Postponed Exam ECON3150/4150: Introductory Econometrics.
2 June 2016; 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.

> **Guideline for correctors:** *In this exam a total of 120 points can be obtained. The number of points that can be obtained by answering a question correctly are indicated in the solution box below the question.*

Question 1

The government of a developing country wants to know whether completing primary school has a positive effect on future income. A government employee has a data set with income ($income_i$) in US dollars of 1000 individuals living in this developing country in 2012. The data set also contains a variable $primary_i$ that equals 1 if individual i completed primary school and zero if individual i did not complete primary school. The government employee estimates the following equation by OLS

$$ln(income_i) = \beta_0 + \beta_1 primary_i + u_i$$

and obtains the following estimation results:

```
. regress ln_income primary, robust

Linear regression                               Number of obs     =      1,000
                                                   F(1, 998)         =     952.76
                                                   Prob > F          =     0.0000
                                                   R-squared         =     0.4838
                                                   Root MSE          =     .22116

------------------------------------------------------------------------------
             |               Robust
   ln_income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primary |    .439359   .0142341    30.87   0.000     .4114268    .4672911
       _cons |   5.255183   .0090622   579.90   0.000       5.2374    5.272966
------------------------------------------------------------------------------
```

a) Interpret the estimated coefficient on $primary_i$.

Solution: *(5 points). The estimated coefficient on* $primary_i$ *equals 0.44. It is a log-linear model which implies that we can interpret the estimated coefficient as follows: completing primary school is associated with an increase in income by about (100**$\widehat{\beta}_{primary}$*=) 44% .*

b) Construct a 90 percent confidence interval for the coefficient on $primary_i$.

Solution: (*5 points*) *90% confidence interval for* $\widehat{\beta}_1$*:*

$$\widehat{\beta_1} \pm 1.65 \times SE(\widehat{\beta}_1)$$

filling in the numbers from the regression output gives

$$0.439 \pm 1.65 \times 0.014$$

$$(0.416 \quad , \quad 0.462)$$

c) Explain whether we can interpret the estimated OLS coefficient on $primary_i$ as the causal effect of completing primary school on future income.

Solution: *(10 points) We can only interpret the OLS coefficient on* $primary_i$ *as the causal effect of completing primary school on future income if the OLS assumptions hold, in particular* $E\left[u_i|primary_i\right] = 0$ *should hold. Completing primary school is however likely related to unobserved characteristics that affect future income. For example children from parents with higher income might be more likely to complete primary school and children from parents with higher income might have higher future income regardless of whether they complete primary school. This implies that there are likely important omitted variables and that* $E\left[u_i|primary_i\right] \neq 0$.

d) The researcher decides to estimate the effect of completing primary school using an instrumental variable approach. He has information on where each individual lived at the age of 10 and computes the distance to the nearest primary school. He uses the variable Z_i as instrumental variable which equals 1 if the primary school was less than 5 kilometers away and zero if the nearest primary school was more than 5 kilometers away. He obtains the following first stage estimation results.

```
. regress primary Z, robust

Linear regression                               Number of obs     =      1,000
                                                [redacted]




------------------------------------------------------------------------------
             |               Robust
     primary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |       .228   .0299638  [redacted] 0.000     .1692008    .2867992
       _cons |       .272   .0199205    13.65   0.000     .2329092    .3110908
------------------------------------------------------------------------------
```

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

Solution: *(10 points) Instrument relevance,* $Corr(primary_i, Z_i) \neq 0$ *can be investigated using the first stage regression. The first stage F-statistic equals* $F = (t)^2 = \left(\frac{\widehat{\pi}_Z}{SE(\widehat{\pi}_Z)}\right)^2 = \left(\frac{0.228}{0.030}\right)^2 = (7.6)^2 = 57.76$, *which is larger than the rule-of-thumb value of 10 so the instrument relevance condition holds and* Z *is not a weak instrument.*

e) Do you think that the instrument exogeneity condition holds?

Solution: *(10 points) The instrument should be uncorrelated with the error term,* $Cov(Z_i, u_i) = 0$. *This assumption can't be tested. The assumption is violated if individuals that lived close to a primary school differ in characteristics from those that live far from a primary school and if these characteristics affect future earnings. For example individuals that live in a big city live close to a primary school but in general there are also better job opportunities in a big city compared to a remote rural area. There are therefore reasons to suspect that the instrument exogeneity condition does not hold.*

f) The following table shows the averages of $ln(income_i)$ and $primary_i$ for those who lived less than 5 kilometers from the nearest primary school ($Z_i = 1$) and for those who lived more than 5 kilometers from the nearest primary school ($Z_i = 0$). Use the results in the table below to obtain the instrumental variable estimate of the effect of completing primary school on future income and interpret the magnitude of this instrumental variable estimate.

	$Z_i = 1$	$Z_i = 0$
$\widehat{E}\left[ln(income_i)\|Z_i = x\right]$	5.446	5.403
$\widehat{E}\left[primary_i\|Z_i = x\right]$	0.50	0.27

Solution: *(10 points) The instrument Z_i is binary, We therefore have that the IV estimator equals the so called Wald estimator:*

$$\hat{\beta}_{IV} = \frac{\widehat{E}\left[ln(income_i)|Z_i = 1\right] - \widehat{E}\left[ln(income_i)|Z_i = 0\right]}{\widehat{E}\left[primary_i|Z_i = 1\right] - \widehat{E}\left[primary_i|Z_i = 0\right]}$$

the instrumental variable estimate of the effect of completing primary school on future income equals:

$$\hat{\beta}_{IV} = \frac{5.446 - 5.403}{0.5 - 0.27} = 0.18$$

this can be interpreted as that completing primary school increases income by about 18%

Question 2

A teacher wants to know whether study time affects the probability of passing an exam. She has a data set with 500 students that contains a variable $passed_{it}$ that equals 1 if a student passed the exam that took place at time t and a variable $studytime_{it}$ that contains the number of hours that student i spent on preparing for the exam that was taken at time t. The data set contains in total 5000 observations, with 500 students ($n = 500$) that each took 10 different exams ($T = 10$). The teacher estimates

$$passed_{it} = \beta_0 + \beta_1 studytime_{it} + u_{it}$$

by OLS and obtains the following estimation results.

```
. regress passed studytime, robust

Linear regression                               Number of obs     =      5,000
                                                   F(1, 4998)        =    2121.16
                                                   Prob > F          =     0.0000
                                                   R-squared         =     0.2854
                                                   Root MSE          =     .36611

------------------------------------------------------------------------------
             |               Robust
      passed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   studytime |   .0804765   .0017474  [redacted] [redacted] [redacted] [redacted]
       _cons |   .3448519   .0127398    27.07   0.000     .3198762    .3698276
------------------------------------------------------------------------------
```

a) Test the null hypothesis that $\beta_1 = 0$ at a 1% significance level.

Solution: (5 points)

$$H_0 : \beta_1 = 0 \quad vs \quad H_0 : \beta_1 \neq 0$$

Compute the t-statistic:

$$t = \frac{0.080}{0.002} = 40$$

The critical value at a 1% significance level is 2.58. Since 40 is bigger than 2.58 we reject the null hypothesis that $\beta_1 = 0$ at a 1% significance level.

b) Interpret the two estimated coefficients.

Solution: (10 points) $\widehat{\beta_0} = 0.34$ *is the expected probability of passing the exam when study time is 0 hours.* $\widehat{\beta_1} = 0.08$ *is the increase in the probability of passing the exam that is associated with an increase in study time by 1 hour.*

c) Explain whether we can interpret the estimated OLS coefficient on $studytime_{it}$ as the causal effect of study time on the probability of passing an exam.

Solution: *(10 points) We can only interpret the OLS coefficient on* $studytime_{it}$ *as the causal effect of study time on the probability of passing the exam if the OLS assumptions hold, in particular* $E\left[u_{it}|studytime_{it}\right] = 0$ *should hold. Students that spent a lot of time on preparing for an exam might differ in characteristics from those that spent very little time on preparing for an exam and these characteristics might affect the probability of passing an exam. For example, motivated students might spent a lot of time on preparing for an exam and pay good attention during lectures, while less motivated students don't pay attention during the lectures and spend very little time on preparing for an exam. There are therefore reasons to suspect that* $E\left[u_{it}|studytime_{it}\right] \neq 0$ *and that we can't interpret the OLS coefficient on* $studytime_{it}$ *as the causal effect of study time on the probability of passing the exam.*

d) The teacher decides to augment the model with student fixed effects (α_i)

$$passed_{it} = \beta_0 + \beta_1 studytime_{it} + \alpha_i + \varepsilon_{it}$$

Explain how you could estimate this model.

Solution: *(10 points) There are two ways they only have to discuss one.*

1. *Least Squares with dummy variables: Create 500 dummy variables for the students* $D1_i, D2_i, ..., D500$ *with* $D1_i = 1$ *if* $i = 1$ *and zero otherwise,* $D2_i = 1$ *if* $i = 2$ *and zero otherwise etc. Augment the model by including* $500 - 1$ *dummy variables, or include* 500 *dummy variables and exclude the constant term.*

2. *Within estimation. First step: demean* $passed_{it}$ *and* $studytime_{it}$*. Second step: regress* $(passed_{it} - \overline{passed}_i)$ *on* $(studytime_{it} - \overline{studytime}_i)$*.*

e) The teacher is confident that by including student fixed effects the estimated coefficient on $studytime_{it}$ cannot suffer from omitted variable bias problems. Explain whether you agree with the teacher.

Solution: *(10 points) No, there can still be omitted variable bias. Student fixed effects control for all (unobserved) variables that are constant over time. However, variables that vary over time that are omitted from the regression can still cause omitted variable bias in a model with student fixed effects. For example if students become more (or less) motivated over time this can affect both study time and the probability of passing the exam. There might also be exam (time) specific characteristics that affect both study time and the probability of passing the exam.*

Question 3

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

a) If the sample size is large and we perform a t-test with a critical value equal to 1.96, the probability of rejecting the null hypothesis when it is true is 5%

Solution: *(5 points) Correct, the t-statistic is approximately normally distributed* $N(0,1)$ *when the sample size is large and the null hypothesis is true. We reject the null hypothesis if* $|t| > 1.96$ *and because the area under the tails of the standard normal distribution outside* ± 1.96 *equals 5%, the probability of rejecting the null hypothesis when it is true equals 5%.*

b) The R^2 can never be equal to 0.

Solution: *(5 points) Incorrect. The* R^2 *is the ratio of the explained sum of squares to the total sum of squares. If a regression model does not include explanatory variables the model does not explain anything and the explained sum of squares is zero and the* R^2 *equals 0.*

c) If we have measurement error in the explanatory variable(s) we can solve this by computing heteroskedasticity robust standard errors.

Solution: (5 points) *Incorrect. Measurement error in the explanatory variable leads in most cases to a violation of the first OLS assumption* ($E\left[u_i|X_i^*\right] \neq 0$) *and the OLS estimator of the coefficient of the explanatory variable that is measured with error will be biased and inconsistent. We can't solve this problem by computing heteroskedasticity robust standard errors, because this will adjust the standard errors such that they are consistent in the presence of heteroskedasticity but this will not solve for any bias in the estimated coefficient due to measurement error.*

Question 4

Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$ with $E\left[u_i | X_i\right] = E\left[u_i | W_i\right] = 0$ and $E\left[W_i | X_i\right] = \alpha$. A researcher does not observe W_i and estimates the following regression model by OLS

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

Show whether the OLS estimator is a biased or unbiased estimator of β_1.

Solution: *(10 points)*

$$True\ model:\ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i, \quad Estimated\ model:\ Y_i = \beta_0 + \beta_1 X_i + v_i$$

$$\begin{aligned}
E\left[\widehat{\beta}_1\right] &= E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right] \\
& && substitute\ for\ (Y_i - \overline{Y}) \\
&= E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\beta_1\left(X_i - \overline{X}\right) + \beta_2\left(W_i - \overline{W}\right) + (u_i - \overline{u})\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right] \\
& && rewrite \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}\beta_2\left(X_i - \overline{X}\right)\left(W_i - \overline{W}\right) + \sum_{i=1}^{n}\left(X_i - \overline{X}\right)(u_i - \overline{u})}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right] \\
&= \beta_1 + E\left[\frac{\beta_2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)W_i + \sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right] \\
& && Law\ of\ it.\ exp. \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)E[W_i | X_i] + \sum_{i=1}^{n}\left(X_i - \overline{X}\right)E[u_i | X_i]}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right] \\
& && E\left[u_i | X_i\right] = 0 \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)E[W_i | X_i]}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}\right]
\end{aligned}$$

so

$$E\left[\widehat{\beta}_1\right] = \beta_1 \quad if \quad E\left[W_i | X_i\right] = 0$$

We know $E\left[W_i | X_i\right] = \alpha$ so $\widehat{\beta}_1$ is biased if $\alpha \neq 0$!