Exam ECON3150/4150: Introductory Econometrics.
Spring 2021

Question 1 - weight 50%

A teacher wants to know the effect of summer schools on the probability of passing the exam. He sets up an experiment in order to estimate the average causal effect of participating in a summer school. The teacher randomly assigns 400 students either to a treatment group or a control group. The 200 students assigned to the treatment group go to a summer school during the summer holidays, while the 200 students in the control group don't go to school. At the end of the summer the students take an exam. The data set collected by the teacher contains the binary variable $passed_i$ which equals one if the student passed the exam and zero if the student failed as well as a binary variable $summer\ school_i$ which equals one if the student participated in the summer school and the variable $disadvantaged_i$ which equals one for students coming from a disadvantaged background.

a) The teacher estimates the following regression model by OLS

$$passed_i = \beta_0 + \beta_1 \cdot summer\ school_i + \beta_2 \cdot disadvantaged_i + u_i \tag{1}$$

and obtains the following estimation results

```
model1 <- lm( passed ~ summer_school+disadvantaged, data = data)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    0.774568   0.036062 21.4786 < 2.2e-16 ***
## summer_school  0.117437   0.042210  2.7822  0.005657 **
## disadvantaged -0.170845   0.042473 -4.0224 6.896e-05 ***
## ---
```

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

b) Construct a 90 percent confidence interval for difference in the probability of passing the exam between disadvantaged and non-disadvantaged students.

c) The teacher wants to test the hypothesis that both participation in the summer school and coming from a disadvantaged background are not related with the probability of passing the exam, using a 1 percent significance level. She obtains the following results:

```
linearHypothesis(model1, c("summer_school", "disadvantaged"),
                 vcov = vcovHC(model1, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## summer_school = 0
## disadvantaged = 0
##
## Model 1: restricted model
## Model 2: passed ~ summer_school + disadvantaged
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1
## 2        12.552
```

What is the conclusion of the teacher?

d) Part of the students that participated in the summer school decide to go on holiday after the summer school and do not take the exam. These students are thus not part of the sample that is used to estimate equation (1). Is the OLS estimator of β_1 an unbiased estimator of the causal effect of participating in the summer school on the probability of passing the exam?

e) The teacher wants to know if disadvantaged and non-disadvantaged students are differentially affected by participation in the summer school. She decides to estimate the following regression model by OLS

$$\begin{aligned} passed_i = \lambda_0 + \lambda_1 \cdot summer\ school_i + \lambda_2 \cdot disadvantaged_i \\ + \lambda_3 \cdot (disadvantaged_i \times summer\ school_i) + \epsilon_i \end{aligned} \tag{2}$$

and obtains the following estimation results

```
model2 <- lm( passed ~ summer_school+disadvantaged+(summer_school*disadvantaged),
              data = data)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                   0.782178   0.041279 18.9487 < 2.2e-16 ***
## summer_school                 0.102437   0.051916  1.9731  0.049176 *
## disadvantaged                -0.186219   0.064504 -2.8869  0.004103 **
## summer_school:disadvantaged   0.030770   0.085029  0.3619  0.717639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the estimated effect of participating in the summer school for disadvantaged students and for non-disadvantaged students (give an interpretation in words)?

f) The teacher decides to estimate a logit model and obtains the following estimation results

```
logit <- glm(passed ~ summer_school+disadvantaged+(summer_school*disadvantaged),
              family = binomial(link = "logit"),
              data = data)

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

```
##
## z test of coefficients:
##
##                              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                   1.27841    0.24228  5.2766 1.316e-07 ***
## summer_school                 0.75848    0.39224  1.9337  0.053151 .
## disadvantaged                -0.88975    0.31792 -2.7987  0.005132 **
## summer_school:disadvantaged -0.15674    0.49951 -0.3138  0.753687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the estimated effect of participating in the summer school for disadvantaged students and for non-disadvantaged students?

g) Does the 99 percent confidence interval around the logit coefficient on the interaction term between $summer\ school_i$ and $disadvantaged_i$ include the value zero?

h) The teacher decides to estimate a probit model and obtains the following estimation results

```
probit <- glm(passed ~ summer_school+disadvantaged+(summer_school*disadvantaged),
              family = binomial(link = "probit"),
              data = data)

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

```
##
## z test of coefficients:
##
##                               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                   0.779571   0.140211  5.5600 2.698e-08 ***
## summer_school                 0.418809   0.214121  1.9559  0.050471 .
## disadvantaged                -0.536668   0.189826 -2.8272  0.004696 **
## summer_school:disadvantaged -0.051417   0.284904 -0.1805  0.856783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the estimated effect of participating in the summer school for disadvantaged students and for non-disadvantaged students ?

i) Some of the students who were assigned to the control group went to school during the summer. Explain the consequences for the interpretation of the estimation results in part a).

j) The teacher claims that she can still estimate the causal effect of participating in the summer school on the probability of passing the exam, because she collected data on assignment to the treatment and control group as well as information on actual participation in the summer school. Do you agree with the teacher, explain why or why not.

Question 2 - weight 50%

A researcher wants to investigate if opening hours of shopping malls have an effect on total sales. She has a panel data set with information on 200 shopping malls for the years 2000-2010. The data set contains the variable $sales_{it}$ which measures the total sales in shopping mall i in year t and the variable $hours_{it}$ which measures the number of hours that shopping mall i was open during year t.

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

$$ln(sales_{it}) = \beta_0 + \beta_1 \cdot hours_{it} + u_{it} \tag{3}$$

She obtains the following estimation results

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.2328491  0.0592211    20.8   <2e-16
## hours       0.0016999  0.0000275    [redacted]
## ---
```

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

b) Is the coefficient on $hours_{it}$ significantly different from zero at a 1 percent significance level?

c) Name and explain two examples of potential threats to the internal validity when estimating equation (3) by OLS.

d) The researcher wants to analyze whether the effect of opening hours differs between large and small shopping malls. Describe in detail how you can test the null hypothesis that the effect of opening hours does not differ between large and small shopping malls.

e) The researcher decides to use an instrumental variable approach to estimate the causal effect of opening hours on the logarithm of total sales. In 2006 there was an economic crisis and many shopping malls reduced their opening hours such that they could lay off part of their store employees. The researcher decides to create a binary variable $crisis_t$ which equals one for all shopping malls in 2006 and zero otherwise. She estimates the following first stage regression model by OLS

$$hours_{it} = \delta_0 + \delta_1 \cdot crisis_t + \epsilon_{it} \tag{4}$$

and obtains the following estimation results

```
first_stage <- lm( hours ~ crisis, data = data2)
coeftest(first_stage,vcovHC(first_stage, type = "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1994.5       20.3   98.13   <2e-16 ***
## crisis        -210.2       66.1   -3.18   0.0015 **
## ---
```

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

f) The researcher wants to control for omitted variables that are common across shopping malls and that vary over time and includes year fixed effects. She creates binary variables for each of the years 2000-2010 and includes all these binary variables in the regression model which results in the following first stage regression model

$$hours_{it} = \theta_0 + \theta_1 \cdot crisis_t + \tau_1 \cdot year2000 + ... + \tau_8 \cdot year2010 + \mu_{it} \tag{5}$$

Explain what issue will arise when estimating equation (5) by OLS.

g) The following table shows the sample means of $ln(sales_{it})$ and $hours_{it}$ separately for the year in which there was an economic crisis and for the other years. Use the results in the table below to obtain the instrumental variable estimate of the effect of opening hours on the logarithm of total sales (using $crisis_t$ as instrument). Give an interpretation, in words, of this instrumental variable estimate.

	Sample mean	
	$ln(sales_{it})$	$hours_{it}$
Year with economic crisis (2006)	4.061	1784.278
Other years	4.644	1994.521

h) Do you think that, when using $crisis_t$ as an instrument to estimate the causal effect of effect of opening hours on the logarithm of total sales, the instrument exogeneity condition holds? Explain why or why not.

i) Instead of using an instrumental variable approach the researcher decides to include shopping mall fixed effects. She estimates the following regression model

$$ln(sales_{it}) = \beta_0 + \beta_1 \cdot hours_{it} + \eta_i + \varepsilon_{it} \tag{6}$$

and obtains the following estimation results.

```
within <- plm(ln_sales ~ hours, data = data2,
          index = c("id"), model = "within")
class(within)
```

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

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

```
##
## t test of coefficients:
##
##        Estimate Std. Error t value Pr(>|t|)
## hours 0.0014797  0.0000276    53.6   <2e-16
## ---
```

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

j) A colleague of the researcher claims that in order to eliminate omitted variable bias the researched should include both shopping mall fixed effects and time fixed effects simultaneously. Do you agree with this colleague, explain why or why not.