UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Exam: **ECON4160 - Econometrics - Modeling and systems estimation**

Date of exam: Monday, December 5, 2011 **Grades are given: December 22, 2011**

Time for exam: 2:30 p.m. – 5:30 p.m.

The problem set covers 7 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.

ECON 4160: ECONOMETRICS –
MODELLING AND SYSTEMS ESTIMATION

PROBLEM SET, EXAM AUTUMN 2011

PROBLEM 1 (weight: 30 %)

We have a set of three observable variables, (y, x, z), and are interested in a relationship between y and x, specified as

$$y = a + bx + u. \tag{1}$$

Assume that the error u is positively correlated with y and negatively correlated with x, because of the random disturbance in the underlying equation and the occurrence of a random measurement error in x. (Proof not required.) Therefore, z, which occurs as an exogenous variable in the model to which equation (1) belongs, is proposed as an instrument variable (IV). You are asked to give your advice about the estimation of b from the results below.

The correlation matrix of the three observable variables, obtained from a sample of 50 observations, is

	y	x	z
y	1.0000		
x	0.9971	1.0000	
z	-0.2538	-0.2385	1.0000

1A. Regressing y on x and regressing x on y by using OLS, give, respectively,

```
No. of obs.  =      50
R-squared    =  0.9941
Root MSE     =  1.0114
```

y	Coef.	Std. Err.	t	P>|t|
x	.8376545	.0093021	90.05	0.000
_cons	-4.342772	.8561157	-5.07	0.000

```
No. of obs.  =      50
R-squared    =  0.9941
Root MSE     =  1.2039
```

x	Coef.	Std. Err.	t	P>|t|
y	1.186785	.0131792	90.05	0.000
_cons	5.687901	.9597318	5.93	0.000

Derive the two OLS estimates of b, show that the former asymptotically underestimates b (plim $< b$) and that the latter asymptotically overestimates b (plim $> b$) under the assumptions above, and give a brief comment on the result.

1B. Using z as IV for x in equation (1), we get

```
Instrumental variables (2SLS) regression
No. of obs.  =      50
Root MSE     =  1.3132
-----------------------------------------------
     y |      Coef.   Std.Err.  Pseudo t value
-------+---------------------------------------
     x |   .8936922   .0506306    17.65
 _cons |  -9.427694   4.598034    -2.05
-----------------------------------------------
Instrumented:  x. Instruments:  z
```

Using z as IV for y in the inverse of equation (1), we get

```
Instrumental variables (2SLS) regression
No. of obs.  =      50
Root MSE     =  1.4694
----------------------------------------------
     x |      Coef.   Std.Err.  Pseudo t-value
-------+--------------------------------------
     y |   1.118954   .0633924    17.65
 _cons |   10.54915   4.547894     2.32
----------------------------------------------
Instrumented:  y. Instruments:   z
```

OLS regressions of y on z and of x on z give respectively,

```
No. of obs.  =      50
R-squared    =  0.0644
Root MSE     =  12.753
------------------------------------------------
     y |      Coef.   Std. Err.      t    P>|t|
-------+----------------------------------------
     z |  -.1150175   .0632814    -1.82   0.075
 _cons |   70.82474   1.862129    38.03   0.000
------------------------------------------------

No. of obs.  =      50
R-squared    =  0.0569
Root MSE     =   15.24
------------------------------------------------
     x |      Coef.   Std. Err.      t    P>|t|
-------+----------------------------------------
     z |  -.1286992   .0756241    -1.70   0.095
 _cons |   89.79874   2.225328    40.35   0.000
------------------------------------------------
```

Derive the two implied IV estimates of b, and comment on the result. What would you say about the quality of the IV z relative to equation (1)? If you were to choose the 'best' estimate of b among the four estimates obtained under **1A** and **1B**, which would you choose? Explain your choice.

PROBLEM 2 (weight: 30 %)

Consider an econometric two-equation model with equations of the form:

$$(1) \qquad y_i = \alpha + \beta x_i \qquad + u_i,$$
$$(2) \qquad x_i = \gamma + \delta y_i + \eta z_i + v_i,$$

where i ($i = 1, \dots, n$) indexes observation number, (y_i, x_i, z_i) are observable variables, $(\alpha, \beta, \gamma, \delta, \eta)$ are constants, (u_i, v_i) are disturbances with zero expectations, variances σ_{uu}, σ_{vv} and covariance σ_{uv}, and $\text{cov}(z_i, u_i) = \text{cov}(z_i, v_i) = 0$. We want to estimate β.

2A. Examine whether β can be estimated, and if so, explain how you would estimate it in the following cases:

Case 1: $(\alpha, \beta, \gamma, \delta, \eta)$ are unknown; $\sigma_{uv} \neq 0$ and unknown.
Case 2: $\eta = 0$, $(\alpha, \beta, \gamma, \delta)$ are unknown; $\sigma_{uv} \neq 0$ and unknown.
Case 3: $\delta = 0$, $(\alpha, \beta, \gamma, \eta)$ are unknown; $\sigma_{uv} = 0$.

2B. Assume that x_i is unobservable, (y_i, z_i) still observable; otherwise the situation is assumed to be as in Case 1. Could you then estimate β and if so, how? Give the reason for your answer.

2C. Assume that z_i can be split into $K\,(\geq 2)$ observable components, such that $z_i = z_{1i} + z_{2i} + \cdots + z_{Ki}$ where $\text{cov}(z_{ki}, u_i) = \text{cov}(z_{ki}, v_i) = 0$ $(k = 1, \dots, K)$. We reformulate the model as:

$$(3) \qquad y_i = \alpha + \beta x_i \qquad + u_i,$$
$$(4) \qquad x_i = \gamma + \delta y_i + \eta(z_{1i} + z_{2i} + \cdots + z_{Ki}) + v_i.$$

Explain how you would then estimate β.

2D. It has been suggested, instead of using (4), to assume that the K components of z_i have different effect on x_i, and to use the model

$$(5) \qquad y_i = \alpha + \beta x_i \qquad + u_i,$$
$$(6) \qquad x_i = \gamma + \delta y_i + \eta_1 z_{1i} + \eta_2 z_{2i} + \cdots + \eta_K z_{Ki} + v_i,$$

where $(\eta_1, \dots, \eta_K)$ are unknown coefficients. Would you recommend the same estimation procedure for β in equation (5) as you proposed for equation (3) in question **2C**, or would you use another one? Explain briefly. **Hint:** Consider the models' reduced forms.

PROBLEM 3 (weight: 40 %)

For this problem we have a data set of $n = 27326$ individual observations from a large health survey in Germany, in the years 1984–1994. We will use the data to examine factors believed to be related to peoples' health satisfaction, a qualitative variable represented in this data set by a binary variable. The variables we use are:

```
SATHIGH =  1 if the individual declares to be satisfied with own health, =  0 othervise.
AGE     =  Age in years.
COH     =  Birth year.
WORK    =  1 if employed, = 0 if not employed.
FEMALE  =  1 if female,   = 0 if male.
MARRIED =  1 if married,  = 0 if unmarried.
CHI     =  1 if there are children in the household, = 0 otherwise.
EDU     =  No. of years of education.
```

Some summary statistics are reported below:

Variable	Mean	Std. Dev.	Min	Max
SATHIGH	0.6095294	0.4878648	0	1
SAT	6.785662	2.293725	0	10
AGE	43.52569	11.33025	25	64
COH	1944.297	11.88667	1920	1969
WORK	0.6770475	0.4676133	0	1
FEMALE	0.4787748	0.4995584	0	1
MARRIED	0.7586182	0.4279291	0	1
CHI	0.40273	0.4904563	0	1
EDU	11.32063	2.324885	7	18

The vector $\boldsymbol{x}$=[AGE, COH, WORK, FEMALE, MARRIED, CHI, EDU] contains the variables to be treated as exogenous in the following. Five printouts from a discrete choice analysis are given at the end of the problem set.

3A. Estimation result from an OLS regression of SATHIGH on $\boldsymbol{x}$ is given in **Printout 1**. Explain what you conclude about the effects on the reported health status of (i) having one year higher age and (ii) being born one year later.

3B. Logit and Probit models are used more frequently than linear regression models in analyzing individuals' discrete choice. Logit and Probit estimation results for the binary health response are given in **Printout 2** and **Printout 3**, respectively. Explain briefly what you conclude from the Logit results about the effect on the health status of: (i) having a one year longer education period, (ii) of being a female compared with a male with the same characteristics, and (iii) of being employed rather than unemployed.

3C. The Logit estimates are substantially higher (in absolute value) than the corresponding Probit estimates, although the underlying problem is the same. Can you explain this?

3D. Marginal effects – i.e., first derivatives of the response probability with respect to the relevant explanatory variables at the sample mean – computed from the Logit estimates are given in **Printout 4**. Explain briefly why the order of magnitude of these effects differs systematically from the estimates in **Printout 2** and **Printout 3**, while they are similar in size to the corresponding estimates in **Printout 1**.

3E. Actually, the data set reports health satisfaction also in the following, more detailed way: The respondents have been asked to indicate the strength of their satisfaction, in the form of assigning an integer variable `SAT`, taking the 11 possible values 0 (=very low declared degree of health satisfaction), 1,2,...,9,10 (= very high declared degree of health satisfaction). The binary health indicator used in questions **3A** through **3D** is related to `SAT` in the following way:

```
SATHIGH = 0 if SAT = 0,1,2,3,4,5,6;    SATHIGH = 1 if SAT = 7,8,9,10.
```

Printout 5 reports the result of a linear regression similar to that in **Printout 1**, with `SAT` as the endogenous variable. Give your comments to the differences between these two sets of results.

Printout 1: Linear regression. Regressand = `SATHIGH`. No. of obs. = 27326

SATHIGH	Coef.	Std.Err.	t-value	P value
AGE	-0.0122657	0.0009582	-12.80	0.000
COH	-0.0041011	0.0009046	-4.53	0.000
WORK	0.0527522	0.006837	7.72	0.000
FEMALE	-0.0196338	0.0062449	-3.14	0.002
MARRIED	0.0122645	0.0073479	1.67	0.095
CHI	0.0266059	0.0067267	3.96	0.000
EDU	0.0208232	0.0012718	16.37	0.000
_cons	8.835158	1.79799	4.91	0.000

Printout 2: Logit regression. No. of obs. = 27326

SATHIGH	Coef.	Std.Err.	Pseudo t-value	P value
AGE	-0.0542254	0.0043363	-12.51	0.000
COH	-0.0186544	0.0040745	-4.58	0.000
WORK	0.2170284	0.030394	7.14	0.000
FEMALE	-0.0864393	0.0281429	-3.07	0.002
MARRIED	0.0515039	0.03307	1.56	0.119
CHI	0.1038393	0.0305083	3.40	0.001
EDU	0.0986473	0.0061489	16.04	0.000
_cons	37.80347	8.098391	4.67	0.000

Printout 3: Probit regression. No. of obs. = 27326

SATHIGH	Coef.	Std.Err.	Pseudo t-value	P value
AGE	-0.0336117	0.00265	-12.68	0.000
COH	-0.0116667	0.0024939	-4.68	0.000
WORK	0.1336249	0.0186961	7.15	0.000
FEMALE	-0.0551729	0.0172021	-3.21	0.001
MARRIED	0.0319183	0.0202209	1.58	0.114
CHI	0.0652121	0.0186252	3.50	0.000
EDU	0.0592194	0.0036404	16.27	0.000
_cons	23.65542	4.956662	4.77	0.000

Printout 4: Marginal effects obtained from the Logit estimates**

Variable	Est.of dP/dx	Std.Err.
AGE	-0.0128145	0.00102
COH	-0.0044084	0.00096
WORK*	0.0516893	0.00729
FEMALE*	-0.0204334	0.00665
MARRIED*	0.0122083	0.00786
CHI*	0.0244762	0.00717
EDU	0.0233123	0.00145

(*) dP/dx for a dummy variable refers to a change from 0 to 1.
(**) Estimated P(SATHIGH) at sample mean = 0.61696426.

Printout 5: Linear regression. Regressand = SAT. No. of obs. = 27326

SAT	Coef.	Std.Err.	t-value	P value
AGE	-0.0769768	0.0044847	-17.16	0.000
COH	-0.0357537	0.0042338	-8.44	0.000
WORK	0.3737091	0.0319988	11.68	0.000
FEMALE	-0.0013297	0.0292277	-0.05	0.964
MARRIED	0.1023993	0.0343902	2.98	0.003
CHI	0.1205228	0.0314826	3.83	0.000
EDU	0.0891128	0.0059523	14.97	0.000
_cons	78.26454	8.415081	9.30	0.000