926

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

Solution (10 points): There is an alternative way of computing the instrumental variable estimator:

$$\begin{aligned}
\hat{\beta}_{IV} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Z_i-\bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Z_i-\bar{Z})/\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})/\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2} \\
&= \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2}
\end{aligned}$$

- $\frac{S_{ZY}}{S_Z^2}$ is the OLS estimator when regressing Y_i on Z_i
- $\frac{S_{ZX}}{S_Z^2}$ is the OLS estimator when regressing X_i on Z_i

This implies that the IV estimator of the effect of $class\ size_i$ on $ln(income_i)$ equals

$$\widehat{\beta}_{IV} = \frac{\widehat{\delta}_1}{\widehat{\pi}_1} = \frac{-0.0002}{1.0064} = -0.0002$$

Question 2

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

a) If the error terms are homoskedastic, hypothesis tests that are based on heteroskedasticity-robust standard errors are invalid.

Solution (5 points) *Incorrect. Homoskedasticity is a special case of heteroskedasticity and hypothesis tests that are based on heteroskedasticity-robust standard errors are valid whether or not the errors terms are heteroskedastic.*

b) An estimator is consistent if the expected value of the estimator equals the true value of the population parameter.

Solution (5 points) *Incorrect. An estimator is consistent if it converges in probability to the true value of the population parameter that it is estimating when $n \to \infty$. An estimator is unbiased if the expected value of the estimator equals the true value of the population parameter.*

c) The p-value is the smallest significance level at which the null hypothesis can be rejected.

Solution (5 points) *Correct. The p-value is the probability of drawing a statistic at least as adverse to the null hypothesis as the one actually computed, assuming the null hypothesis is correct. The significance level is the prescribed rejection probability of a statistical hypothesis test when the null hypothesis is true. The p-value is therefore the smallest significance level at which the null hypothesis can be rejected.*

d) In a regression model with a only a constant term and no explanatory variables the R^2 equals zero.

Solution (5 points) *Correct. The R^2 is the fraction of the sample variance of Y_i explained/predicted by the explanatory variables ($R^2 = \frac{ESS}{TSS}$). When there are no explanatory variables this fraction is zero (the explained sum of squares $EES = 0$)*

Question 3

A researcher wants to estimate the effect of a job training program (T_i) on wages (W_i).

$$W_i = \beta_0 + \beta_1 T_i + u_i$$

A colleague of the researcher thinks that the wage affects participation the job training program:

$$T_i = \delta_0 + \delta_1 W_i + v_i$$

If the colleague is right, will the researcher obtain a consistent estimate of the causal effect of the job training program (T_i) on wages (W_i) when he estimates the following regression model $W_i = \beta_0 + \beta_1 T_i + u_i$ by OLS? Show why or why not (Hint: first derive $Cov\left(T_i, u_i\right)$ and assume that $Cov\left(u_i, v_i\right) = 0$).

Solution (20 points):

$$\begin{aligned} Cov(T_i, u_i) &= Cov(\delta_0 + \delta_1 W_i + v_i,\ u_i) \\ &\qquad\qquad\qquad\qquad assuming\, Cov(v_i, u_i) = 0 \\ &= Cov(\delta_1 W_i,\ u_i) \\ &\qquad\qquad\qquad\qquad substitute\, for\, W_i \\ &= Cov(\delta_1 \left(\beta_0 + \beta_1 T_i + u_i\right), u_i) \\ &= \delta_1 \beta_1 Cov(T_i, u_i) + \delta_1 Var(u_i) \end{aligned}$$

Solving for $Cov(T_i, u_i)$ gives

$$Cov(T_i, u_i) = \frac{\delta_1}{1 - \delta_1 \beta_1} Var(u_i)$$

$$\begin{aligned} \widehat{\beta_1} = \tfrac{s_{WT}}{s_T^2} \xrightarrow{p} \tfrac{Cov(W_i, T_i)}{Var(T_i)} &= \tfrac{Cov(\beta_0 + \beta_1 T_i + u_i, T_i)}{Var(T_i)} \\ &= \beta_1 + \tfrac{Cov(u_i, T_i)}{Var(T_i)} \\ &= \beta_1 + \tfrac{\frac{\delta_1}{1 - \delta_1 \beta_1} Var(u_i)}{Var(T_i)} \neq \beta_1 \end{aligned}$$

If the colleague is right, the researcher will not obtain a consistent estimate of the causal effect of the job training program (T_i) on wages (W_i) when he estimates the following regression model $W_i = \beta_0 + \beta_1 T_i + u_i$ by OLS.