Exam ECON3150/4150: Introductory Econometrics.

25 May 2018; 14:30h-17.30h.

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.

Guidelines for correctors: *The exam has 15 sub-questions and for each sub-question a maximum of 10 points can be obtained. This means that a total of 150 points can be obtained in this exam. Last year the following cut-offs were used to convert points to grades:*

A	*135≤points*
B	*119≤points≤134*
C	*89≤points≤118*
D	*68≤points≤88*
E	*46≤points≤67*
F	*points≤45*

Question 1

A researcher wants to investigate whether parents' participation in a welfare program increases the probability that their child will also participate in a welfare program as an adult. She has a data set with information on 10 000 children and their parents. The dependent variable $Wchild_i$ is a binary variable that equals 1 if the child receives welfare benefits when he is between 18 and 30 years old. The explanatory variable $Wparent_i$ equals 1 if the parents received welfare benefits when the child was between 12 and 18 years old.

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

$$Wchild_i = \beta_0 + \beta_1 \cdot Wparent_i + u_i \tag{1}$$

and obtains the following estimation result

```
. regress Wchild Wparent, robust

Linear regression                               Number of obs     =     10,000
                                                F(1, 9998)        =      10.00
                                                Prob > F          =     0.0016
                                                R-squared         =     0.0013
                                                Root MSE          =     .21783

------------------------------------------------------------------------------
             |               Robust
      Wchild |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Wparent |   .0219976   .0069551     3.16   0.002     .0083641    .0356311
       _cons |   .0467356   .0022875    20.43   0.000     .0422516    .0512195
------------------------------------------------------------------------------
```

Give an interpretation, in words, of the two estimated coefficients, $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Solution (10 points): $\widehat{\beta}_0 = 0.046$ *is the fraction of children that participates in a welfare program among those that don't have a parent that participated in a welfare program.* $\widehat{\beta}_1 = 0.022$ *is the difference in the fraction of children that that participates in a welfare program between those with and without a parent that participated in a welfare program. The fraction of children that participates in a welfare program among those that have a parent that participated in a welfare program is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 0.069$.

b) Is the coefficient on $Wparent_i$ significantly different from zero at a 1 percent significance level?

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

$$t = \frac{0.022 - 0}{0.007} = 3.16$$

The absolute value of the t-statistic is bigger than 2.58 so we reject H_0. *The coefficient on* $Wparent_i$ *is significantly different from zero at a 1 percent significance level.*

c) Do you think that the OLS estimator of β_1 is an unbiased estimator of the causal effect of parents' welfare participation on child's welfare participation as an adult? Explain why or why not.

Solution (10 points): *To answer this question students need to think about potential threats to internal validity. One potential threat to the internal validity is omitted variable bias. Parents that participate in a welfare program likely differ in characteristics, such as educational attainment, health and ability, from parents that do not participate in a welfare program. If these characteristics affect child's welfare participation, for example because these characteristics are passed on from parents to children, they will create omitted variable bias in the OLS estimator of β_1 in equation (1). Another potential threat to internal validity that will cause the OLS estimator to be biased is measurement error (in case of survey data).*

d) The data set also includes the variable $edu\,parent_i$ which contains the average number of years of education completed by the parents. The variable $edu\,parent_i$ is negatively correlated with parents' welfare participation ($Wparent_i$) and has a negative effect on child's welfare participation ($Wchild_i$). Explain what will happen with the estimated coefficient on $Wparent_i$ when $edu\,parent_i$ is included as control variable in the OLS regression of $Wchild_i$ on $Wparent_i$?

Solution (10 points):
Suppose the following holds:

$$\begin{aligned} True\ model: \quad & Wchild_i = \beta_0 + \beta_1 \cdot Wparent_i + \beta_2 \cdot edu\,parent_i + v_i \\ & E\left(v_i | Wparent_i, edu\,parent_i\right) = 0 \end{aligned}$$

$$Estimated\ model\ part\ (a): \quad Wchild_i = \beta_0 + \beta_1 \cdot Wparent_i + u_i$$

Then it can be shown that

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(Wparent_i, edu\,parent_i)}{Var(Wparent_i)}$$

Since the variable $edu\,parent_i$ is negatively correlated with parent's welfare participation ($Wparent_i$) we have that $Cov\left(Wparent_i, edu\,parent_i\right) < 0$. In addition $edu\,parent_i$ has a negative effect on child's welfare participation ($Wchild_i$) which implies that $\beta_2 < 0$. A variance is never negative, we therefore have that the probability limit of $\widehat{\beta}_1 > \beta_1$ in part (a) where $edu\,parent$ is not included. If we include $edu\,parent_i$ as a control variable this will therefore reduce the coefficient estimate on $Wparent_i$.

e) Since the dependent variable $Wchild_i$ is a binary variable, the researcher decides to estimate a probit model and obtains the following estimation result

```
Probit regression                               Number of obs     =     10,000
                                                Wald chi2(2)      =     376.66
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -1716.3527               Pseudo R2         =     0.1354
```

Wchild	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Wparent	.1411273	.0596087	2.37	0.018	.0242964	.2579582
edu_parent	-.2019093	.010536	-19.16	0.000	-.2225594	-.1812591
_cons	.6865118	.117618	5.84	0.000	.4559849	.9170388

What is the estimated effect of parent's welfare participation on the probability that the child participates in a welfare program, given that the parent has obtained 12 years of education?

Solution (10 points): *The estimated effect of parent's welfare participation on the probability that the child participates in a welfare program, given that the parent has obtained 12 years of education, equals:*

$$\begin{aligned} \triangle Pr(\widehat{Wchild_i} = 1) \quad = \quad & \Pr(Wchild = 1|\widehat{Wparent_i} = 1,\ edu\,parent_i = 12) \\ - \ & \Pr(Wchild = 1|\widehat{Wparent_i} = 0,\ edu\,parent_i = 12) \end{aligned}$$

$$\begin{aligned} \triangle Pr(\widehat{Wchild_i} = 1) &= \Phi\,(0.687 + 0.141 * 1 - 0.202 * 12) - \Phi\,(0.687 + 0.141 * 0 - 0.202 * 12) \\ &= \Phi\,(-1.60) - \Phi\,(-1.74) \\ &= 0.0548 - 0.0409 \\ &= 0.0139 \end{aligned}$$

f) Construct a 90 percent confidence interval around the coefficient on $Wparent_i$ in the probit regression model.

Solution (10 points): *90% confidence interval for $\beta_{Wparent}$ is*

$$\left[\widehat{\beta}_{Wparent} - 1.64 \times SE\left(\widehat{\beta}_{Wparent}\right), \widehat{\beta}_{Wparent} + 1.64 \times SE\left(\widehat{\beta}_{Wparent}\right)\right]$$

Using the results in the Stata output gives:

$$[0.141 - 1.64 \times 0.060\ ,\ 0.141 + 1.64 \times 0.060]$$

$$[-0.043\ ,\ -0.239]$$

g) The researcher also estimates a logit model and obtains the following estimation results

```
Logistic regression                             Number of obs     =     10,000
                                                Wald chi2(2)      =     418.46
                                                [redacted]
Log pseudolikelihood = -1718.0542               Pseudo R2         =     0.1345
```

Wchild	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Wparent	.2732244	.1199389	2.28	0.023	.0381485	.5083004
edu_parent	-.4377319	.0217626	-20.11	0.000	-.4803858	-.3950781
_cons	2.040748	.2336515	8.73	0.000	1.5828	2.498697

```
. test Wparent edu_parent,[redacted]

 ( 1)  [Wchild]Wparent = 0
 ( 2)  [Wchild]edu_parent = 0

       F( [redacted] ) =  209.23
       [redacted]
```

What is the estimated effect of parent's welfare participation on the probability that the child participates in a welfare program, given that the parent has obtained 12 years of education?

Solution (10 points): *The estimated effect of parent's welfare participation on the probability that the child participates in a welfare program, given that the parent has obtained 12 years of education, equals:*

$$\begin{aligned}\triangle Pr(\widehat{Wchild_i = 1}) \quad &= \quad \Pr(Wchild = 1 | \widehat{Wparent_i = 1},\ edu\, parent_i = 12) \\ &- \ \Pr(Wchild = 1 | \widehat{Wparent_i = 0},\ edu\, parent_i = 12)\end{aligned}$$

$$\begin{aligned}\triangle Pr(\widehat{Wchild_i = 1}) \quad &= \left(1/\left(1+e^{-(2.040+0.273\cdot 1-0.438\cdot 12)}\right)\right) - \left(1/\left(1+e^{-(2.040+0.273\cdot 0-0.438\cdot 12)}\right)\right) \\ &= 0.050 - 0.039 \\ &= 0.012\end{aligned}$$

h) Test the null hypothesis that both the coefficients on $Wparent_i$ and $edu\, parent_i$ are zero using a 5 percent significance level.

Solution (10 points): $H_0 : \beta_{Wparent} = 0$ & $\beta_{edu\, parent} = 0$ *vs* $H_1 : \beta_{Wparent} \neq 0$ *and/or* $\beta_{edu\, parent} \neq 0$
The F-statistic is given in the Stata output and equals F=209.23. There are 2 restrictions under the null hypothesis and the number of observations is large which implies that we can use the following critical value $F^{5\%}_{2,\infty} = 3.00$. *Since 209,23>3 we reject the null hypothesis at a 5% significance level.*

i) A reform took place that made it more difficult to participate in a welfare program. This reform affected about half of the parents. The researcher decides to use this reform as an instrument for parent's welfare participation and estimates the following first stage regression by OLS

$$Wparent_i = \pi_0 + \pi_1 \cdot reform_i + \varepsilon_i$$

She obtains the following estimation results

```
. regress Wparent reform, robust

Linear regression                               Number of obs     =     10,000
                                                [redacted]
                                                R-squared         =     0.0431
                                                Root MSE          =     .34778
```

Wparent	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
reform	-.1476753	.0069603	-21.22	0.000	-.1613189	-.1340316
_cons	.2223558	.005886	37.78	0.000	.210818	.2338935

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

Solution (10 points): *Instrument relevance,* $Corr(Wparent_i, reform_i) \neq 0$*, can be investigated using the first stage regression. The first stage F-statistic equals* $F = (t)^2 = \left(\frac{-0.148}{0.007}\right)^2 = 450$*, which is bigger than the rule-of-thumb value of 10. The instrument relevance condition holds and* $reform_i$ *is a not a weak instrument.*

j) The following table shows the fraction of children and the fraction of parents that participated in a welfare program separately for the children with parents that were affected by the reform ($reform_i = 1$) and for the children with parents that were not affected by the reform ($reform_i = 0$) . Use the results in the table below to obtain the instrumental variable estimate of the effect of $Wparent_i$ on $Wchild_i$.

	$reform_i = 1$	$reform_i = 0$
$\widehat{E}\left[Wchild_i \mid reform_i = x\right]$	0.049	0.050
$\widehat{E}\left[Wparent_i \mid reform_i = x\right]$	0.075	0.222

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

$$\hat{\beta}_{IV} = \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2} = \frac{\widehat{E}\left[Wchild_i \mid reform_i = 1\right] - \widehat{E}\left[Wchild_i \mid reform_i = 0\right]}{\widehat{E}\left[Wparent_i \mid reform_i = 1\right] - \widehat{E}\left[Wparent_i \mid reform_i = 0\right]}$$

The instrumental variable estimate of the effect of $Wparent_i$ on years of education ($Wchild_i$) equals

$$\hat{\beta}_{IV} = \frac{0.049 - 0.050}{0.075 - 0.222} = 0.007$$

Question 2

A policy maker wants to know whether the inflow of immigrants affects the wages of native workers. The country is divided into two regions, region A and region B. There was a sudden influx of immigrants into region A but not in region B. The policy maker has information about wages of native workers in regions A and B both before and after the influx of immigrants. The following stata output shows the averages of the logaritm of wages of native workers (*lnwage*):

```
. bys region time: sum lnwage

-> region = A, time = after

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lnwage |      3,040    2.890215    .0553958   2.699678   3.059546

-> region = A, time = before

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lnwage |      2,942    2.994545    .0489649   2.797889   3.160841

-> region = B, time = after

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lnwage |      1,984    3.064744    .0463179    2.87966   3.239686

-> region = B, time = before

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lnwage |      2,034    3.090116    .0460191   2.919806   3.215129
```

a) Compute the difference-in-differences estimate of the effect of the inflow of immigrants on the logarithm of wages of native workers.

Solution (10 points)

$$\begin{aligned}
\widehat{\beta}_{DID} &= \Big(E[\widehat{ln(wage)}_{i\,A\,after}] - E[\widehat{ln(wage)}_{i\,A\,before}]\Big) - \Big(E[\widehat{ln(wage)}_{i\,B\,after}] - E[\widehat{ln(wage)}_{i\,B\,before}]\Big) \\
&= (2.890 - 2.995) - (3.065 - 3.090) \\
&= -0.080
\end{aligned}$$

b) Interpret the sign and magnitude of the difference-in-differences estimate obtained in part (a).

Solution (10 points) *The inflow of immigrants reduced the wages of native workers on average by about 8 percent.*

c) Explain the common trend assumption in the context of the application in this exercise.

Solution (10 points) *In absence of the inflow of immigrants the trend in the logaritm of wages (ln(wage)) should have been the same in region A and region B.*

Question 3

A researcher wants to estimate the effect of an additional year of schooling (S_i) on yearly earnings (E_i). Consider the following population regression model $E_i = \beta_0 + \beta_1 S_i + u_i$ with $Cov\left(S_i, u_i\right) = 0$. The researcher has a large data set with i.i.d observations on years of schooling S_i and on yearly earnings reported to the tax authority E_i^*. According to a colleague of the researcher, individuals under-report their earnings to the tax authority to reduce the amount of taxes they have to pay. This means that the observed taxable earnings differ from true earnings, more specifically $E_i^* = \gamma \cdot E_i$ with $0 < \gamma < 1$. The researcher wants to estimate the causal effect of an additional year of schooling on *true* earnings. He estimates the following equation by OLS

$$E_i^* = \beta_0 + \beta_1 S_i + v_i$$

a) Express v_i in terms of $\beta_0,\ \beta_1,\ \gamma,\ S_i,\ u_i$ and show that $Cov\left(S_i, v_i\right) = \left(\gamma - 1\right)\beta_1 Var\left(S_i\right)$

Solution (10 points)

$$\begin{array}{ccc}
E_i^* & = & \beta_0 + \beta_1 S_i + v_i \\
\gamma E_i & = & \beta_0 + \beta_1 S_i + v_i \\
\gamma\left(\beta_0 + \beta_1 S_i + u_i\right) & = & \beta_0 + \beta_1 S_i + v_i \\
\left(\gamma\beta_0 + \gamma\beta_1 S_i + \gamma u_i\right) - \left(\beta_0 + \beta_1 S_i\right) & = & v_i \\
\left(\gamma - 1\right)\beta_0 + \left(\gamma - 1\right)\beta_1 S_i + \gamma u_i & = & v_i
\end{array}$$

This implies that $v_i = \left(\gamma - 1\right)\beta_0 + \left(\gamma - 1\right)\beta_1 S_i + \gamma u_i$

$$\begin{aligned}
Cov\left(S_i, v_i\right) &= Cov\left(S_i,\ \left(\gamma - 1\right)\beta_0 + \left(\gamma - 1\right)\beta_1 S_i + \gamma u_i\right) \\
&= Cov\left(S_i,\ \left(\gamma - 1\right)\beta_0\right) + Cov\left(S_i,\ \left(\gamma - 1\right)\beta_1 S_i\right) + Cov\left(S_i,\ \gamma u_i\right) \\
&= 0 + \left(\gamma - 1\right)\beta_1 Var\left(S_i\right) + \gamma Cov\left(S_i,\ u_i\right) \\
&= \left(\gamma - 1\right)\beta_1 Var\left(S_i\right)
\end{aligned}$$

b) Is the OLS estimator of β_1 a consistent estimator of the causal effect of an additional year of schooling on *true* earnings? Show why or why not.

Solution (10 points):

$$\widehat{\beta_1} = \frac{s_{E^*S}}{s_S^2} \xrightarrow{p} \frac{Cov(E_i^*, S_i)}{Var(S_i)}$$

$$\begin{aligned}
\widehat{\beta_1} \xrightarrow{p} \frac{Cov(E_i^*, S_i)}{Var(S_i)} &= \frac{Cov(\beta_0+\beta_1 S_i+v_i,\, S_i)}{Var(S_i)} \\
&= \frac{Cov(\beta_0, S_i)+\beta_1 Cov(S_i, S_i)+Cov(S_i, v_i)}{Var(S_i)} \\
&= \beta_1 + \frac{Cov(S_i, v_i)}{Var(S_i)} \\
&= \beta_1 + \frac{(\gamma-1)\beta_1 Var(S_i)}{Var(S_i)} \\
&= \beta_1 + (\gamma - 1)\,\beta_1 \\
&= \gamma\beta_1
\end{aligned}$$

This means that the OLS estimator of β_1 is inconsistent because it converges to $\gamma \cdot \beta_1$ so it underestimates the true causal effect by $100\,(1-\gamma)\%$.