UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Postponed exam: **ECON4136 – Applied statistical analysis for the social sciences**

Date of exam: Tuesday, January 14, 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, is allowed

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

Postponed exam ECON4136 – Fall 2013

In this exercise, we will consider how workplace smoking bans affect the incidence of smoking. We have data on 10,000 US indoor workers from 1991 to 1993 used in Evans, Farrelly and Montgomery, *American Economic Review,* 89(4), 1999: p. 728–747. Basic information about the data is in the output at the end of the exercise, along with all output that is needed to solve the exercise.

1. In the output, we provide summary statistics for workers at workplaces with and without smoking bans separately, as well as results from two alternative regressions of -smoker- on -smkban-.

   (a) Interpret the coefficient on -smkban- from the first (bivariate) regression, and explain what assumptions you need to interpret this as the causal effect of smoking bans on smoking.

   (b) Use the output to calculate the standard error, the t-value and a 90 % confidence interval for the coefficient on -smkban-.

   (c) What is the OLS estimate of the coefficient on -smkban- if we omit the constant?

   (d) Results from an alternative regression model including controls for age, education, race and gender is also provided in the output. Compare the estimates on -smkban-. Explain why you think including covariates is or is not important here.

2. Assume that the regression model in (1a) above is correctly specified.

   (a) Show that the model is heteroskedastic, and explain how this affects your interpretation of the results (consequences for coefficient estimates and standard errors).

   (b) Explain precisely how we can use feasible GLS to get a more efficient estimator.

   (c) How would you take heteroskedasticity into account if you were unable or unwilling to assume that the regression model is correctly specified?

3. The above estimates were from a linear model. The output below gives the results from a probit regression of -smoker- on -smkban- and control variables.

   (a) Write down a probit model corresponding to the full model including covariates.

   (b) Explain why we might prefer a probit model to the linear model in this application?

   (c) Interpret the coefficient on -smkban- from the probit model and compare it to the previous estimate.

(d) What is the estimator for the average partial effect of -age- and of -smkban- in the linear regression model and in the probit model?

(e) Explain precisely how you would calculate the average partial effects for the full sample, and separately for high school graduates and college attendees. How would you calculate the standard error of the average partial effects?

4. You are concerned that -smkban- is not exogenous. You learn that workplaces with an average age above 50 are legally required to enforce smoking bans, and decide to estimate the following linear regression model

$$smkban = \alpha + \zeta \cdot D\left(\overline{age} > 50\right) + f\left(\overline{age}\right) + u$$

where $D\left(\overline{age} > 50\right)$ is a dummy variable equal to one if average age at the workplace is above 50, and $f\left(\overline{age}\right)$ is a flexible continuous function in average age at the workplace.

(a) How do you interpret an estimate of ζ that is 0.8?

(b) How would you estimate the effect of smoking bans on smoking?

(c) Explain why this may give a more reliable estimate of how smoking bans affect smoking than the estimate in (1d).

(d) Your fellow students worry that other variables may also be important for smoking, and therefore want you to include control variables in the regression. Explain why you think including covariates is or is not important here.

(e) Name one or two key threats to the identification strategy used here. How could you in practice investigate whether these are likely to be real or not?

(f) In practice, you need to specify the form of the function $f(\cdot)$. One alternative is to use local polynomial regression. Explain how local linear regression estimates a nonlinear relationship between -smoker- and -age-.

(g) Local linear regression requires you to choose an appropriate bandwidth. What is the tradeoff you need to consider? What are some ways to use the data to help you decide on the bandwidth?

```
Variable      Obs Unique      Mean   Min   Max  Label
------------------------------------------------------------------------------
smoker      10000      2    .2423     0     1  =1 if a current smoker
smkban      10000      2    .6098     0     1  =1 if there is a work area smoking bans
age         10000     65  38.6932    18    88  age in years
black       10000      2    .0769     0     1  =1 if black
female      10000      2    .5637     0     1  =1 if female
highschool  10000      2    .3266     0     1  =1 if high school graduate
college     10000      2    .4774     0     1  =1 if attended college
------------------------------------------------------------------------------

. sum if smkban == 0,

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      smoker |      3902    .2895951    .4536326          0          1
      smkban |      3902           0           0          0          0
         age |      3902    38.08713    12.49925         18         81
       black |      3902    .0745771    .2627415          0          1
      female |      3902    .4923116     .500005          0          1
  highschool |      3902    .3721169    .4834313          0          1
     college |      3902     .426448    .4946239          0          1

. sum if smkban == 1,

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      smoker |      6098    .2120367    .4087842          0          1
      smkban |      6098           1           0          1          1
         age |      6098    39.08101    11.84533         18         88
       black |      6098    .0783864    .2688006          0          1
      female |      6098    .6093801    .4879293          0          1
  highschool |      6098    .2974746    .4571846          0          1
     college |      6098    .5100033    .4999409          0          1

. reg smoker smkban

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) =   78.56
       Model |   14.313036     1   14.313036           Prob > F      =  0.0000
    Residual |  1821.59406  9998  .182195846           R-squared     =  0.0078
-------------+------------------------------           Adj R-squared =  0.0077
       Total |   1835.9071  9999  .183609071           Root MSE      =  .42684

------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smkban |  -.0775583          ?         ?     0.000     ?              ?
       _cons |   .2895951   .0068332    42.38   0.000     .2762006    .3029896
------------------------------------------------------------------------------

. reg smoker smkban age highschool college black female

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  6,  9993) =   38.25
       Model |  41.2188555     6  6.86980925           Prob > F      =  0.0000
    Residual |  1794.68824  9993  .179594541           R-squared     =  0.0225
-------------+------------------------------           Adj R-squared =  0.0219
       Total |   1835.9071  9999  .183609071           Root MSE      =  .42379
```

```
------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smkban |  -.0647838   .0087925    -7.37   0.000    -.0820189   -.0475487
         age |  -.0012684   .0003533    -3.59   0.000    -.0019609   -.0005759
  highschool |   .0883002   .0122338     7.22   0.000     .0643194     .112281
     college |  -.0229265    .011509    -1.99   0.046    -.0454865   -.0003666
       black |  -.0008369   .0159234    -0.05   0.958      -.03205    .0303761
      female |  -.0256619   .0086751    -2.96   0.003    -.0426669   -.0086569
       _cons |   .3275198    .018275    17.92   0.000     .2916972    .3633425
------------------------------------------------------------------------------

. probit smoker smkban age highschool college black female

Iteration 0:   log likelihood = -5537.1662
Iteration 1:   log likelihood = -5426.7393
Iteration 2:   log likelihood = -5426.6167
Iteration 3:   log likelihood = -5426.6167

Probit regression                                 Number of obs   =      10000
                                                  LR chi2(6)      =     221.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -5426.6167                       Pseudo R2       =     0.0200

------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smkban |  -.2076269   .0283519    -7.32   0.000    -.2631955   -.1520583
         age |  -.0040217   .0011595    -3.47   0.001    -.0062944   -.0017491
  highschool |    .272297   .0394637     6.90   0.000     .1949495    .3496445
     college |  -.0770965   .0382241    -2.02   0.044    -.1520143   -.0021786
       black |   .0025281   .0517443     0.05   0.961    -.0988889    .1039451
      female |  -.0813567   .0282668    -2.88   0.004    -.1367586   -.0259548
       _cons |  -.4371721    .059269    -7.38   0.000    -.5533371    -.321007
------------------------------------------------------------------------------
```