UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Exam: **ECON4136 – Applied statistical analysis for the social sciences**

Date of exam: Wednesday, December 18, 2013 **Grades are given: January 6, 2014**

Time for exam: 09.00 a.m. – 12.00 noon

The problem set covers 5 pages (incl. cover sheet)

Resources allowed:

- All written and printed resources, as well as calculator are allowed

The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

Exam ECON4136 – Fall 2013

1. You are interested in the relationship between income and education for a sample of 21 to 35-year-old women, and plan to estimate the following equation using OLS:

$$incomem_i = \beta_0 + \beta_1 educm_i + \epsilon_i \tag{1}$$

where $incomem_i$ is income from work in \$, and $educm_i$ is years of schooling.

Consider the following descriptive statistics and regression results

```
. sum incomem morekids educm

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     incomem |    689200    8170.689     11362.6          0   304810.5
    morekids |    689200    .2459228    .4306333          0          1
       educm |    689200    12.32615    2.424783          0         20

. reg incomem educm

      Source |       SS       df       MS              Number of obs =  689200
-------------+------------------------------           F(  1,689198) =21841.84
       Model |  2.7334e+12     1  2.7334e+12           Prob > F      =  0.0000
    Residual |  8.6248e+13689198   125142940           R-squared     =  0.0307
-------------+------------------------------           Adj R-squared =  0.0307
       Total |  8.8982e+13689199   129108743           Root MSE      =   11187

------------------------------------------------------------------------------
     incomem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educm |   821.3014   5.557224   147.79   0.000     810.4095    832.1934
       _cons |  -1952.798   69.81201   -27.97   0.000    -2089.628   -1815.969
------------------------------------------------------------------------------
```

(a) Interpret the coefficient on `educm` and calculate the 90% confidence interval.

(b) Under what condition can we give the coefficient on `educm` a causal interpretation? Do you think this condition holds in practice? Discuss briefly.

You decide to extend your specification by adding a control for whether women have more than 2 children:

$$incomem_i = \beta_0 + \beta_1 educm_i + \beta_2 morekids_i + \epsilon_i \tag{2}$$

where $morekids_i$ is a dummy variable that equals 1 if the mother has 3 kids or more.

```
. reg incomem educm morekids

      Source |       SS       df       MS              Number of obs =  689200
-------------+------------------------------           F(  2,689197) =14669.85
       Model |  3.6333e+12     2  1.8167e+12           Prob > F      =  0.0000
    Residual |  8.5348e+13689197   123837259           R-squared     =  0.0408
-------------+------------------------------           Adj R-squared =  0.0408
```

```
       Total |  8.8982e+13689199   129108743           Root MSE      =   11128

------------------------------------------------------------------------------
     incomem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educm |   743.8146   5.602383   132.77   0.000     732.8341    754.7951
    morekids |   -2689.26   31.54555   -85.25   0.000    -2751.088   -2627.432
       _cons |  -336.3333   71.98892    -4.67   0.000    -477.4292   -195.2373
------------------------------------------------------------------------------
```

(c) What is the correlation between `morekids` and `educm`?

(d) Interpret the coefficient on `educm`.

(e) Suppose $E[\epsilon|morekids, educm] = E[\epsilon|morekids]$. What does this imply for the causal interpretation of your estimated coefficients?

(f) Suppose `educm` was initially randomly assigned. Discuss why (or why not) you may want to control for `morekids`.

You decide to add an interaction between `educm` and `morekids`:

```
. gen morekidseducm = morekids * educm

. reg incomem educm morekids morekidsedu

      Source |       SS       df       MS              Number of obs =  689200
-------------+------------------------------           F(  3,689196) =10236.90
       Model |  3.7959e+12     3  1.2653e+12           Prob > F      =  0.0000
    Residual |  8.5186e+13689196   123601593           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0427
       Total |  8.8982e+13689199   129108743           Root MSE      =   11118

------------------------------------------------------------------------------
     incomem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educm |   869.3505   6.581079   132.10   0.000     856.4518    882.2492
    morekids |   2705.309   152.0607    17.79   0.000     2407.275    3003.343
morekidseducm |  -453.7047   12.51123   -36.26   0.000    -478.2263   -429.1831
       _cons |   -1911.91   84.02523   -22.75   0.000    -2076.597   -1747.224
------------------------------------------------------------------------------

// the covariance matrix of the estimated coefficients
. mat l e(V)

                  educm          morekids   morekidseducm         _cons
       educm      43.3106
    morekids    543.58324       23122.471
morekidseducm     -43.3106      -1861.1575       156.53076
       _cons    -543.58324      -7060.2395       543.58324     7060.2395
```

(g) Interpret the coefficient on the interaction `morekidseducm`

(h) Test the null hypothesis that the return to schooling for women with more than 2 children equals zero.

2. Angrist and Evans (1998, AER), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" are interested in the impact of children on mothers' labor supply. They estimate equations of the following type:

$$incomem_i = \beta_0 + \beta_1 morekids_i + \epsilon_i \tag{3}$$

where $incomem_i$ is a mother's income from work, and $morekids_i$ is a dummy variable that equals 1 if the mother has 3 kids or more. The worry here is that $morekids_i$ and ϵ_i are correlated. Angrist and Evans note that parents who have two children with the same sex after the first two births – two boys or two girls – are more likely to have a third child. They propose to use $samesex_i$, which equals 1 if a mother has two boys or two girls after the first two births and is 0 otherwise, as an instrumental variable for $morekids_i$.

The following table shows data from the 1980 US census for mothers aged 21 to 35 who have at least two children. Use this table (when necessary) to answer the questions below.

```
. tabulate morekids samesex if kidcount>=2 & inrange(agem, 21, 35 ), s(incomem) nost

        Means (top) and Frequencies (bottom) of mothers labor income

           |       samesex
  morekids |         0          1 |     Total
-----------+----------------------+----------
         0 |   7733.27    7783.11 |   7757.43
           |   157,234    147,898 |   305,132
-----------+----------------------+----------
         1 |   5666.69    5599.68 |   5630.38
           |    77,658     91,832 |   169,490
-----------+----------------------+----------
     Total |   7050.03    6946.72 |   6997.85
           |   234,892    239,730 |   474,622
```

(a) Assume β_1 is the same for all women. Discuss the validity of the $samesex_i$ instrument in the context of equation (3).

(b) Calculate the first stage estimate and interpret this coefficient in words.

(c) If you were to judge instrument relevance, how would you do that?

(d) Calculate the reduced form effect and interpret this coefficient in words.

(e) Calculate the IV estimate of the effect of having more than 2 kids on mothers' income and interpret this coefficient in words, assuming β_1 is the same for all women.

Suppose now that the effect of interest is heterogeneous: $\beta_1 = \beta_{1i}$.

(f) What assumptions does an instrument need to fulfill if we are interested in estimating a local average treatment effect.

(g) Explain, in the context of this application, who the compliers are and how this affects your interpretation of the IV estimate.

(h) What are the fractions of compliers, always takers and never takers?

(i) What counterfactual outcomes for compliers, never-takers and always-takers are identified?

(j) Calculate all possible average counterfactual outcomes for compliers, never-takers and always-takers.

3. The new government is planning to invest 300 million NOK in the further training of school teachers, and promises that 10,000 math teachers will receive extra training in the next five years.

   You are asked to estimate the causal effect of the training on pupils' achievement. Line out and motivate your preferred estimation strategy for the following cases, and highlight important assumptions and limitations.

   (a) Capacity for training is limited, and the government decides to allocate the available training slots on the basis of teacher age. In the first year the 5,000 oldest teachers are offered training slots. In the next year, the 5,000 teachers who follow in age are offered slots, and so on.

   (b) Suppose the government allocates the available training funds to municipalities. Because the government's annual budget for this program is 60 million, it stages the implementation so that some municipalities receive the funds in the first year, other municipalities in the second year, and so on.