**Exam ECON4150: Introductory Econometrics.**
**11 May 2015; 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. In the grading, each sub-question will count for 1/12$^{th}$ of the total grade.*

**Guideline for correctors:** *In this exam a total of 120 points can be obtained. With each sub-question a maximum of 10 points can be obtained.*

## Question 1

One of the oldest questions in labor economics is the effect of union membership on wages. In order to investigate this question a researcher uses data on 5000 individuals observed in the years 2000-2010 with information on their hourly wage ($Wage_{it}$) in year $t$, their years of completed education ($Education_i$) and whether or not they are member of a union in year $t$ ($Union_{it}$). The researcher estimates the following equation by OLS

$$ln(Wage_{it}) = \beta_0 + \beta_1 Union_{it} + \beta_2 Education_i + u_{it}$$

and obtains the following regression results

```
. regress ln_Wage Union Education, robust

Linear regression                               Number of obs =      55000
                                                F(  2, 54997) =       4.08
                                                [redacted]
                                                Root MSE      =     .89992
```

| ln_Wage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Union | .0237997 | .0091306 | 2.61 | 0.009 | .0059036 | .0416957 |
| Education | .0152618 | .0132334 | 1.15 | 0.249 | -.0106757 | .0411992 |
| _cons | 2.346704 | .0078609 | 298.53 | 0.000 | 2.331296 | 2.362111 |

**a)** Interpret the sign and magnitude of the coefficient on $Union_{it}$.

**Solution (10 points):** *The coefficient on $Union_{it}$, $\widehat{\beta}_1 = 0.024$. It is a log-linear model and $Union_{it}$ is a binary variable, this implies that the coefficient can be interpreted as follows; being a member of a union is associated with an increase in the hourly wage by about 2.4 percent.*

**b)** Test the null hypothesis that the coefficients on $Union_{it}$ and $Education_i$ are both equal to zero using a 1 percent significance level.

**Solution (10 points):** *Test* $H_0: \quad \beta_1 = \beta_2 = 0$ *versus* $H_1: \quad \beta_1 \neq 0$ *and/or* $\beta_2 \neq 0$. *This is a joint null hypothesis which means that we should an F-test. The F-statistic is given in the Stata output and equals* $F = 4.08$. *The critical value of the F-statistic equals* $F^{1\%}_{2,\infty} = 4.61$. *Since the value of F-statistic is lower than the critical value we do not reject the null hypothesis at a 1% significance level.*

**c)** The researcher is mainly interested in the effect of union membership on wages. Describe one potential threat to the internal validity of the current regression results.

**Solution (10 points):** *A potential threat to the internal validity is omitted variable bias. Individuals that are a member of a union might be more motivated or of higher ability and would earn more than non-members even if they would not be a union member. Another potential threat to internal validity is measurement error; union membership might be measured with error.*

**d)** The researcher decides to augment the model by including individual fixed effects and obtains the following estimation results.

```
. xtreg ln_Wage Union Education, fe i(id) robust

Fixed-effects (within) regression               Number of obs      =         55000
Group variable:  id                             Number of groups   =          5000

R-sq:  within  =  0.0045                        Obs per group: min =            11
       between =  0.3087                                       avg =          11.0
       overall =  0.0375                                       max =            11

                                                F(1,4999)          =        245.03
corr(u_i, Xb)  =  0.2693                        Prob > F           =        0.0000

                                    (Std. Err. adjusted for    5000 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     ln_Wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Union |   .0232507   .0014853    15.65   0.000     .0203388    .0261626
   Education |           0  (omitted)
       _cons |   2.070391     .00034  6089.28   0.000     2.069724    2.071057
-------------+----------------------------------------------------------------
     sigma_u |  .06462443
     sigma_e |  .12884386
         rho |  .20100616   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Explain why the variable $Education_i$ is omitted from the regression.

**Solution (10 points):** *The variable $Education_i$ measures completed years of education and this variable does not vary over time for individuals that work (it does not have a subscript t) it is therefore perfectly multicollinear with the individual fixed effects. It is not possible to estimate a model that includes individual fixed effects and $Education_i$, Stata therefore omits the variable $Education_i$ from the regression.*

**Question 2**

A researcher wants to know whether a preschool program for immigrant children has an effect on their future educational attainment. In 1990 the government set up an experiment where 1000 (5-year old) immigrant children were randomly assigned to a treatment group (the preschool program) and to a control group (no preschool program), 25 years later the researcher uses data on these 1000 immigrant children (who are now 30 years old) to investigate the effect of participation in the preschool program ($preschool_i$) on years of completed education ($education_i$). The researcher decides to estimate the following regression model by OLS

$$education_i = \beta_0 + \beta_1 \cdot preschool_i + u_i \tag{1}$$

and obtains the following regression results

```
. regress education preschool, robust

Linear regression                               Number of obs =        1000
                                                F(  1,   998) =      169.11
                                                Prob > F      =      0.0000
                                                R-squared     =      0.1449
                                                Root MSE      =      1.1155
```

| education | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| preschool | .9187161 | .0706472 | 13.00 | 0.000 | .7800821 | 1.05735 |
| _cons | 12.76565 | .0512387 | 249.14 | 0.000 | 12.6651 | 12.8662 |

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

**Solution (5 points):** $\widehat{\beta}_0 = 12.77$ *is the average number of years of schooling of the children who did not participate in the preschool program and* $\widehat{\beta}_1 = 0.92$ *is the difference between the average years of schooling of the children who participated in the preschool program and the average years of schooling of those who did not participate. The average years of schooling of the children who participated in the preschool program is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 13.68$

**b)** The researcher wants to analyze whether there is a difference in the effect of the preschool program between boys and girls. Describe in detail how you would extend model (1), such that you can test the null hypothesis that the effect of the preschool program on years of education does not depend on gender.

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

$$education = \delta_0 + \delta_1 preschool_i + \delta_2 girl_i + \delta_3 (preschool_i \times girl_i) + \epsilon_i$$

*whereby* $girl_i$ *equals one for girls and zero for boys. The hypothesis can be tested by using a* $t$ *or* $F$ *test testing* $H_0$*:* $\delta_3 = 0$.

**c)** The researcher also wants to know whether participating in the preschool program increases the likelihood of having a job at age 30. The data set contains two additional variables; $employed_i$ which is equal to one if an individual has a job in 2015 and is zero otherwise and $girl_i$ which is equal to one for girls and zero for boys. The researcher estimates the following regression model

$$employed_i = \beta_0 + \beta_1 \cdot preschool_i + \beta_2 \cdot girl_i + u_i \tag{2}$$

and obtains the following estimation results

```
. regress employed preschool girl, robust noheader
```

| employed | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| preschool | .1371742 | .0290887 | 4.72 | 0.000 | .0800921 | .1942564 |
| girl | .3726898 | .0289697 | 12.86 | 0.000 | .3158413 | .4295384 |
| _cons | .2655838 | .0251674 | 10.55 | 0.000 | .2161966 | .314971 |

On the basis of these estimation results, what is the change in the probability of being employed that is associated with participating in the preschool program *for boys*?

**Solution (10 points):** *It is a linear probability model which means that the coefficient on the variable preschool measures the change in the probability of being employed that is associated with participating in the preschool program. Since the model does not include an interaction term the estimated change in the probability of being employed that is associated with participating in the preschool program does not depend on gender. This means that the estimated change in the probability of being employed that is associated with participating in the preschool program for boys equals 0.137 (or 13.7%).*

**d)** The researcher also estimates a logit model and obtains the following estimation results

```
. logit employed preschool girl, robust noheader

Iteration 0:   log pseudolikelihood =   -692.80914
Iteration 1:   log pseudolikelihood =   -613.31915
Iteration 2:   log pseudolikelihood =   -613.13556
Iteration 3:   log pseudolikelihood =   -613.13555
```

| employed | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| preschool | .6451726 | .1396817 | 4.62 | 0.000 | .3714015 | .9189436 |
| girl | 1.609523 | .1398102 | 11.51 | 0.000 | 1.3355 | 1.883546 |
| _cons | -1.030281 | .1265701 | -8.14 | 0.000 | -1.278354 | -.7822083 |

On the basis of these estimation results, what is the change in the probability of being employed that is associated with participating in the preschool program *for boys*?

**Solution (10 points):** *On the basis of these estimation results the change in the probability of being employed that is associated with participating in the preschool program for boys is equal to 0.142 or (14.2%).*

$$\Pr(employed_i = 1|\widehat{preschool}_i = 1,\ girl_i = 0) \ - \ \Pr(employed_i = 1|\widehat{preschool}_i = 0,\ girl_i = 0)$$
$$= \frac{1}{1+e^{-(-1.03+0.645)}} - \frac{1}{1+e^{-(-1.03)}}$$
$$= 0.4049 - 0.2630$$
$$= 0.1418$$

**e)** Some children who were assigned to the control group did participate in the preschool program, because their parents managed to convince the preschool teacher to let their child participate. Do you think that the OLS estimator of $\beta_1$ in model (2) (estimated in part (c)) is a consistent estimator of the causal effect of participating in the preschool program on the probability of being employed at age 30? Explain why or why not.

**Solution (10 points):** *This is an example of failure to follow the randomized treatment protocol or partial compliance. Although assignment to the treatment or control group is random, participating in the preschool program is not random and likely related to unobserved variables that affect employment. An example of such an unobserved variable is motivation of the parents. If motivated parents decide to send their child to preschool although they are assigned to the control group (because they think it is important for their child) and motivated parents also do other things to increase the probability that their child has a job at age 30 (such as investing more in the child's education), the OLS estimator will be inconsistent (omitted variable bias).*

**f)** The researcher decides to use an instrumental variable approach to estimate the effect of participating the preschool program on the probability of being employed at age 30. He uses the assignment to the treatment ($assignment_i = 1$) or control group ($assignment_i = 0$) as an instrument for whether or not a child participated in the preschool program. The researcher obtains the following first stage OLS estimates.

```
. regress preschool assignment girl, robust

Linear regression                               Number of obs =        1000
                                                F(  2,   997) =       30.38
                                                Prob > F      =      0.0000
                                                R-squared     =      0.0571
                                                Root MSE      =      .48567
```

| preschool | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| assignment | .2349113 | .0307472 | 7.64 | 0.000 | .1745745 | .295248 |
| girl | -.0362907 | .0308057 | -1.18 | 0.239 | -.0967422 | .0241607 |
| _cons | .4236373 | .0267232 | 15.85 | 0.000 | .3711972 | .4760774 |

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

**Solution (10 points):** *Instrument relevance,* $Corr(preschool_i, assignment_i) \neq 0$ *can be investigated using the first stage regression. The first stage F-statistic equals* $F = (t)^2 = (7.64)^2 = 58.4$*, which is bigger than the rule-of-thumb value of 10. The instrument relevance condition holds and* $assignment_i$ *is a not a weak instrument.*

**Question 3**

Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ with $Cov(X_i, u_i) = 0$. The researcher observes $Y_i$ but does not observe $X_i$, instead he observed a noisy measure $X_i^* = X_i + \varepsilon_i$ , where $\varepsilon_i = \varepsilon$ (it is identical for all $i$). The researcher has a large sample with i.i.d observations on $Y_i$ and $X_i^*$ and estimates the following equation by OLS

$$Y_i = \beta_0 + \beta_1 X_i^* + v_i$$

**a)** What is $Cov(X_i^*, v_i)$?

**Solution (10 points; 5 points for obtaining covariance expression, another 5 for knowing that all terms are 0):**

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 (X_i^* - \varepsilon_i) + u_i \\ &= \beta_0 + \beta_1 X_i^* + (u_i - \beta_1 \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i^* + v_i \end{aligned}$$

This implies that $v_i = u_i - \beta_1 \varepsilon_i$

$$\begin{aligned} Cov(X_i^*, v_i) &= Cov(X_i + \varepsilon_i,\ u_i - \beta_1 \varepsilon_i) \\ &= Cov(X_i, u_i) + Cov(X_i, -\beta_1 \varepsilon_i) + Cov(\varepsilon_i, u_i) + Cov(\varepsilon_i, -\beta_1 \varepsilon_i) \\ (5\,points) \quad &= Cov(X_i, u_i) - \beta_1 Cov(X_i, \varepsilon_i) + Cov(\varepsilon_i, u_i) - \beta_1 Var(\varepsilon_i) \\ (5\,points) \quad &= 0 + 0 + 0 + 0 \\ &= 0 \end{aligned}$$

Since $\varepsilon_i = \varepsilon$ for all $i$; $Cov(X_i, \varepsilon_i) = Cov(\varepsilon_i, u_i) = Var(\varepsilon_i) = 0$, in addition $Cov(X_i, u_i) = 0$ as given in the exercise.

**b)** Is the OLS estimator of $\beta_1$ consistent?

**Solution (10 points):**

$$\widehat{\beta}_1 = \frac{s_{X^*Y}}{s^2_{X^*}} \xrightarrow{p} \frac{Cov(X_i^*, Y_i)}{Var(X_i^*)}$$

$$\begin{aligned}
\widehat{\beta}_1 \xrightarrow{p} \frac{Cov(X_i^*, Y_i)}{Var(X_i^*)} &= \frac{Cov(X_i^*, \beta_0 + \beta_1 X_i^* + v_i)}{Var(X_i^*)} \\
&= \frac{Cov(X_i^*, \beta_0) + \beta_1 Cov(X_i^*, X_i^*) + Cov(X_i^*, v_i)}{Var(X_i^*)} \\
&= \beta_1 + \frac{Cov(X_i^*, v_i)}{Var(X_i^*)} \\
&= \beta_1
\end{aligned}$$

As shown in part a) $Cov\left(X_i^*, v_i\right) = 0$. This means that the OLS estimator of $\beta_1$ is consistent.