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

--------------------------------------------------------------------------------
                |               Robust
         passed |      Coef.   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 $ln(school\ size_i)$.

Solution (10 points): $\widehat{\beta}_1 = -0.799$. *This is a linear (probability) log model. If school size increases with 1% this is associated with a decrease in the probability of passing the exam by on average 0.01 *0.799≈0.008 (0.8 percentage points).*

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

Solution (10 points): *90% confidence interval for* $\widehat{\beta}_1$*:*

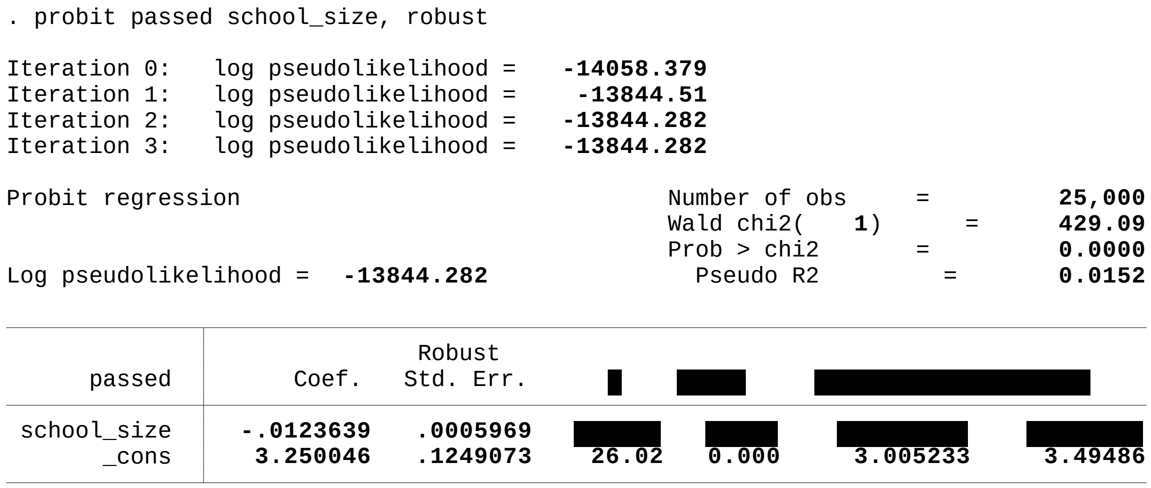
$$\widehat{\beta}_1 \pm 1.64 \times SE(\widehat{\beta}_1)$$

filling in the numbers from the regression output gives

$$-0.799 \pm 1.64 \times 0.038$$

$$(-0.861 \quad , \quad -0.737)$$

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

------------------------------------------------------------------------------------
                |               Robust
         passed |      Coef.   Std. Err.   [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?

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

$$t = \frac{-0.0123639 - 0}{0.0005969} = -20.71$$

The absolute value of the t-statistic is bigger than 1.96 so we reject H_0*. The the coefficient on school size*$_i$ *is 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?

Solution (10 points): *The predicted change in the probability of passing the exam that is associated with an increase in school size from 200 to 220 students equals:*

$$\triangle Pr(\widehat{passed}_i = 1) = \ \Pr(passed_i = 1|\widehat{school}\ size_i = 220)\ -\ \Pr(passed_i = 1|\widehat{school}\ size_i = 200)$$

$$\begin{aligned}\triangle Pr(\widehat{Passed}_i = 1) \ &= \Phi\,(3.25 - 0.012 \cdot 220) - \Phi\,(3.25 - 0.012 \cdot 200)\\ &= \Phi\,(0.61) - \Phi\,(0.85)\\ &= 0.729 - 0.802\\ &= -0.073\end{aligned}$$

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

------------------------------------------------------------------------------
             |               Robust
ln_testscore |      Coef.   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.

Solution (10 points): *Equation 2 is a log-linear model, we can therefore interpret the coefficient on school size (approximately) as follows: if school size increases by 1 student this is associated with, on average, a reduction in test scores by about 0.36 percent (0.0036*100).*

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

------------------------------------------------------------------------------
             |               Robust
ln_testscore |      Coef.   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.

Solution (10 points): *We can test the linear against the quadratic model by performing the following hypothesis test:* $H_0 : \beta_2 = 0$ *vs* $H_1 : \beta_2 \neq 0$.

$$t = 1.42$$

The absolute value of the t-statistic is smaller than 1.64 so we do not reject H_0 *at a 10% (nor 5% or 1%) level. This means that we do not reject the linear model in favor of the quadratic model and the model in equation 3 does not seem to be better than the model in equation 2.*

g) 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 omitted variables that affect test scores and that are related to school size. For example students that are enrolled in big schools might come from different backgrounds than students enrolled in smaller schools. There might also be reversed causality if schools with high average test scores attract more students.*

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 $school\,size_i$. 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

------------------------------------------------------------------------------
             |               Robust
 school_size |      Coef.   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?

Solution (10 points): *The first stage F-statistic equals* $F = (t)^2 = (167.49)^2 = 28052$ *(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* $treated_i$ *is therefore not 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.

Solution (10 points): *Instrument exogeneity:* $Cov(treated_i, u_i) = 0$. *Municipalities are randomly assigned to a treatment or control group and characteristics of students and schools that are contained in the error term* u_i *should therefore be uncorrelated with the variable* $treated_i$. *The instrument exogeneity condition might however be violated if there is a direct effect of* $treated_i$ *on test scores, for example if schools merge this might lead to a period of disruptions and changes for the students enrolled in these schools which might affect their test scores.*

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
   ------------------------------------------------------------------------------
               |               Robust
  ln_testscore |      Coef.   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
   ------------------------------------------------------------------------------
               |               Robust
   school_size |      Coef.   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$_i$)*?

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 *school size$_i$* on *ln(testscore$_i$)* equals

$$\widehat{\beta}_{IV} = \frac{\widehat{\delta}_1}{\widehat{\pi}_1} = \frac{-0.074}{21.174} = -0.0035$$

Question 2

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

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

Solution (5 points) *Incorrect. Perfect multicollinearity is a situation in which one of the regressors is an exact linear function of the other regressors. In this situation the OLS estimator is not biased, but it is impossible to compute the OLS estimator.*

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

Solution (5 points) *Incorrect. Imperfect multicollinearity is a situation in which two or more regressors are highly (but not perfectly) correlated. This does not violate any of the OLS assumptions and therefore does not lead to biased OLS estimators. Instead the coefficient on at least one of the regressors will be imprecisely estimated.*

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

Solution (5 points) *Incorrect. A confidence interval is an interval that contains the true value of a population parameter with a prespecified probability when computed over repeated samples.*

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

Solution (5 points) *Correct. Let X_i be a time-invariant characteristic. When the following model is estimated*

$$Y_{it} = \beta_1 X_i + \alpha_i + u_{it}$$

with α_i an individual fixed effect, X_i and α_i are perfectly multicollinear and it is impossible to estimate β_1.

Question 3

Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ with $Cov\left(X_i, u_i\right) = 0$. 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 $Cov\left(X_i^*, v_i\right)$?

Solution (10 points)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 \left(2X_i^*\right) + u_i \\ &= \beta_0 + \beta_1 X_i^* + \left(u_i + \beta_1 X_i^*\right) \\ &= \beta_0 + \beta_1 X_i^* + v_i \end{aligned}$$

This implies that $v_i = u_i + \beta_1 X_i^*$

$$\begin{aligned} Cov\left(X_i^*, v_i\right) &= Cov\left(X_i^*,\ u_i + \beta_1 X_i^*\right) \\ &= Cov(0.5X_i, u_i) + Cov\left(0.5X_i, \beta_1 0.5X_i\right) \\ &= 0 + 0.5 \cdot 0.5 \cdot \beta_1 Var\left(X_i\right) \\ &= 0.25\beta_1 Var\left(X_i\right) \end{aligned}$$

b) Is the OLS estimator of β_1 consistent?

Solution (10 points):

$$\widehat{\beta}_1 = \frac{s_{X^*Y}}{s^2_{X^*}} \xrightarrow{p} \frac{Cov(X_i^*,Y_i)}{Var(X_i^*)}$$

$$\begin{aligned}
\widehat{\beta}_1 \xrightarrow{p} \frac{Cov(X_i^*,Y_i)}{Var(X_i^*)} &= \frac{Cov(X_i^*,\beta_0+\beta_1X_i^*+v_i)}{Var(X_i^*)} \\
&= \frac{Cov(X_i^*,\beta_0)+\beta_1Cov(X_i^*,X_i^*)+Cov(X_i^*,v_i)}{Var(X_i^*)} \\
&= \beta_1 + \frac{Cov(X_i^*,v_i)}{Var(X_i^*)} \\
&= \beta_1 + \frac{0.25\beta_1Var(X_i)}{Var(0.5X_i)} \\
&= \beta_1 + \frac{0.25\beta_1Var(X_i)}{0.25Var(X_i)} \\
&= \beta_1 + \beta_1
\end{aligned}$$

This means that the OLS estimator of β_1 is inconsistent because it converges to $2 \cdot \beta_1$ so it overestimates the true causal effect by 100%.