Exam ECON3150/4150: Introductory Econometrics.
Spring 2021

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:

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 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

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

Solution: $\widehat{\beta}_1 = 0.117437$ is the estimated change in the probability of passing the exam when the variable *summer school* increases from zero to one. Participation in the summer school is thus associated with an increase in the probability of passing the exam by 11.7 percentage points.

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

Solution: 90% confidence interval for difference in the probability of passing the exam between disadvantaged and non-disadvantaged students.:

$$\left(\widehat{\beta_2} \pm 1.645 \times SE(\widehat{\beta}_2)\right)$$

filling in the numbers from the regression output gives

$$(-0.170845 \pm 1.645 \times 0.042473)$$

$$(-0,2407 \quad , \quad -0,1010)$$

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?

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=12.552. There are 2 restrictions under the null hypothesis and the number of observations is large (n=400) which implies that we can use the following critical value $F_{2,\infty}^{1\%} = 4.61$. Since 12.552>4.61, the teacher rejects the null hypothesis at a 1% significance level.

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?

Solution: This is an example of attrition. Whether or not the attrition causes bias in the estimated coefficient on $summer\ school_i$ depends on whether the students that decide to go on holiday differ in characteristics (that affect test scores) from those that do not go on holiday. If the students that go on holiday would have had a different probability of passing the exam compared to the students who do not go on holiday, this will make $\widehat{\beta}_1$ a biased estimator of the causal effect of participating in the summer school.

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)?

Solution: The estimated effect of participating in the summer school for non-disadvantaged students equals $\widehat{\lambda_1} = 0.102437$. Participating in the summer school increases the probability of passing the exam for non-disadvantaged students by about 10.2 percentage points. The estimated effect of participating in the summer school for disadvantaged students equals $\widehat{\lambda_1} + \widehat{\lambda_3} = 0.133207$. Participating in the summer school increases the probability of passing the exam for disadvantaged students by about 13.3 percentage points.

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?

Solution (10 points): the estimated effect of participating in the summer school for non-disadvantaged students*:*

$$\begin{aligned}\triangle Pr(\widehat{passed=1}|disadv=0) &= \left(1/\left(1+e^{-(1.278+0.758)}\right)\right) - \left(1/\left(1+e^{-(1.278)}\right)\right) \\ &= 0.885 - 0.782 \\ &= 0.103\end{aligned}$$

the estimated effect of participating in the summer school for disadvantaged students*:*

$$\begin{aligned}\triangle Pr(\widehat{passed=1}|disadv=1) &= \left(1/\left(1+e^{-(1.278+0.758-0.890-0.157)}\right)\right) - \left(1/\left(1+e^{-(1.278-0.890)}\right)\right) \\ &= 0.729 - 0.596 \\ &= 0.133\end{aligned}$$

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?

Solution (10 points): The 99% confidence interval is

$$[-0.15674 - 2.58 \times 0.49951\ ,\ -0.15674 + 2.58 \times 0.49951]$$

$$[-1.445\ ,\ 1.132]$$

The confidence interval includes 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 ?

Solution (10 points): the estimated effect of participating in the summer school for non-disadvantaged students*:*

$$\begin{aligned}
\triangle Pr(\widehat{passed = 1}|disadv = 0) &= \Phi(0.780 + 0.419) - \Phi(0.780) \\
&= \Phi(1.20) - \Phi(0.78) \\
&= 0.8849 - 0.7823 \\
&= 0.103
\end{aligned}$$

the estimated effect of participating in the summer school for disadvantaged students*:*

$$\begin{aligned}
\triangle Pr(\widehat{passed = 1}|disadv = 1) &= \Phi(0.780 + 0.419 - 0.537 - 0.051) - \Phi(0.780 - 0.537) \\
&= \Phi(0.61) - \Phi(0.24) \\
&= 0.7291 - 0.5948 \\
&= 0.134
\end{aligned}$$

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).

Solution: This is an example of failure to follow the treatment protocol, or partial compliance. Since students who are assigned to the control group and decide to go to school might differ in characteristics from the other students in the control group, the estimated coefficient on $summer\ school_i$ might pick up the effect of these characteristics on exam results and therefore not provide an unbiased and consistent estimate of the causal effect of participating in the summer school on the probability of passing the exam.

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.

Solution: The teacher is right. She can use the instrumental variable approach. She can use the assignment to the treatment and control group as instrument for $summer\ school_i$ and estimate the following model

$$summer\ school_i = \pi_0 + \pi_1 \cdot treatmentgroup_i + \varepsilon_i$$

$$passed_i = \beta_0 + \beta_1 \cdot summer\ school_i + u_i$$

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

Question 2

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 $\widehat{\beta}_1$.

Solution: $\widehat{\beta}_1 = 0.0016999$ .This is a log-linear model. The (approximate) interpretation of $\widehat{\beta}_1$ is that if a shopping mall is open for 1 additional hour this is associated with an increase in total sales by about 0.17 percent ($100^*\beta_1\%$).

b) Is the coefficient on $hours_{it}$ 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:

$$t = \frac{0.0016999 - 0}{0.0000275} = 61.9$$

The absolute value of the t-statistic is bigger than 2.58 so we reject H_0. The coefficient on $hours_{it}$ is 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.

Solution: One potential threat to the internal validity is omitted variable bias. Shopping malls that are open for more hours might also differ in other characteristic that affect sales, they might for example have more shops and therefore higher total sales regardless of the opening hours. Reversed causality might also be a possible threat to internal validity. If a shopping mall has many customers who buy a lot it might decide to have longer opening hours to make sure that all customers have a chance to shop without the shops being overcrowded.

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.

Solution: The researcher should first create a binary variable which equals 1 for large shopping malls ($large\ mall_i$) and zero otherwise. The regression should next be augmented to include an interaction term between $hours_{it}$ and the dummy variable $large\ mall_i$ as follows:

$$ln(sales_{it}) = \lambda_0 + \lambda_1 \cdot hours_{it} + \lambda_2 large\ mall_i + \lambda_3 (hours_{it} \cdot large\ mall_i) + \varepsilon_{it}$$

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

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?

Solution: Instrument relevance, $Cov(hours_{it}, crisis_t) \neq 0$, 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. The first stage F-statistic equals $F = (t)^2 = (-3.18)^2 = 10.11$, which is just above the rule-of-thumb value of 10, which implies that $crisis_t$ is a not weak instrument (although also not very strong).

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.

Solution: The issue that will arise is perfect multicollinearity. The variable $crisis_t$ equals 1 for the year 2006 and zero otherwise. The binary variable $year2006$ equals 1 for the year 2006 and zero otherwise. This implies that the variables $crisis_t$ and $year2006$ are identical and cannot both be included in the regression model, the two variables are perfectly multicollinear. It is not possible to estimate 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

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

$$\hat{\beta}_{IV} = \frac{\widehat{E}\left[ln(sales_{it})|crisis_t = 1\right] - \widehat{E}\left[ln(sales_{it})|crisis_t = 0\right]}{\widehat{E}\left[hours_{it}|crisis_t = 1\right] - \widehat{E}\left[hours_{it}|crisis_t = 0\right]}$$

the instrumental variable estimate of the effect of opening hours on the logarithm of total sales equals:

$$\hat{\beta}_{IV} = \frac{4.061 - 4.644}{1784.278 - 1994.521} = 0.0028$$

this can be interpreted as that a one hour increase in opening hours increases sales by about 0.28 percent.

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.

Solution: The instrument exogeneity condition is likely violated because during an economic crisis total sales will likely go down regardless of the opening hours because people consume and buy less.

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.

Solution: The estimated coefficient on the variable $hours_{it}$ when including individual fixed effects is smaller than the estimated coefficient on $hours_{it}$ 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. Shopping malls with long opening hours seem to differ in time-invariant characteristics from shopping malls with shorter opening hours, and these characteristics affect total sales.

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.

Solution: Including shopping mall fixed effects will control for omitted variables that vary across shopping malls but that are constant over time. Including time fixed effects will control for omitted variables that vary over time but that are constant across shopping malls. Including both shopping mall fixed effects and time fixed effects simultaneously will however not eliminate omitted variable bias, because there could still be omitted variables affecting total sales that vary both across shopping malls and over time and that are correlated with opening hours. The colleague is thus wrong.