Postponed Exam ECON3150/4150: Introductory Econometrics.
Spring 2021

Question 1 - 50%

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 results

```
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 $\widehat{\beta}_1$.

b) Is the coefficient on $smoke\,parent_i$ 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.

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*$_i$?

e) Since the dependent variable *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?

f) Construct a 90 percent confidence interval around the coefficient on *smoke parent*$_i$ in the probit regression model.

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

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]
```

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?

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.

Question 2 - 20%

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

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

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

Question 3 - 30%

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 $\widehat{\beta}_1$.

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

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.

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

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

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.

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)?

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

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