Exam ECON3150/4150: Introductory Econometrics.
18 May 2016; 09:00h-12.00h.

This is an open book examination where all printed and written resources, in addition to a calculator, are allowed. 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.

Question 1

The head of department wants to investigate whether a mandatory term paper increases the probability that a student passes the exam. He has asked the econometrics teacher to set up an experiment where students are randomly assigned to a treatment group that has to write a mandatory term paper and to a control group that does not have to write a mandatory term paper. At the end of the course the teacher has a data set with information on 150 students about whether they passed the exam ($Passed_i$) and about whether they had written a mandatory term paper ($Termpaper_i = 1$) or not ($Termpaper_i = 0$). The teacher decides to estimate the following regression model by OLS

$$Passed_i = \beta_0 + \beta_1 \cdot Termpaper_i + u_i \tag{1}$$

and obtains the following estimation results

```
. regress Passed Termpaper, robust

Linear regression                               Number of obs     =        150
                                                F(1, 148)         =       4.23
                                                Prob > F          =     0.0415
                                                R-squared         =     0.0278
                                                Root MSE          =     .39706
```

Passed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
Termpaper	.1333333	.0648398	2.06	0.042	[redacted]	[redacted]
_cons	.7333333	.0514066	14.27	0.000	.6317475	.8349192

a) Give an interpretation, in words, of the two estimated coefficients.

Solution (5 points): $\widehat{\beta}_0 = 0.73$ *is the fraction of students in the control group (that did not write a term paper) that passed the exam and* $\widehat{\beta}_1 = 0.13$ *is the difference in the fraction of students that passed the exam between those that wrote a term paper, and those that did not write a term paper. The fraction of students in the treatment group (that wrote a term paper) that passed the exam is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 0.867$.

b) Compute a 95 percent confidence interval for $\widehat{\beta}_1$.

Solution (5 points): *95% confidence interval for* $\widehat{\beta}_1$*:*

$$\widehat{\beta}_1 \pm 1.96 \times SE(\widehat{\beta}_1)$$

filling in the numbers from the regression output gives

$$0.133 \pm 1.96 \times 0.065$$

$$(0.005 \quad , \quad 0.261)$$

c) The teacher wants to analyze whether the effect of the mandatory term paper depends on the age of the student. Describe in detail how you would extend model (1), such that you can test the null hypothesis that the effect of the mandatory term paper does not depend on the age of the student.

Solution (10 points): *The regression should be augmented to include an interaction term as follows:*

$$Passed_i = \beta_0 + \beta_1 \cdot TermPaper_i + \beta_2 \cdot Age_i + \beta_3 \cdot (TermPaper_i \times Age_i) + \epsilon_i$$

whereby Age_i *is the age of the student. The hypothesis can be tested by using a* t *or* F *test testing* H_0*:* $\beta_3 = 0$*.*

d) Because the dependent variable is binary the teacher decides to estimate a logit model and obtains the following estimation results.

```
. logit Passed Termpaper, robust

Iteration 0:   log pseudolikelihood =   -75.060364
Iteration 1:   log pseudolikelihood =   -72.978111
Iteration 2:   log pseudolikelihood =   -72.944236
Iteration 3:   log pseudolikelihood =   -72.944223
Iteration 4:   log pseudolikelihood =   -72.944223

Logistic regression                             Number of obs     =        150
                                                Wald chi2(1)      =       4.00
                                                Prob > chi2       =     0.0454
Log pseudolikelihood =  -72.944223              Pseudo R2         =     0.0282
```

Passed	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Termpaper	.8602013	.4298819	2.00	0.045	.0176483	1.702754
_cons	1.011601	.2619912	3.86	0.000	.4981075	1.525094

What is effect of writing a term paper on the predicted probability of passing the exam?

Solution (10 points): *The effect of writing a term paper on the predicted probability of passing the exam equals:*

$$\triangle Pr(\widehat{Passed}_i = 1) = \Pr(Passed_i = 1|\widehat{Termpaper}_i = 1) - \Pr(Passed_i = 1|\widehat{Termpaper}_i = 0)$$

$$\begin{aligned} \triangle Pr(\widehat{Passed}_i = 1) &= \left(1/\left(1 + e^{-(1.01+0.86\cdot 1)}\right)\right) - \left(1/\left(1 + e^{-(1.01+0.86\cdot 0)}\right)\right) \\ &= 0.866 - 0.733 \\ &= 0.133 \end{aligned}$$

e) The teacher finds out that 180 students enrolled in the course and 90 were randomly assigned to the treatment group and the other 90 to the control group. The teacher doesn't observe exam results for all 180 students but only for 150 students. She suspects that some of the students in the treatment group didn't want to write a term paper and dropped out of the course. Explain whether in this case the OLS estimator of β_1 (from part a)) is a consistent estimator of the causal effect of writing a mandatory term paper on the probability of passing the exam.

Solution (10 points): *This is an example of attrition. Since the attrition is related to the treatment it can lead to sample selection problems. If the less motivated students are the ones that decide to drop out of the course this implies that students in the treatment group that take the exam are on average more motivated than those in the control group that take the exam. This will result in an inconsistent OLS estimator of the causal effect of writing a mandatory term paper on the probability of passing the exam.*

f) The administration informs the teacher that 30 students got the flu and therefore did not take the exam. Explain whether in this case the OLS estimator of β_1 (from part a)) is a consistent estimator of the causal effect of writing a mandatory term paper on the probability of passing the exam.

Solution (10 points): *This is an example of attrition. Since the attrition is due to students getting the flu it is unlikely to be related to the treatment and it will therefore not lead to sample selection problems. This type of attrition will therefore not (or very unlikely) result in an inconsistent OLS estimator of the causal effect of writing a mandatory term paper on the probability of passing the exam.*

Question 2

The federal government of the U.S. wants to know whether increasing seat belt usage reduces traffic fatalities. A government employee has data on the number of traffic fatalities per million of traffic miles ($fatalityrate_{it}$) and on the seat belt useage rate ($sb\,useage_{it}$) for 51 U.S. States, for the years 1990-1997. He estimates the following regression model by OLS

$$fatalityrate_{it} = \beta_0 + \beta_1 \cdot sb\,usage_{it} + u_{it} \tag{2}$$

and obtains the following estimation results

```
. regress fatalityrate sb_useage, robust

Linear regression                               Number of obs     =        408
                                                F(1, 406)         =
                                                Prob > F          =
                                                R-squared         =     0.0309
                                                Root MSE          =     .00448

------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0060436   .0018963
       _cons |   .0220271   .0012039    18.30   0.000     .0196605    .0243937
------------------------------------------------------------------------------
```

a) Is the relation between seat belt usage and the traffic fatality rate significantly different from zero at a 5 percent significance level?

Solution (5 points):$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$. *Construct the t-statistic:*

$$t = \frac{-0.006 - 0}{0.0018} = 3.19$$

The absolute value of the t-statistic is bigger than 1.96 so we reject H_0. The relation between seat belt usage and the traffic fatality rate is significantly different from zero at a 5 percent significance level.

b) The government employee decides to include state fixed effects and obtains the following results.

```
. xtreg fatalityrate sb_useage, fe i(State) robust

Fixed-effects (within) regression               Number of obs      =       408
Group variable: State                           Number of groups   =        51

R-sq:                                           Obs per group:
     within  = 0.3010                                         min =         8
     between = 0.0033                                         avg =       8.0
     overall = 0.0309                                         max =         8

                                                F(1,50)            =     65.43
corr(u_i, Xb)  = -0.2176                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 51 clusters in State)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0131094   .0016207    -8.09   0.000    -.0163646   -.0098542
       _cons |   .0261786   .0009522    27.49   0.000      .024266    .0280912
-------------+----------------------------------------------------------------
     sigma_u |  .00433259
     sigma_e |  .00167444
         rho |  .87004661   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Are the results different from the results without fixed effects? Explain why or why not.

Solution (10 points): *Yes the estimated coefficient on seat belt usage in the regression with fixed effects is more negative and more than twice as large in absolute value. This indicates that there are time-invariant state characteristics that caused omitted variable bias in the regression without fixed effects. These characteristics are correlated with seat belt usage and affect the traffic fatality rate.*

c) Explain an alternative way to estimate the effect of seat belt usage on the traffic fatality rate while including state fixed effects.

Solution (10 points): *Instead of within estimation we can also include dummy variables for 50 states and include a constant term. Exclude 1 of the dummy variables to avoid perfect multicollinearity:*

$$fatalityrate_{it} = \beta_0 + \beta_1 \cdot sb\,usage_{it} + \gamma_2 D2_i + \ldots + \gamma_{51} D51_i + u_{it}$$

$$Di_i = 1\ for\ state\ i\ and\ 0\ otherwise$$

Alternatively we can include 51 dummies and exclude the constant term:

$$fatalityrate_{it} = \beta_1 \cdot sb\,usage_{it} + \gamma_1 D1_i + \gamma_2 D2_i + \ldots + \gamma_{51} D51_i + u_{it}$$

$$Di_i = 1\ for\ state\ i\ and\ 0\ otherwise$$

d) The government employee is worried that there are omitted variables that vary within states over time that cause omitted variable bias. He decides to use an instrumental variable approach and to use the presence of a mandatory seat belt law as instrument. The variable $primary_{it}$ is a binary variable that equals 1 if in State i in year t there is a law which makes it possible for a police officer to stop a car and ticket the driver if the officer observes that someone in the car is not wearing a seat belt. He obtains the following first stage estimation results

```
. xtreg sb_useage primary, fe i(State)

Fixed-effects (within) regression               Number of obs      =       408
Group variable:  State                          Number of groups   =        51

R-sq:                                           Obs per group:
     within  =  0.0306                                         min =         8
     between =  0.3273                                         avg =       8.0
     overall =  0.2212                                         max =         8

                                                F(   1,356)        =    [redacted]
corr(u_i, Xb)  =  -0.4552                       Prob > F           =    0.0009
```

sb_useage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
primary	.2957143	.0882233	[redacted]	0.001	.12221	.4692186
_cons	.5418801	.0142222	38.10	0.000	.5139101	.5698502
sigma_u	.0992207					
sigma_e	.08252533					
rho	.5910923	(fraction of variance due to u_i)				

Is $primary_{it}$ a weak instrument?

Solution (10 points): *Compute the first stage F-statistic. Since there is only one instrument the first stage F-statistic is the square of the the t-statistic:* $F = t^2 = \left(\frac{\widehat{\pi_1}}{SE(\widehat{\pi_1})}\right)^2 = \left(\frac{0.2957}{0.0882}\right)^2 = 11.2$. *Since 11.2 is bigger than the rule of thumb value of 10,* $primary_{it}$ *is not a weak instrument.*

e) The government employee decides to use two instruments. Next to the variable $primary_{it}$ the data set contains the variable $secondary_{it}$. The binary variable $secondary_{it}$ equals 1 if in State i in year t there is a law that makes it possible that a police officer can write a ticket if an occupant is not wearing a seat belt, but only if the police officer has another reason to stop the car. The government official estimates a first stage regression using both instruments and obtains the following first stage estimation results

```
. xtreg sb_useage primary secondary, fe i(State)

Fixed-effects (within) regression               Number of obs      =       408
Group variable:  State                          Number of groups   =        51

R-sq:                                           Obs per group:
     within  =  0.0372                                         min =         8
     between =  0.3319                                         avg =       8.0
     overall =  0.2259                                         max =         8

                                                F(   2,355)        =      6.86
corr(u_i, Xb)  =  -0.4735                       Prob > F           =    0.0012
```

sb_useage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
primary	.3055089	.0882704	3.46	0.001	.1319103	.4791074
secondary	.0035196	.0022574	1.56	0.120	-.00092	.0079591
_cons	.5380134	.0144087	37.34	0.000	.5096761	.5663506
sigma_u	.10001657					
sigma_e	.08235998					
rho	.59591494	(fraction of variance due to u_i)				

Are there weak instrument problems when using these two instrument? Explain whether it is better to use both $primary_{it}$ and $secondary_{it}$ as instruments or whether it is better to use only $primary_{it}$ as instrument for $sb\,usage_{it}$.

Solution (10 points): *Compute the first stage F-statistic. Since there are two instruments and no control variables we can use the overall F statistic:* $F = 6.86$. *Since 6.86 is smaller than the rule of thumb value of 10, there are weak instrument problems. It is better to use only* $primary_{it}$ *as instrument because in this case the first stage F-statistic is higher (higher than 10), while using both instruments results in a lower F-statistic and weak instruments problems.*

Question 3

Discuss whether each of the following statements is correct or not.

a) If we perform a J-test and we don't reject the null hypothesis we know that the instrument relevance condition holds.

Solution (5 points) *Incorrect. The J-test test the null hypothesis of instrument exogeneity, not instrument relevance. If we do not reject the null hypothesis of instrument exogeneity this does not imply that we know that instrument exogeneity holds, we can never accept the null hypothesis. We can use the first stage regression (regression of the endogenous regressor on the instrument(s)) to test for instrument relevance.*

b) If the $R^2 = 0.5$ the total sum of squares is twice as large as the sum of squared residuals.

Solution (5 points) *Correct.* $R^2 = 1 - \frac{SSR}{TSS}$ *so* $\frac{SSR}{TSS} = 1 - 0.5 \Longrightarrow TSS = \frac{SSR}{0.5} = 2 \cdot SSR$

c) When the sample size is small we can estimate the Root Mean Squared Forecast Error by the Standard Error of the Regression.

Solution (5 points) *Incorrect. The* $RMSFE = \sqrt{E\left[\left(Y_{T+1} - \widehat{Y}_{T+1|T}\right)^2\right]}$ *has two sources of error:*

1. *The error arising because future values of* u_t *are unknown*
2. *The error in estimating the coefficients*

If the sample size is large the first source of error will be (much) larger than the second and $RMSFE \approx \sqrt{Var\left(u_t\right)}$ *and* $\sqrt{Var\left(u_t\right)}$ *can be estimated by the* $SER = \frac{1}{T-2}\sum_{t=1}^{T}\widehat{u}_t^2$. *However if the sample size is small the second type of error can be large and it is not a good idea to estimate the RMSFE by the SER.*

Question 4

A researcher wants to estimate the effect of a job training program (X_i) on wages in NOK (Y_i). Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ with $E[u_i|X_i] = 0$. The researcher observes X_i but by accident wages are measured in Euro's instead of NOK. Let α be NOK/Euro exchange rate. The researcher performs an OLS regression of Y_i^* (the observed wage rate in Euro's) on X_i. The researcher wants to know the causal effect of the job training program on wages in NOK, show whether the OLS estimator is a consistent or an inconsistent estimator of this causal effect.

Solution (10 points):

$$\alpha = \frac{Y_i}{Y_i^*} \Longrightarrow Y_i^* = \frac{1}{\alpha} Y_i$$

$$\widehat{\beta_{ols}} = \frac{s_{Y^*X}}{s_X^2} \xrightarrow{p} \frac{Cov(Y_i^*, X_i)}{Var(X_i)} = \frac{Cov(\frac{1}{\alpha}Y_i, X_i)}{Var(X_i)}$$

$$= \frac{Cov(\frac{1}{\alpha}(\beta_0 + \beta_1 X_i + u_i), X_i)}{Var(X_i)}$$

$$= \frac{Cov\left(\frac{1}{\alpha}\beta_0, X_i\right) + Cov\left(\frac{1}{\alpha}\beta_1 X_i, X_i\right) + Cov\left(\frac{1}{\alpha}u_i, X_i\right)}{Var(X_i)}$$

$$= \frac{0 + \frac{1}{\alpha}\beta_1 Cov(X_i, X_i) + 0}{Var(X_i)}$$

$$= \frac{1}{\alpha}\beta_1$$

If $\alpha \neq 1$ the OLS estimator of β_1 is an inconsistent estimator of the causal effect of the job training program on wages in NOK.