UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

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

Date of exam: Friday, November 23, 2012 **Grades are given: 18. December 2012**

Time for exam: 09:00 a.m. – 12:00 noon

The problem set covers 6 pages (incl. cover page)

Resources allowed:

- All written and printed resources, including calculator, is allowed.

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

Exam ECON4136

1. Consider the following three variables

```
. sum lbwght  lfaminc  fatheduc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lbwght |      1192    4.767457    .1879538   3.135494   5.602119
     lfaminc |      1192    3.275699    .7157919  -.6931472   4.174387
    fatheduc |      1192    13.18624    2.745985          1         18
```

where -lbwght- is the log of birth weight of a child (in ounces), -lfaminc- is the log of family income at birth, and -fatheduc- is the education of the father of the child in years

You are interested in the relationship between (family) income and birthweight and estimate the following regression

$$lbwght = \beta_0 + \beta_1 lfaminc + u \tag{1}$$

which gives you the following results:

```
. reg  lbwght  lfaminc

      Source |       SS       df       MS              Number of obs =    1192
-------------+------------------------------           F(  1,  1190) =    7.16
       Model |  .251709024     1  .251709024           Prob > F      =  0.0075
    Residual |  ?????????? 1190  .035144808           R-squared     =  ??????
-------------+------------------------------           Adj R-squared =  0.0051
       Total |    42.07403  1191  .035326641           Root MSE      =  .18747

------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lfaminc |   .0203099   .0075891      2.7   0.008     ????????    ????????
       _cons |   4.700928   .0254456    184.7   0.000     4.651004    4.750851
------------------------------------------------------------------------------
```

(a) What is the R-squared of this regression? How do you interpret it?

(b) What is the residual sum of squares?

(c) What is the 95% confidence interval for the coefficient on -lfaminc-? Can you reject at the 5% level that the coefficient on -lfaminc- equals 0.035?

(d) What is the t-value corresponding to the null-hypothesis that the intercept equals 5?

(e) What is the estimate of the residual variance?

(f) Interpret the coefficient on -lfaminc-

(g) 1 ounce is about 28.35 grams. If you would measure birth weight in grams, how would this affect your estimates?

The following regression adds the education of the father in years (-fatheduc-) as a regressor

```
. reg  lbwght  lfaminc  fatheduc

      Source |       SS       df       MS              Number of obs =    1192
-------------+------------------------------           F(  2,  1189) =    4.91
       Model |  .34438718     2  .17219359           Prob > F      =  0.0076
    Residual |  41.7296428  1189  .035096419          R-squared     =  0.0082
-------------+------------------------------           Adj R-squared =  0.0065
       Total |    42.07403  1191  .035326641          Root MSE      =  .18734

------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
     lfaminc |   .0147377    .008323      1.8   0.077    -.0015917    .0310671
    fatheduc |   .0035255   .0021695      1.6   0.104     -.000731    .0077821
       _cons |   4.672692   .0307978    151.7   0.000     4.612267    4.733116
------------------------------------------------------------------------------
```

(h) What is the correlation between -lfaminc- and -fatheduc-?

We change the specification and now regress the birth weight (in ounces) on -lfaminc- and -fatheduc- and estimate

$$bwght = \beta_0 + \beta_1 lfaminc + \beta_2 fatheduc + u$$

```
. reg bwght lfaminc fatheduc

      Source |       SS       df       MS              Number of obs =    1192
-------------+------------------------------           F(  2,  1189) =    5.85
       Model |  4703.73859     2  2351.86929           Prob > F      =  0.0030
    Residual |  478199.818  1189  402.186558           R-squared     =  0.0097
-------------+------------------------------           Adj R-squared =  0.0081
       Total |  482903.556  1191  405.460585           Root MSE      =  20.055

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lfaminc |     1.59812    .890968      1.8   0.073    -.1499252    3.346164
    fatheduc |   .4445568   .2322474      1.9   0.056    -.0111035    .9002172
       _cons |   108.4223   3.296877     32.9   0.000      101.954    114.8906
------------------------------------------------------------------------------

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

symmetric e(V)[3,3]
             lfaminc    fatheduc       _cons
 lfaminc    .79382406
fatheduc   -.08525025   .05393884
   _cons   -1.4761984  -.43199644   10.869396
```

(i) Interpret the coefficient on -lfaminc-

(j) The R-squared increased compared to the specification in (g). What do you conclude?

(k) A friend says that you cannot interpret your estimates in a causal manner because the R-squared is too low. What do you reply?

(l) "10 % extra family income is equivalent to a year of extra paternal education." What is the corresponding formal null hypothesis?

(m) Calculate the Wald statistic corresponding to $H_0 : \beta_1 - \beta_2 = 1$

(n) You want to test whether income has the same effect on birthweight for kids with fathers who have completed high school (fatheduc>=12) compared to children with fathers who do not have a high school diploma (fatheduc<12). Explain i) the specification of the regression that you will estimate, and ii) the null hypothesis you will be testing.

2. You are still interested in the population relationship between (family) income and birthweight in

$$lbwght = \beta_0 + \beta_1 lfaminc + u \tag{2}$$

but you want to estimate the causal effect of $lfaminc$ on $lbwght$

You think that unobserved family effects α_f may bias your OLS estimates. You therefore want to estimate the following equation

$$lbwght_{if} = \beta_0 + \beta_1 lfaminc_{if} + \alpha_f + u_{if} \tag{3}$$

where i indexes individuals and f families.

(a) Would you prefer fixed or random effects estimates? Explain.

(b) What type of data do you need to estimate (3)?

You decide to use instrumental variables (IV) to estimate β_1. You think that you have an instrument z. Now

$$\sum_i z_i x_i = 52452 \quad \sum_i z_i y_i = 74981 \quad \sum_i y_i x_i = 18628$$
$$\sum_i x_i = 3905 \quad \sum_i z_i = 15718 \quad \sum_i y_i = 5683 \quad N = 1192$$

where $x = lfaminc$ and $y = lbwght$

(c) What is the IV estimate of β_1?

(d) The instrument used in (c) was $fatheduc$ (see above). Do you think that this is a good instrument? Discuss.

(e) Assume now that you have two instruments, z_1 and z_2. Explain precisely what regressions you will run to estimate β_1 using two-stage least squares (2SLS).

(f) You want to test the null hypothesis that the exclusion restriction of z_2 is valid. What test would you use? What crucial assumption does this test rely on? Explain the intuition behind this test.

Suppose that people report their income with error:

$$lfaminc^* = lfaminc + v$$

where v is i.i.d., and independent of u and $lfaminc$.

With your data you would therefore be estimating the following regression

$$lbwght = \alpha_0 + \alpha_1 lfaminc^* + u' \tag{4}$$

(g) Explain how this affects the consistency of your IV estimate in (c).

3. At the end of this exercise, you can find a data description for an extract of the 1991 CPS, the US labor force survey, with information on wife and husband earnings, along with family demographic information.

You are interested in the relationship between union membership and wages for women. To start your investigation you estimate a logit of union membership for the women in your dataset:

```
. logit union educ exp* black kid*

Logistic regression                                     Number of obs   =       3558
                                                        LR chi2(6)      =     120.70
                                                        Prob > chi2     =     0.0000
Log likelihood = -1444.1694                             Pseudo R2       =     0.0401

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1842391   .0211966     8.69   0.000     .1426945    .2257838
       exper |   .1119823   .0243188     4.60   0.000     .0643183    .1596464
     expersq |  -.0017761   .0005567    -3.19   0.001    -.0028674   -.0006849
       black |   .5744795   .1726811     3.33   0.001     .2360307    .9129283
      kidge6 |   .1291035   .1235557     1.04   0.296    -.1130612    .3712682
      kidlt6 |   .3730176   .1474664     2.53   0.011     .0839888    .6620463
       _cons |  -5.801925    .410292   -14.14   0.000    -6.606083   -4.997768
------------------------------------------------------------------------------
```

(a) What is the marginal effect (at the sample average) of education? What is the marginal effect (at the sample average) of experience for a woman with 10 years of experience?

(b) You want to estimate the equivalent linear probability model in Stata. Give the complete command you would use.

You think that wages are set differently for union and non-union members:

$$\begin{aligned} y_i^{\text{union}} &= x_i\beta_1 + \epsilon_{i1} \\ y_i^{\text{non-union}} &= x_i\beta_2 + \epsilon_{i2} \end{aligned}$$

and that the wage you observe is determined as follows

$$y_i = \begin{cases} y_i^{\text{union}} & \text{if } I_i^* > 0 \\ y_i^{\text{non-union}} & \text{otherwise} \end{cases}$$

where people's choice of union membership is determined by

$$I_i^* = x_i\gamma + z_i\delta + \epsilon_{i3}$$

you also assume that

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim \mathcal{N}(0, \Omega) \qquad \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{pmatrix}$$

(c) What is $E[y|\text{union member}]$? and $E[y|\text{no union member}]$?

(d) How would you estimate β_1 and β_2 using a two-step method?

(e) Write down the likelihood function for the above model.

(f) Do you need any normalization in Ω to estimate the model? Explain.

(g) You plan to use the union status of the woman's husband as the exclusion restriction z. Discuss this choice.

```
. d

Contains data from cps91.dta
  obs:         5,634
 vars:            26                          11 Nov 2012 21:35
 size:       281,700
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
husunion        byte   %8.0g                  =1 if hus. in union
kidge6          byte   %8.0g                  =1 if child >= 6
earns           float  %8.0g                  wife's weekly earnings
age             byte   %8.0g                  wife's age
black           byte   %8.0g                  =1 if wife black
educ            byte   %8.0g                  wife's yrs schooling
union           byte   %8.0g                  =1 if wife in union
exper           byte   %8.0g                  age - educ - 6
kidlt6          byte   %8.0g                  =1 if child < 6
hours           int    %9.0g                  wife's weekly hours
expersq         int    %8.0g                  exper^2
hrwage          float  %9.0g                  earns/hours
lwage           float  %9.0g                  log(hrwage)
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    husunion |      4148    .2324012    .4224143          0          1
```

```
      kidge6 |       5634    .3075967    .4615396           0           1
       earns |       5634     232.833    263.3265           0      2884.5
         age |       5634    39.42758     9.98761          18          59
       black |       5634    .0573305    .2324937           0           1
        educ |       5634    12.98403    2.615436           0          18
       union |       3558    .1500843    .3572042           0           1
       exper |       5634    20.44391    10.44549           0          52
      kidlt6 |       5634    .2793752     .448732           0           1
       hours |       5634    20.72222    19.39618           0         120
    expersq |       5634    527.0424    468.2888           0        2704
      hrwage |       3286    10.36721    7.034759   .0333333         200
       lwage |       3286    2.195801    .5250468  -3.401197    5.298317
```