Postponed Exam ECON3150/4150: Introductory Econometrics.
Spring 2021

Guidelines for correctors: The exam has 19 sub-questions and for each sub-question a maximum of 5 points can be obtained, except for question 2 a) which has a maximum of 10 points. This means that a total of 100 points can be obtained in this exam. Based on student performance in previous years I suggest to use the following cut-offs to convert points to grades:

A	90≤points
B	80≤points≤89
C	60≤points≤79
D	46≤points≤59
E	36 ≤points≤45
F	points≤ 35

Question 1

A researcher wants to investigate whether parents' smoking behavior affects the probability that their child smokes as an adult. She has a data set with information on 10 000 children and their parents. The dependent variable $smoke\,child_i$ is a binary variable that equals 1 if the child smokes when she is between 18 and 30 years old and zero otherwise. The explanatory variable $smoke\,parent_i$ equals 1 if at least one of the parents smoked when the child was between 12 and 18 years old and zero otherwise.

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

$$smoke\,child_i = \beta_0 + \beta_1 \cdot smoke\,parent_i + u_i \tag{1}$$

and obtains the following estimation result

```
model1 <- lm( smoke_child ~ smoke_parent, data = data)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.0913574  0.0031224 29.2583 < 2.2e-16
## smoke_parent 0.0582382  0.0097720 [redacted]
```

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

Solution: $\widehat{\beta}_0 = 0.091$ is the fraction of children that smokes as an adult among those whose parents did not smoke. $\widehat{\beta}_1 = 0.058$ is the difference in the fraction of children that smoke as an adult between those with and without at least one parent that smoked during the child's adolescence. The fraction of children that smokes among those that have a parent that smoked is equal to $\widehat{\beta}_0 + \widehat{\beta}_1 = 0.149$.

b) Is the coefficient on $smoke\,parent_i$ significantly different from zero at a 1 percent significance level?

Solution: $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$. Construct the t-statistic:

$$t = \frac{0.058 - 0}{0.0098} = 5.9$$

The absolute value of the t-statistic is bigger than 2.58 so we reject H_0. The coefficient on $smoke\,parent_i$ 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' smoking behavior on the probability that the child smokes as an adult? Explain why or why not.

Solution: 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 smoke likely differ in characteristics, such as educational attainment and ability, from parents that do not smoke. If these characteristics affect the likelihood that a child smokes as an adult, 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 measures the average number of years of education completed by the parents. Parents that smoke are on average lower educated than parents that do not smoke and parents' education has a negative relation with the probability that the child smokes as an adult. Explain what will happen with the estimated coefficient on $smoke\,parent_i$ when $edu\,parent_i$ is included as control variable in the OLS regression of $smoke\,child_i$ on $smoke\,parent_i$?

Solution:

Suppose the following holds:

$$\textit{True model}: \quad \textit{smoke child}_i = \beta_0 + \beta_1 \cdot \textit{smoke parent}_i + \beta_2 \cdot \textit{edu parent}_i + v_i$$
$$E\left(v_i | \textit{smoke parent}_i, \textit{edu parent}_i\right) = 0$$

$$\textit{Estimated model part } (a): \quad \textit{smoke child}_i = \beta_0 + \beta_1 \cdot \textit{smoke parent}_i + u_i$$

Then it can be shown that

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(\textit{smoke parent}_i, \textit{edu parent}_i)}{Var(\textit{smoke parent}_i)}$$

Since the variable $\textit{edu parent}_i$ is negatively correlated with $\textit{smoke parent}_i$ we have that $Cov\left(\textit{smoke parent}_i, \textit{edu parent}_i\right) < 0$. In addition $\textit{edu parent}_i$ is negatively related with $\textit{smoke child}_i$ 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 $\textit{edu parent}_i$ as a control variable this will therefore reduce the coefficient estimate on $\textit{smoke parent}_i$.

e) Since the dependent variable $\textit{smoke child}_i$ is a binary variable, the researcher decides to estimate a probit model and obtains the following estimation results

```
probit <- glm(smoke_child ~ smoke_parent + edu_parent,
              family = binomial(link = "probit"),
              data = data)

coeftest(probit,vcovHC(probit, type = "HC1"))
```

```
##
## z test of coefficients:
##
##                Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)   1.1726470  0.0910926  12.8731 < 2.2e-16
## smoke_parent  0.2407112  0.0475460   5.0627 4.134e-07
## edu_parent   -0.2101799  0.0079185 -26.5430 < 2.2e-16
```

What is the estimated effect of having a parent that smokes (compared to having nonsmoking parents) on the probability that the child smokes as an adult, given that the parents obtained on average 14 years of education?

Solution: The estimated effect of having a parent that smokes on the probability that the child smokes as an adult, given that the parents obtained on average 14 years of education, equals:

$$\begin{aligned}
\triangle Pr(\widehat{smoke\ child}_i = 1) \quad &= \quad \Pr(smoke\ child_i = 1 | \widehat{smoke\ parent}_i = 1,\ edu\ parent_i = 14) \\
&\quad - \ \Pr(smoke\ child_i = 1 | \widehat{smoke\ parent}_i = 0,\ edu\ parent_i = 14) \\
\triangle Pr(\widehat{smoke\ child}_i = 1) \quad &= \Phi\left(1.1726 + 0.2407 - 0.2102 \cdot 14\right) - \Phi\left(1.1726 - 0.2102 \cdot 14\right) \\
&= \Phi\left(-1.53\right) - \Phi\left(-1.77\right) \\
&= 0.0630 - 0.0384 \\
&= 0.0246
\end{aligned}$$

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

Solution: 90% confidence interval for $\beta_{smoke\ parent}$ is

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

Using the results in the R output gives:

$$[0.2407 - 1.64 \times 0.0475\ ,\ 0.2407 + 1.64 \times 0.0475]$$

$$[0.1628\ ,\ 0.3186]$$

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

```
logit <- glm(smoke_child ~ smoke_parent + edu_parent,
             family = binomial(link = "logit"),
             data = data)

coeftest(logit,vcovHC(logit, type = "HC1"))
```

```
##
## z test of coefficients:
##
##                Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)   2.550546   0.168967  15.0950 < 2.2e-16
## smoke_parent  0.451192   0.087752   5.1416 2.723e-07
## edu_parent   -0.412443   0.015276 -26.9990 < 2.2e-16
```

What is the estimated effect of having a parent that smokes (compared to having nonsmok-

ing parents) on the probability that the child smokes as an adult, given that the parents obtained on average 14 years of education?

Solution: The estimated effect of having a parent that smokes on the probability that the child smokes as an adult, given that the parents obtained on average 14 years of education, equals*:*

$$\begin{aligned}
\triangle Pr(\widehat{smoke}\, child_i = 1) &= \Pr(smoke\, child_i = 1 | \widehat{smoke}\, parent_i = 1,\ edu\, parent_i = 14) \\
&\quad - \Pr(smoke\, child_i = 1 | \widehat{smoke}\, parent_i = 0,\ edu\, parent_i = 14) \\
\triangle Pr(\widehat{smoke}\, child_i = 1) &= \left(1/\left(1 + e^{-(2.55+0.45-0.41\cdot 14)}\right)\right) - \left(1/\left(1 + e^{-(2.55-0.41\cdot 14)}\right)\right) \\
&= 0.061 - 0.040 \\
&= 0.021
\end{aligned}$$

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

```
linearHypothesis(logit, c("smoke_parent", "edu_parent"),
                 test=c("F"), vcov = vcovHC(logit, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## smoke_parent = 0
## edu_parent = 0
##
## Model 1: restricted model
## Model 2: smoke_child ~ smoke_parent + edu_parent
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   9999
## 2   [redacted]  [redacted] 378.72 [redacted]
```

Solution: $H_0 : \beta_{smoke\, parent} = 0\ \&\ \beta_{edu\, parent} = 0$ vs $H_1 : \beta_{smoke\, parent_i} \neq 0\ and/or\ \beta_{edu\, parent} \neq 0$ The F-statistic is given in the R output and equals F=378.72. 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 378.72>3 we reject the null hypothesis at a 5% significance level.

i) The government implemented a smoking ban in the public sector, but not in the private sector. All parents that worked in the public sector were no longer allowed to smoke

during work time. The researcher decides to use this implementation of a smoking ban as an instrument for parents' smoking behaviour and estimates the following first stage regression by OLS

$$smoke\ parent_i = \pi_0 + \pi_1 \cdot smoke\ ban_i + \varepsilon_i \tag{2}$$

She obtains the following estimation results

```
FirstStage<- lm( smoke_parent ~ smoke_ban, data = data)
coeftest(FirstStage,vcovHC(FirstStage, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.2223558  0.0058860  37.777 < 2.2e-16 ***
## smoke_ban   -0.1476753  0.0069603 -21.217 < 2.2e-16 ***
```

Do you think that the instrument relevance condition holds? Is $smoke\ ban_i$ a weak instrument?

Solution: Instrument relevance, $Corr(smoke\ parent_i, smoke\ ban_i) \neq 0$, can be investigated using the first stage regression. The first stage F-statistic equals $F = (t)^2 = (-21.217)^2 = 450.16$, which is much bigger than the rule-of-thumb value of 10. The instrument relevance condition holds and $smoke\ ban_i$ is a not a weak instrument.

j) The researcher estimates the following equation by OLS

$$smoke\ child_i = \delta_0 + \delta_1 smoke\ ban_i + \epsilon_i \tag{3}$$

and obtains the following estimation results.

```
ReducedForm<- lm( smoke_child ~ smoke_ban, data = data)
coeftest(ReducedForm,vcovHC(ReducedForm, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1019631  0.0042833 23.8050   <2e-16 ***
## smoke_ban   -0.0039200  0.0060006 -0.6533   0.5136
```

Use these results in combination with the first stage estimation results from part i) to obtain the instrumental variable estimate of the effect of $smoke\ parent_i$ on $smoke\ child_i$. Give an interpretation of this instrumental variable estimate in words.

Solution:

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

- $\frac{S_{ZY}}{S_Z^2}$ 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 *smoke parent*$_i$ on *smoke child*$_i$ equals

$$\widehat{\beta}_{IV} = \frac{\widehat{\delta}_1}{\widehat{\pi}_1} = \frac{-0.0039}{-0.1477} = 0.026$$

Having at least one parent that smoked during the child adolescence is estimated to increase the probability that the child smokes as an adult by 2.6 percentage points.

Question 2

The directorate of education wants to know whether the time of the day that an exam takes place affects exam scores. The country is divided into two regions, region A and region B. Initially the exam took place in the afternoon both in regions A and B, but region A decided to move the exam to the morning. The directorate of education has information about exam scores of students in regions A and B both before, when the exam took place in the afternoon in both regions A and B, and after region A decided to have the exam take place in the morning. The following R output shows the averages of the logarithm of exam scores (*ln examscore*):

```
aggregate(ln_examscore ~ time + region, data = data, mean)
```

```
##     time region ln_examscore
## 1  after      A     2.770598
## 2 before      A     2.705574
## 3  after      B     2.600211
## 4 before      B     2.561860
```

a) Compute the difference-in-differences estimate of the effect of the time of the day the exam takes place on the logarithm of exam scores

Solution:

$$\begin{aligned}\widehat{\beta}_{DID} &= \Big(E[\widehat{ln(examscore)}_{i\ A\ after}] - E[\widehat{ln(examscore)}_{i\ A\ before}]\Big) \\ &\quad - \Big(E[\widehat{ln(examscore)}_{i\ B\ after}] - E[\widehat{ln(examscore)}_{i\ B\ before}]\Big) \\ &= (2.771 - 2.706) - (2.600 - 2.562) \\ &= 0.027\end{aligned}$$

b) Interpret the sign and magnitude of the difference-in-differences estimate obtained in 2(a).

Solution: Taking the exam in the morning instead of the afternoon is estimated to increase exam scores on average by about 2.7 percent.

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

Solution: In absence of a change in the time of day the exam takes place the trend in the logaritm of exam scores should have been the same in region A and region B.

Question 3

A researcher wants to investigate if the number of hours students spend on preparing for a test has an effect on test scores. She has information on test scores, the level of difficulty of each test and she collects information on test preparation time by conducting surveys among students. This results in a panel data set with information on 200 students and for each student she observes the score obtained on 10 different tests. The data set contains the variable $score_{it}$ which measures the test score obtained by student i on test t and the variable $hours_{it}$ which measures the number of hours that student i spent on preparing for test t.

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

$$ln\left(score_{it}\right) = \beta_0 + \beta_1 \cdot ln\left(hours_{it}\right) + u_{it} \tag{4}$$

She obtains the following estimation results

```
model1 <- lm( ln_score ~ ln_hours, data = data2)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3559     0.0998    3.56  0.00037
## ln_hours      0.7317     0.0387   18.89  < 2e-16
## ---
```

Give an interpretation, in words, of the estimated coefficient $\widehat{\beta}_1$.

Solution: $\widehat{\beta}_1 = 0.7317$ .This is a log-log model. The (approximate) interpretation of $\widehat{\beta}_1$ is that if the number of hours a student spent preparing for the exam increases by 1 percent this is associated with an increase in the test score by about 0.73 percent.

b) Name and explain two examples of potential threats to the internal validity when estimating equation (4) by OLS.

Solution: One potential threat to the internal validity is omitted variable bias. Students that spent many hours preparing for a test might also differ in other characteristic that affect the test score, they might be more motivated, or be of higher (or lower) ability compared to students who spent less time on preparing for a test. It might also be that students spent more time preparing for difficult tests than for easier tests (also an example of omitted variable bias). Measurement error is also a potential problem. It might be difficult to collect accurate information on the number of hours each student spent on preparing for each of the tests.

c) The researcher wants to analyze whether the effect of test preparation time differs between difficult and easy tests. Describe in detail how you can test the null hypothesis that the effect of the logaritm of the hours a student spent on preparing for a test does not differ between difficult and easy tests.

Solution: The researcher should first create a binary variable which equals 1 for a difficult test ($difficult_t$) and zero for an easy test. The regression should next be augmented to include an interaction term between $ln\,(hours_{it})$ and the dummy variable $difficult_t$ as follows:

$$ln(score_{it}) = \lambda_0 + \lambda_1 \cdot ln\,(hours_{it}) + \lambda_2 difficult_t + \lambda_3 (ln\,(hours_{it}) \cdot difficult_t) + \varepsilon_{it}$$

The hypothesis can be tested by using a t test testing H_0: $\lambda_3 = 0$.

d) The researcher decides to include test fixed effects. She estimates the following regression model

$$ln\,(score_{it}) = \lambda_1 \cdot ln\,(hours_{it}) + \eta_t + \epsilon_{it} \tag{5}$$

and obtains the following estimation results.

```
within <- plm(ln_score ~ ln_hours, data = data2,
          index = c("test"), model = "within")
class(within)
```

```
## [1] "plm"         "panelmodel"
```

```
coeftest(within,vcovHC(within, type = "HC1"))
```

```
##
## t test of coefficients:
##
##          Estimate Std. Error t value Pr(>|t|)
## ln_hours   0.9791     0.0501    19.5   <2e-16
##
```

Compare these results to the results in part a) and explain whether the results differ and if so why.

Solution: The estimated coefficient on the variable $ln\,(hours_{it})$ when including test fixed effects is bigger than the estimated coefficient on $ln\,(hours_{it})$ in the regression model without fixed effects in part a). This indicates that the regression model without test fixed effects in part a) suffers from omitted variable bias. Tests for which students spent more hours to prepare seem to differ in characteristics from test for which students spent less preparation time and these characteristics affect test scores. It could for example be that students score on average lower on difficult tests, but they also spent more time on preparing for these difficult tests.

e) A colleague of the researcher constructs binary variables for each of the tests and estimates the following regression model by OLS.

$$ln\left(score_{it}\right) = \lambda_1 \cdot ln\left(hours_{it}\right) + \tau_1 test1_t + ... + \tau_{10} test10_t + \epsilon_{it} \tag{6}$$

How will the estimated effect of $ln\left(hours_{it}\right)$ on $ln\left(score_{it}\right)$ obtained by the colleague compare to the estimate obtained by the researcher in part d)?

Solution: The estimates will be identical. The fixed effect regression model can be estimated by within estimation as is done in part (d) or by the Least Squares with Dummy variable method, both estimation methods will give identical estimates of λ_1, the effect of $ln\left(hours_{it}\right)$ on $ln\left(score_{it}\right)$.

f) What will happen if the colleague estimates the following equation by OLS instead of equation 6?

$$ln\left(score_{it}\right) = \lambda_0 + \lambda_1 \cdot ln\left(hours_{it}\right) + \tau_1 test1_t + ... + \tau_{10} test10_t + \epsilon_{it} \tag{7}$$

Solution: Equation 7 cannot be estimated by OLS because of perfect multicollinearity. The (regressor on) the constant term is a perfect linear combination of the 10 included test-dummy variables. Either the constant term or one of the dummy variables for the tests should be omitted from the regression model.