Exam ECON3150/4150: Introductory Econometrics.
25 May 2020; 09:00 (5 hours)

If you are asked to derive something, give all intermediate steps. Do not answer questions with a "yes" or "no" only, but carefully motivate your answer.

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

Question 1

A researcher wants to investigate if the number of hours that children go to school during a year affects test scores. She has a panel data set with information on 150 regions for the years 2000-2010. The dependent variable $testscore_{it}$ is the average test score (in points) obtained by students in region i in year t. The explanatory variable $hours\,in\,school_{it}$ is the number of hours that students spent in school in region i in year t.

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

$$testscore_{it} = \beta_0 + \beta_1 \cdot hours\,in\,school_{it} + u_{it} \tag{1}$$

She obtains the following estimation results

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

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -16.349402   2.494307 -6.5547 7.437e-11
## hours_in_school   0.432520   0.014356 [redacted]
##
```

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

Solution (10 points): $\widehat{\beta}_1 = 0.43$ *is the estimated change in region average test scores when the number of hours that children go to school during a year increases by 1. An additional hour in school is thus associated with an increase in average test scores by 0.43 points.*

b) Is the coefficient on *hours in school*$_{it}$ significantly different from zero at a 1 percent significance level?

Solution (10 points): $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$. *Construct the t-statistic:*

$$t = \frac{0.43252 - 0}{0.014356} = 30.1$$

The absolute value of the t-statistic is bigger than 2.58 so we reject H_0. *The coefficient on hours in school*$_{it}$ *is significantly different from zero at a 1 percent significance level.*

c) The researcher decides to take the logarithm of test scores as dependent variable and estimates the following regression model by OLS

$$ln(testscore_{it}) = \pi_0 + \pi_1 \cdot hours\ in\ school_{it} + \varepsilon_{it} \tag{2}$$

She obtains the following estimation results

```
model2 <- lm( ln_testscore ~ hours_in_school, data = data)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     2.55937207 0.05911901  43.292 < 2.2e-16
## hours_in_school 0.00845617 0.00033237  25.442 < 2.2e-16
```

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

Solution (10 points): *This is a log-linear model. The (approximate) interpretation of* $\widehat{\pi}_1$ *is that if students spent an additional hour in school this is associated with an increase in average test scores by about 0.8 percent (100**$\widehat{\pi}_1$*%).*

d) Compute a 95 percent confidence interval around π_1.

Solution (10 points): *95% confidence interval for* π_1 *is*

$$[\widehat{\pi}_1 - 1.96 \times SE(\widehat{\pi}_1), \widehat{\pi}_1 + 1.96 \times SE(\widehat{\pi}_1)]$$

Using the results in the R output gives:

$$[0.00845517 - 1.96 \times 0.00033237 \ , \ 0.00845517 + 1.96 \times 0.00033237]$$

$$[0.0078 \ , \ 0.0091]$$

e) Do you think that the OLS estimator of π_1 is an unbiased estimator of the causal effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$? Explain why or why not.

Solution (10 points): *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. Regions in which students spent many hours in school might differ in characteristics from regions in which students spent fewer hours in school. For example if regions that care a lot about the education of children offer both high quality education and many school hours and regions that care less about education offer lower quality education and fewer hours in school, the OLS estimator of* π_1 *will be biased due to omitted variable bias. Another potential threat to internal validity that will cause the OLS estimator to be biased is measurement error. Especially measurement error in hours spent in school will lead to a biased OLS estimator of* π_1.

f) The researcher decides to use an instrumental variable approach to estimate the causal effect of hours spent in school on average test scores. In 2005 there was a pandemic and all schools were closed for part of the year. She decides to create a binary variable $pandemic_t$ which equals one for all regions in 2005 and zero otherwise. She estimates the following first stage regression model by OLS

$$hours\ in\ school_{it} = \delta_0 + \delta_1 \cdot pandemic_t + \epsilon_{it} \tag{3}$$

and obtains the following estimation results

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 173.72667    0.42246  411.23 < 2.2e-16 ***
## pandemic    -59.52667    1.29461 [redacted]
```

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

Solution (10 points): *Instrument relevance,* $Cov(hours\ in\ school_{it}, pandemic_t) \neq 0$, *can be investigated using the first stage regression. The first stage F-statistic equals* $F = (t)^2 = \left(\frac{-59.52667}{1.29461}\right)^2 = 2114.2$, *which is much bigger than the rule-of-thumb value of 10. The instrument relevance condition holds and* $pandemic_t$ *is a not a weak instrument.*

g) The researcher estimates the following regression model by OLS

$$ln(testscore_{it}) = \gamma_0 + \gamma_1 \cdot pandemic_t + \upsilon_{it}$$

and obtains the following estimation results.

```
reduced_form<- lm( ln_testscore ~ pandemic, data = data)
coeftest(reduced_form,vcovHC(reduced_form, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  4.004575   0.008133 492.3865 < 2.2e-16 ***
## pandemic    -0.240918   0.035967  -6.6983 2.884e-11 ***
```

Use these results in combination with the first stage results from part f) to compute the instrumental variable estimate of the effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$.

Solution (10 points): There is an alternative way of computing the instrumental variable estimator:

$$\begin{aligned}\hat{\beta}_{IV} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Z_i-\bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Z_i-\bar{Z})/\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})/\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2} \\ &= \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2}\end{aligned}$$

- $\frac{S_{ZY}}{S_Z^2}$ is the OLS estimator when regressing Y_i on Z_i
- $\frac{S_{ZX}}{S_Z^2}$ is the OLS estimator when regressing X_i on Z_i

This implies that the IV estimator of the effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$ equals

$$\widehat{\beta}_{IV} = \frac{\widehat{\gamma}_1}{\widehat{\delta}_1} = \frac{-0.240918}{-59.52667} = 0.004$$

h) Do you think that, when using $pandemic_t$ as an instrument to estimate the causal effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$, the instrument exogeneity condition holds? Explain why or why not.

Solution (10 points): *Instrument exogeneity:* $Cov(pandemic_t, \varepsilon_{it}) = 0$. *The instrument exogeneity condition consists of two components: independence and the exclusion restriction. Independence might be violated because* $pandemic_t$ *is equal to one for all regions in 2005 and might therefore pick up cohort effects. If there is something specific in 2005 which is unrelated to the pandemic this will be picked up by the instrument. The exclusion restriction might also be violated. If students become ill due to the pandemic this might have a direct effect on their test scores. (describing one potential reason for a violation of instrument exogeneity is enough to get full points).*

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

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\ in\ school_{it} + \eta_i + \nu_{it} \tag{4}$$

and obtains the following estimation results.

```
within <- plm(ln_testscore ~ hours_in_school, data = data,
          index = c("region_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_in_school 0.00379138 0.00032263  11.751 < 2.2e-16
```

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

Solution (10 points): *The estimated coefficient on the variable hours in school$_{it}$ when including region fixed effects is much smaller than the estimated coefficient on hours in school$_{it}$ in the regression model without fixed effects in part c). This indicates that the regression model without fixed effects in part c) suffers from omitted variable bias. Regions where students spent on average many hours in school seem to differ in (time-invariant) characteristics from regions where students spent on average fewer hours in school, and these characteristics affect average test scores.*

j) The researcher thinks that estimating the following model by OLS

$$ln(testscore_{it}) - ln(testscore_{it-1}) = \theta_1 \cdot (hours\ in\ school_{it} - hours\ in\ school_{it-1}) + (\nu_{it} - \nu_{it-1}) \tag{5}$$

will give an identical estimate of the causal effect of *hours in school*$_{it}$ on $ln(testscore_{it})$ as the estimate shown in the R-output in part i). Is she right, explain why or why not.

Solution (10 points): *When the number of time periods is exactly 2, so T=2, the within (or entity-demeaned) estimation procedure and the first-differences estimation procedure will give identical estimates of the causal effect of hours in school$_{it}$ on $ln(testscore_{it})$. In this exercise the number of time periods is not 2 but 11 (2000-2010), estimating equation (5) is therefore unlikely to give an identical estimate as the estimate obtained by the within estimation procedure in part i) (they might be very similar though).*

k) The test is in English and in some regions students don't have English as their native language. The researcher thinks this might affect test scores and decides to include the binary variable *no english*$_i$ which equals one for regions where students don't have English as their native language and zero otherwise. She estimates the following regression model

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\ in\ school_{it} + \theta_2 \cdot no\ english_i + \eta_i + \omega_{it} \tag{6}$$

and obtains the following estimation results.

```
within2 <- plm(ln_testscore ~ hours_in_school + no_english,
          data = data, index = c("region_id"), model = "within")
class(within)
```

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

```
coeftest(within2,vcovHC(within2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## hours_in_school 0.00379138 0.00032263  11.751 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' :
```

Explain why the R-output does not show an estimated coefficient on *no english*$_i$. Is it possible to estimate the coefficient on the variable *no english*$_i$ when region fixed effects are included in the regression model?

Solution (10 points): *The variable noenglish*$_i$ *varies between regions but it does not vary over time (it does not have a subscript t) it is therefore perfectly multicollinear with the region fixed effects. It is not possible to estimate a model that includes region fixed effects and no english*$_i$*. R therefore omits the variable no english*$_i$ *from the regression.*

l) The researcher wants to control for omitted variables that are common across regions but that vary over time and decides to include year fixed effects. She estimates the following regression model

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\,in\,school_{it} + \eta_i + \tau_1 \cdot year2001 + ... + \tau_{10} \cdot year2010 + \mu_{it} \quad (7)$$

She wants to test whether the time fixed effects are jointly significantly different from zero and performs an F-test with the following results:

```
linearHypothesis(within3, c("year2001", "year2002", "year2003", "year2004",
                            "year2005", "year2006", "year2007", "year2008",
                            "year2009","year2010" ),
                 test=c("F"), vcov = vcovHC(within3, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## year2001 = 0
## year2002 = 0
## year2003 = 0
## year2004 = 0
## year2005 = 0
## year2006 = 0
## year2007 = 0
## year2008 = 0
## year2009 = 0
## year2010 = 0
##
## Model 1: restricted model
## Model 2: ln_testscore ~ hours_in_school + year
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1 [redacted]
## 2 [redacted] 1.4715 [redacted]
```

Are the time fixed effects jointly significantly different from zero at a 1 percent significance level?

Solution (10 points): $H_0 : \tau_1 = 0 \,\&\, \tau_2 = 0 \,\&\, \tau_3 = 0 \,\&\, \tau_4 = 0 \,\&\, \tau_5 = 0 \,\&\, \tau_6 = 0 \,\&\, \tau_7 = 0 \,\&\, \tau_8 = 0 \,\tau_9 = 0 \,\&\, \tau_{10} = 0$ *vs* H_1 *:at least one of the coefficients* $\tau_1, ..., \tau_{10}$ *is unequal to zero.*
The F-statistic is given in the R output and equals F=1.47. There are 10 restrictions under the null hypothesis and the number of observations is large (n=1650) which implies that we can use the following critical value $F^{1\%}_{10,\infty} = 2.32$. *Since 1.47<2.32 we do not reject the null hypothesis at a 1% significance level.*

Question 2

A business owner wants to know if bonus payments will increase the work effort of the employees. He asks his research department to set up an experiment in order to estimate the average causal effect of bonus payments on work effort. The research department randomly assigns 500 employees either to a treatment group or a control group. The 250 employees assigned to the treatment group receive a bonus if they meet the target, the 250 employees in the control group do not get a bonus if they meet the target. The experiment lasts for 3 months and at the end of the period the research department collects information on work effort. They construct a binary variable $effort_i$ which equals one if the worker exerted high effort during the 3 months and zero if the worker exerted low effort. The data set collected by the research department contains in addition a binary variable $bonus_i$ which equals one if the worker was assigned to the treatment group and zero if assigned to the control group and the variable $female_i$ which equals one for female employees and zero for male employees.

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

$$effort_i = \beta_0 + \beta_1 \cdot bonus_i + u_i \tag{8}$$

and obtains the following estimation results

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

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.440000   0.031457 13.9872  < 2e-16
## bonus       0.096000   0.044591  2.1529  0.03181
## ---
```

Give an interpretation, in words, of the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Solution (10 points): $\widehat{\beta}_0 = 0.44$ *is share of employees in the control group that exert high effort.* $\widehat{\beta}_1 = 0.096$ *is the difference in the share of employees that exert high effort between the treatment and control group, it is the estimated average causal effect of the bonus payment on the probability of exerting high effort. The share of employees in the treatment group that exert high effort is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 0.536$

b) The experiment started during the summer holiday and all 500 workers of the firm were randomly assigned to the treatment or control group. Part of these workers are students. Midway during the experiment, the summer holiday ends and all students quit their job

and go back to school. These students are therefore not part of the data set collected by the research department. Do you think that the OLS estimator of β_1 in model (8) (estimated in part (a)) is a consistent estimator of the causal effect of bonus payments on the probability of exerting high effort? Explain why or why not.

Solution (10 points): *This is an example of attrition. The attrition is related to a predetermined characteristic (whether an employee is a student) and all workers (and thus also all workers who are students) are randomly assigned to the treatment and control group. The attrition is therefore not related to the treatment and it will therefore not (or very unlikely) result in an inconsistent OLS estimator of the causal effect of bonus payments on the probability of exerting high effort. The attrition is not a threat to internal validity, but it can be a problem for the external validity depending on whether the students are part of the population of interest or not.*

c) The business owner wants to know if men and women respond differently to bonus payments. In order to answer this question the research department decides to estimate the following regression model by OLS

$$effort_i = \beta_0 + \beta_1 \cdot bonus_i + \beta_2 \cdot female_i + \beta_3(bonus_i \times female_i) + \epsilon_i \tag{9}$$

and obtains the following estimation results

```
model2 <- lm( effort ~ bonus + female + (female*bonus), data = data2)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   0.338843   0.043202 [redacted]
## bonus         0.234800   0.061462
## female        0.196041   0.061729
## bonus:female -0.273816   0.088342
## ---
```

Give an interpretation, in words, of the estimated coefficient $\widehat{\beta}_3$. What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?

Solution (10 points): $\widehat{\beta}_3 = -0.27$ *is the estimated coefficient on the interaction term between* $bonus_i$ *and* $female_i$*, it measures the differential effect of bonus payments on the probability of exerting high effort between men and women. The estimated average casual effect of bonus payments on the probability of exerting high effort for men equals 0.23, or 23 percentage points. The estimated average causal effect of bonus payments on the probability of exerting high effort for women equals -0.039 (0.2348-0.2738), or minus 3.9 percentage points. The estimated average causal effect of bonus payments is thus 0.2738, or 27.38 percentage points, lower for women.*

d) The business owner wants to know if the probability of exerting high effort differs significantly between men and women in absence of a bonus payment. Use the results of part c) and a 5 percent significance level to answer the question of the business owner.

Solution (10 points): *In order to answer the question you should conduct a t-test with the following null and alternative hypotheses:*

$$H_0 : \beta_2 = 0 \qquad vs \qquad H_1 : \beta_2 \neq 0$$

Construct the t-statistic:

$$t = \frac{0.196 - 0}{0.0625} = 3.14$$

The absolute value of the t-statistic is bigger than 1.96 so we reject H_0*. The probability of exerting high effort differs significantly between men and women in absence of a bonus payment*

e) The business owner wants to know if the probability of exerting high effort differs significantly between men and women in case workers receive a bonus payment if they meet the target. Explain how the research department can answer the question of the business owner.

Solution (10 points): *In order to answer the question the research department can conduct a t-test with the following null and alternative hypotheses:*

$$H_0 : \beta_2 + \beta_3 = 0 \qquad vs \qquad H_1 : \beta_2 + \beta_3 \neq 0$$

Alternatively, the research department can estimate the following model, using only the workers assigned to the treatment group:

$$effort_i = \pi_0 + \pi_1 \cdot female_i + \upsilon_i$$

and perform a t-test with the following null and alternative hypotheses:

$$H_0 : \pi_1 = 0 \qquad vs \qquad H_1 : \pi_1 \neq 0$$

f) The research department decides to estimate a logit model and they obtain the following estimation results

```
logit <- glm(effort ~ bonus + female + (female*bonus),
             family = binomial(link = "logit"),
             data = data2)

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

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.66845    0.19284 -3.4663 0.0005276
## bonus         0.96519    0.26294  3.6707 0.0002419
## female        0.80822    0.26191  3.0858 0.0020299
## bonus:female -1.12148    0.36589 -3.0651 0.0021763
## ---
```

What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?

Solution (10 points): the estimated effect of bonus payments on the probability of exerting high effort for men*:*

$$\begin{aligned}\triangle Pr(\widehat{effort = 1}|female_i = 0) &= \left(1/\left(1+e^{-(-0.668+0.965)}\right)\right) - \left(1/\left(1+e^{-(-0.668)}\right)\right) \\ &= 0.574 - 0.339 \\ &= 0.235\end{aligned}$$

the estimated effect of bonus payments on the probability of exerting high effort for women*:*

$$\begin{aligned}\triangle Pr(\widehat{effort = 1}|female_i = 1) &= \left(1/\left(1+e^{-(-0.668+0.965+0.808-1.121)}\right)\right) - \left(1/\left(1+e^{-(-0.668+0.808)}\right)\right) \\ &= 0.496 - 0.535 \\ &= -0.039\end{aligned}$$

g) Does the 99 percent confidence interval around the logit coefficient on the interaction term between $female_i$ and $bonus_i$ include the value zero?

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

$$[-1.12148 - 2.58 \times 0.36589\ ,\ -1.12148 + 2.58 \times 0.36589]$$

$$[-2.065\ ,\ -0.177]$$

The confidence interval does not include the value zero.

h) The research department decides to estimate a probit model and they obtain the following estimation results

```
probit <- glm(effort ~ bonus + female + (female*bonus)
              family = binomial(link = "probit"),
              data = data2)

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

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.41562    0.11806 -3.5204 0.0004308
## bonus         0.60128    0.16238  3.7029 0.0002132
## female        0.50318    0.16201  3.1058 0.0018975
## bonus:female -0.69919    0.22752 -3.0731 0.0021182
## ---
```

What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?

Solution (10 points): the estimated effect of bonus payments on the probability of exerting high effort for men*:*

$$\begin{aligned}\triangle Pr(\widehat{effort = 1}|female_i = 0) &= \Phi(-0.416 + 0.601) - \Phi(-0.416) \\ &= 0.5714 - 0.3372 \\ &= 0.234\end{aligned}$$

the estimated effect of bonus payments on the probability of exerting high effort for women*:*

$$\begin{aligned}\triangle Pr(\widehat{effort = 1}|female_i = 0) &= \Phi(-0.416 + 0.601 + 0.503 - 0.699) - \Phi(-0.416 + 0.503 - 0.699) \\ &= 0.4960 - 0.5359 \\ &= -0.039\end{aligned}$$