Postponed Exam ECON3150/4150: Introductory Econometrics.
Spring 2020

Guidelines for correctors: The exam has 20 sub-questions and for each sub-question a maximum of 5 points can be obtained. 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 (but since this is a home exam instead of a regular exam we need to see whether this is indeed the best way to convert points to grades):

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

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 $\widehat{\beta}_1$.

Solution:$\widehat{\beta}_1 = 0.084$ is the estimated change in the probability of having bad health when the variable $job\ loss_{it}$ increases from zero to one. Job loss is thus associated with an increase in the probability of having bad health by 8.4 percentage points.

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

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

$$t = \frac{0.0839469 - 0}{0.0086148} = 9.7$$

The absolute value of the t-statistic is bigger than 1.96 so we reject H_0. The coefficient on $job\ loss_{it}$ is 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.

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. Individuals that lose their job likely differ in characteristics, such as motivation and skills, from individuals that do not lose their job. If these characteristics affect health, for example because less motivated individuals live less healthy, they will create omitted variable bias in the OLS estimator of β_1 in equation (1).

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.

Solution: The researcher should first create a binary variable which equals 1 for workers older than 45 and zero otherwise. The regression should next be augmented to include an interaction term between job loss and the dummy variable $older45_{it}$ as follows:

$$bad\ health_{it} = \lambda_0 + \lambda_1 \cdot job\ loss_{it} + \lambda_2 \cdot older45_{it} + \lambda_3 \cdot (job\ loss_{it} \times older45_{it}) + \varepsilon_{it}$$

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

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

```r
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?

Solution: the estimated effect of job loss on the probability of having bad health for an individual who is 30 years old*:*

$$\begin{aligned}
\triangle Pr(bad\ \widehat{health} = 1|age = 30) &= \left(1/\left(1 + e^{-(-2.97+0.419+0.073\times 30)}\right)\right) - \left(1/\left(1 + e^{-(-2.97+0.073\times 30)}\right)\right) \\
&= 0.411 - 0.314 \\
&= 0.096
\end{aligned}$$

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

```r
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?

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

$$\begin{aligned}\triangle Pr(\widehat{bad\ health} = 1|age = 30) &= \Phi\left(-1.778 + 0.251 + 0.043 \times 30\right) - \Phi\left(-1.778 + 0.043 \times 30\right)\\ &= \Phi\left(-0.24\right) - \Phi\left(-0.49\right)\\ &= 0.4052 - 0.3121\\ &= 0.093\end{aligned}$$

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

Solution: 95% confidence interval:

$$\widehat{\beta_{age}} \pm 1.96 \times SE(\widehat{\beta}_{age})$$

filling in the numbers from the regression output gives

$$0.0431650 \pm 1.96 \times 0.0022484$$

$$(0.039 \quad , \quad 0.048)$$

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

$$job\ loss_{it} = \delta_0 + \delta_1 \cdot crisis_t + \epsilon_{it} \tag{2}$$

and obtains the following estimation results

```r
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?

Solution: Instrument relevance, $Cov(job\ loss_{it}, crisis_t) \neq 0$, can be investigated using the first stage regression. The estimated coefficient on $crisis_t$ is significantly different from zero at a 1 percent significance level, so the instrument relevance condition seems to hold. However, the first stage F-statistic equals $F = (t)^2 = \left(\frac{0.0121750}{0.0040609}\right)^2 = 8.99$, which is smaller than the rule-of-thumb value of 10, which implies that $crisis_t$ is 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?

Solution: The variable $crisis_t$ equals 1 for the year 2005 and zero otherwise. The binary variable $year2005$ equals 1 for the year 2005 and zero otherwise. This implies that the variables $crisis_t$ and $year2005$ are identical and cannot both be included in the regression model, the two variables are perfectly multicollinear. It is not possible to estimate the coefficient on the binary variable for the year 2005 in equation 3, because this equation also includes the variable $crisis_t$.

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

Solution: The instrument $crisis_t$ is binary, we therefore have that the IV estimator equals the so called Wald estimator:

$$\hat{\beta}_{IV} = \frac{\widehat{E}\left[bad\ health_{it}|crisis_t = 1\right] - \widehat{E}\left[bad\ health_{it}|crisis_t = 0\right]}{\widehat{E}\left[job\ loss_{it}|crisis_t = 1\right] - \widehat{E}\left[job\ loss_{it}|crisis_t = 0\right]}$$

the instrumental variable estimate of the effect of job loss on the probability of having bad health equals:

$$\hat{\beta}_{IV} = \frac{0.251 - 0.250}{0.081 - 0.069} = 0.083$$

this can be interpreted as that job loss increases the probability of bad health by about 8.3 percentage points.

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.

Solution: The instrument exogeneity condition might be violated because the financial crisis can have a direct impact on health independent of the effect via job loss. A financial crisis might lead for example to more stress which can result in bad health.

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.

```r
within <- plm(bad_health ~ job_loss, data = data,
              index = c("id"), model = "within")
class(within)
```

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

```r
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.

Solution: The estimated coefficient on the variable $job\ loss_{it}$ when including individual fixed effects is much smaller than the estimated coefficient on $job\ loss_{it}$ in the regression model without fixed effects in part a). This indicates that the regression model without fixed effects in part a) suffers from omitted variable bias. Individuals that lose their job seem to differ in time-invariant characteristics from individuals that do not lose their job, and these characteristics affect health.

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

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

Solution: The estimated coefficient on $digital_i$ equals -0.04. It is a log-linear model which implies that we can interpret the estimated coefficient as follows: digital teaching decreases test scores by on average 4 percent. (100*$\widehat{\beta}_1$)= -4% .

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

Solution: 99% confidence interval for the (approximate) percentage difference in test scores between female and male students:

$$100 * \left(\widehat{\beta_2} \pm 2.58 \times SE(\widehat{\beta}_2)\right)$$

filling in the numbers from the regression output gives

$$100 * (0.101 \pm 2.58 \times 0.003)$$

$$(9.3 \quad , \quad 10.9)$$

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:

```r
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?

Solution: $H_0 : \beta_1 = 0 \;\&\; \beta_2 = 0$ vs H_1 :at least one of the coefficients is unequal to zero.
The F-statistic is given in the R output and equals F=703.04. There are 2 restrictions under the null hypothesis and the number of observations is large (n=1000) which implies that we can use the following critical value $F^{5\%}_{2,\infty} = 3.00$. Since 703.04>3.00, the teacher rejects the null hypothesis at a 5% significance level.

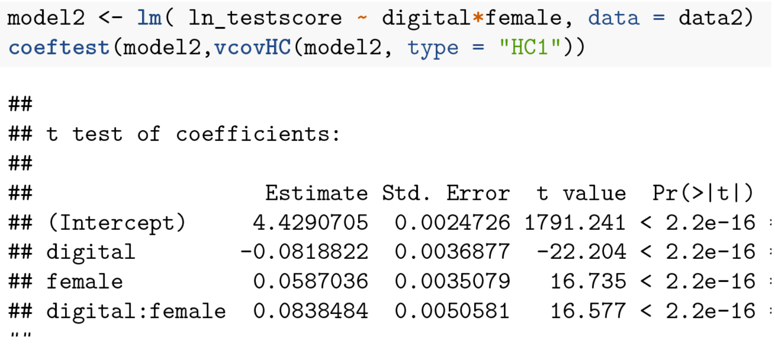
d) All lectures are in English, but the test is in Norwegian. Part of the students is foreign and they have difficulties reading Norwegian. Does this affect the interpretation of the estimated coefficient on $digital_i$ in part (a), is β_1 an unbiased estimator of the causal effect of digital teaching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

Solution: All students are randomly assigned to the treatment and control group. This implies that foreign students are equally likely to end up in the treatment or control group. The treatment and control group make the same test. This implies that the fact that the test is in Norwegian does not affect the interpretation of the estimated coefficient on $digital_i$ in part (a), and β_1 is 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

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

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)?

Solution: The estimated effect of digital teaching on the logarithm of test scores for male students equals $\widehat{\lambda_1} = -0.0818822$. Digital teaching decreases test scores of male students on average by about 8.1 percent.

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

Solution: $\widehat{\lambda}_3$ measures the difference in the effect of digital teaching on the logaritm of test scores between female and male students. Digital teaching is estimated to increase test scores of female students on average by about (100*(-0.0818822 + 0.0838484)) 0.2 percent, while digital teaching is estimated to decreases test scores of male students on average by about 8.1 percent.

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).

Solution: This is an example of failure to follow the treatment protocol, or partial compliance. Since students who are assigned to the treatment group and decide to attend the physical lectures might differ in characteristics from the other students in the treatment group, the estimated coefficient on $digital_i$ might pick up the effect of these characteristics on test scores and therefore not provide an unbiased and consistent estimate of the causal effect of digital teaching on test scores.

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.

Solution: The teacher can use the instrumental variable approach. He can use the assignment to the treatment and control group as instrument for $digital_i$ and estimate the following model

$$digital_i = \pi_0 + \pi_1 \cdot treatmentgroup_i + \varepsilon_i$$

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

where $treatmentgroup_i$ equals 1 for students randomly assigned to the treatment group en zero for students assigned to the control group.