Postponed Exam ECON3150/4150: Introductory Econometrics.
Spring 2020

Question 1

A researcher wants to investigate if losing your job has an effect on health. She has a panel data set with information on 5000 individuals for the years 2001-2009. The dependent variable $bad\ health_{it}$ is a binary variable equal to one if individual i has bad health in year t and zero otherwise. The explanatory variable $job\ loss_{it}$ is a binary variable equal to one if individual i lost his job in year t and zero otherwise and age_{it} is the age (in years) of individual i in year t.

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

$$bad\ health_{it} = \beta_0 + \beta_1 \cdot job\ loss_{it} + u_{it} \tag{1}$$

She obtains the following estimation results

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 0.2441088  0.0021000 116.2408 < 2.2e-16
## job_loss    0.0839469  0.0086148 [redacted]
## ---
```

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

b) Is the coefficient on $job\ loss_{it}$ significantly different from zero at a 5 percent significance level?

c) Do you think that the OLS estimator of β_1 is an unbiased estimator of the causal effect of job loss on the probability of having bad health? Explain why or why not.

d) The researcher wants to analyze whether the effect of job loss differs between workers who are older than 45 and workers who are younger than 45. Describe in detail how you can test the null hypothesis that the effect of job loss does not differ between workers who are older than 45 and workers who are younger than 45.

e) The researcher decides to estimate a logit model and obtains the following estimation results

```
logit <- glm(bad_health ~ job_loss + age,
             family = binomial(link = "logit"),
             data = data)

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

```
##
## z test of coefficients:
##
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -2.9741065  0.0975084 -30.501 < 2.2e-16
## job_loss     0.4191794  0.0398514  10.519 < 2.2e-16
## age          0.0732721  0.0038254  19.154 < 2.2e-16
##
```

What is the estimated effect of job loss on the probability of having bad health for an individual who is 30 years old?

f) The researcher decides to estimate a probit model and obtains the following estimation results

```
probit <- glm(bad_health ~ job_loss + age,
             family = binomial(link = "probit"),
             data = data)

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

```
##
## z test of coefficients:
##
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -1.7780485  0.0570372 -31.174 < 2.2e-16
## job_loss     0.2514164  0.0241925  10.392 < 2.2e-16
## age          0.0431650  0.0022484  19.198 < 2.2e-16
##
```

What is the estimated effect of job loss on the probability of having bad health for an individual who is 30 years old?

g) Construct a 95 percent confidence interval for the probit coefficient on age_{it}.

h) The researcher decides to use an instrumental variable approach to estimate the causal effect of job loss on the probability of having bad health. In 2005 there was a financial crisis and many companies had to lay off part of their employees. The researcher decides to create a binary variable $crisis_t$ which equals one for all individuals in 2005 and zero otherwise. She estimates the following first stage regression model by OLS

$$job\ loss_{it} = \delta_0 + \delta_1 \cdot crisis_t + \epsilon_{it} \qquad (2)$$

and obtains the following estimation results

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.0688250  0.0012658 54.3723 < 2.2e-16
## crisis      0.0121750  0.0040609  2.9981  0.002718
## ---
```

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

i) The researcher wants to control for omitted variables that are common across individuals and that vary over time and includes year fixed effects. She creates binary variables for the years 2002, 2003, 2004, 2005, 2006, 2007, 2008 and 2009 estimates the following first stage regression model

$$job\ loss_{it} = \theta_0 + \theta_1 \cdot crisis_t + \tau_1 \cdot year2002 + ... + \tau_8 \cdot year2009 + \mu_{it} \tag{3}$$

and obtains the following estimation results.

```
data$year <- factor(data$year)
levels(data$year)
```

```
## [1] "2001" "2002" "2003" "2004" "2005" "2006" "2007" "2008" "2009"
```

```
first_stage <- lm( job_loss ~ crisis + year, data = data)
coeftest(first_stage,vcovHC(first_stage, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0678000  0.0035557 19.0679  < 2e-16 ***
## crisis       0.0132000  0.0052473  2.5156  0.01189 *
## year2002    -0.0042000  0.0049555 -0.8476  0.39669
## year2003    -0.0008000  0.0050148 -0.1595  0.87325
## year2004     0.0050000  0.0051133  0.9778  0.32816
## year2006    -0.0042000  0.0049555 -0.8476  0.39669
## year2007     0.0070000  0.0051465  1.3601  0.17379
## year2008    -0.0026000  0.0049835 -0.5217  0.60187
## year2009     0.0080000  0.0051630  1.5495  0.12127
## ---
```

Explain why the R-output does not show an estimated coefficient on $year2005$. Is it possible to estimate the coefficient on the binary variable for the year 2005 in equation 3?

j) The following table shows the sample means of $bad\ health_{it}$ and $job\ loss_{it}$ separately for the year in which there was a financial crisis and for the other years. Use the results in the table below to obtain the instrumental variable estimate of the effect of job loss on the probability of having bad health (using $crisis_t$ as instrument). Give an interpretation, in words, of this instrumental variable estimate.

	Sample mean	
	$bad\ health_{it}$	$job\ loss_{it}$
Year with financial crisis (2005)	0.251	0.081
Other years	0.250	0.069

k) Do you think that, when using $crisis_t$ as an instrument to estimate the causal effect of $job\ loss_{it}$ on $bad\ health_{it}$, the instrument exogeneity condition holds? Explain why or why not.

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

$$bad\ health_{it} = \beta_0 + \beta_1 \cdot job\ loss_{it} + \eta_i + \varepsilon_{it} \tag{4}$$

and obtains the following estimation results.

```
within <- plm(bad_health ~ job_loss, data = data,
          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|)
## job_loss 0.0174389  0.0085994  2.0279  0.04258
""
```

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

Question 2

A teacher wants to know the effect of digital teaching on student test scores. He sets up an experiment in order to estimate the average causal effect of digital instead of physical teaching on student performance. The teacher randomly assigns 1000 students either to a treatment group or a control group. The 500 students assigned to the treatment group watch recorded lectures on their computer at home, while the 500 students in the control group follow regular teaching in a class room. At the end of the course all students make the same test. The data set collected by the teacher contains the test scores of the students as well as a binary variable $digital_i$ which equals one if the student watched the recorded lectures and zero if the student attended the physical lectures and the variable $female_i$ which equals one for female students and zero for male students.

a) The teacher constructs a variable which is the logarithm of test scores and estimates the following regression model by OLS

$$ln\left(testscore_i\right) = \beta_0 + \beta_1 \cdot digital_i + \beta_2 \cdot female_i + u_i \tag{5}$$

and obtains the following estimation results

```
model1 <- lm( ln_testscore ~ digital + female, data = data2)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  4.4081287  0.0024440 1803.624 < 2.2e-16 ***
## digital     -0.0408812  0.0028594  -14.297 < 2.2e-16 ***
## female       0.1005872  0.0028564   35.215 < 2.2e-16 ***
## ---
```

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

b) Construct a 99 percent confidence interval for the (approximate) percentage difference in test scores between female and male students.

c) The teacher wants to test the hypothesis that both the coefficients on $digital_i$ and $female_i$ are zero versus the alternative that at least one of these coefficients is nonzero, using a 5 percent significance level. She obtains the following results:

```
linearHypothesis(model1, c("digital", "female"),
                 test=c("F"), vcov = vcovHC(model1, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## digital = 0
## female = 0
##
## Model 1: restricted model
## Model 2: ln_testscore ~ digital + female
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1 [redacted]
## 2 [redacted] 703.04 [redacted]
## ---
```

What is the conclusion of the teacher?

d) All lectures are in English, but the test is in Norwegian. Part of the students is foreign and they have difficulties reading Norwegian. Does this affect the interpretation of the estimated coefficient on $digital_i$ in part (a), is β_1 an unbiased estimator of the causal effect of digital teaching on test scores?

e) The teacher wants to know if male and female students are differentially affected by digital teaching. She decides to estimate the following regression model by OLS

$$ln\,(testscore_i) = \lambda_0 + \lambda_1 \cdot digital_i + \lambda_2 \cdot female_i + \lambda_3 \cdot (digital_i \times female_i) + u_i \qquad (6)$$

and obtains the following estimation results

```
model2 <- lm( ln_testscore ~ digital*female, data = data2)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)     4.4290705  0.0024726 1791.241 < 2.2e-16
## digital        -0.0818822  0.0036877  -22.204 < 2.2e-16
## female          0.0587036  0.0035079   16.735 < 2.2e-16
## digital:female  0.0838484  0.0050581   16.577 < 2.2e-16
""
```

What is the estimated effect of digital teaching on test scores for male students (give an interpretation in words)?

f) Give an interpretation, in words, of the estimated coefficient $\widehat{\lambda}_3$.

g) Teacher discovers that some students who were assigned to the digital teaching attend the physical lectures. Explain the consequences for the interpretation of the estimation results in part a).

h) Can the teacher still use the data with information on test scores, assignment to digital and physical teaching and lecture attendance to estimate the causal effect of digital teaching on test scores? Explain why not or explain how the teacher should do this.