Exercise sheet 6

Models with endogenous explanatory variables

Note: Some of the exercises include estimations and references to the data files. Use these to compare them to the results you obtained with Gretl.

1. Recall that in the simple lineal model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

when X is endogenous but we have an instrument Z that can be correlated with ε, we have that

$$p\,\text{lím}\,\widetilde{\beta}_1 = \beta_1 + \frac{C(Z,\varepsilon)}{C(Z,X)} \quad \text{(estimator simple VI)}$$

$$p\,\text{lím}\,\widehat{\beta}_1 = \beta_1 + \frac{C(X,\varepsilon)}{V(X)} \quad \text{(estimator OLS)}$$

Assume that $\sigma_X = \sigma_\varepsilon$, so that the population variation in the error term is the same as that of X. Suppose that the instrumental variable Z, is slightly correlated with ε, $\rho(Z,\varepsilon) = 0{,}1$. Suppose also that Z and X have a somewhat stronger correlation, $\rho(X,Z) = 0{,}2$.

a) What is the asymptotic bias in the IV estimator that uses Z as an instrument?

b) How much correlation would have to exist between X and ε so that the OLS estimator is more asymptotically biased than the previous IV estimator?

2. [Problem 16.10 in Wooldridge Textbook] Use the data in `MROZ.gdt` and consider the example of women labor supply.
Let *hours* be the total annual hours worked by a woman, *wage* is the wage rate per hour, *educ* years of education, *age* is age in years, *kidslt6* is the number of kids a woman has under the age of 6, $nwifeinc$, is the household income (excluding womens' wage income), *exper* is years of work experience, *motheduc* and $fatheduc$ the years of education of her mother and father respectively.

a) Estimate the labor supply equation

$$\begin{aligned}\ln(hours) = \alpha_1 \ln(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age \\ + \beta_{13} kidslt6 + \beta_{14} nwifeinc + u_1\end{aligned}$$

by 2SLS using *exper* and $exper^2$ as instruments for $\log(wage)$, and compare the result with that obtained with *hours* as the dependant variable.

b) In the labor supply equation in part *(a)*, it is possible that *educ* is endogenous, given that ability is omitted. Use *motheduc* and $fatheduc$ as instruments for *educ*. Now there are 2 endogenous explanatory variables in this equation.

c) Test the overidentifying restrictions in the 2SLS estimation from part *(b)*. What can you conclude about the validity of the instruments?

3. We would like to study the returns of education (ED) to wages (W). We are interested in knowing on the effect of years of education on wages. We have a sample of 3010 US young men in 1976 from the *National Longitudinal Survey of Young Men* (NLSYM) of the National Longitudinal Surveys in the USA for the year 1976. For each individual, we observe ED (Years of education), EX (Experience, in years), EX^2 (Experience squared), $WHITE$ (Binary variable that takes the value of 1 if the young men is white and 0 otherwise).
The following model is considered to analyze the return to education:
$$\ln W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 EX^2 + \varepsilon_1 \quad (1)$$
Notice that the error term ε_1 may include unobserved factors not included in the model that can affect wages. In particular, $ABIL$ (ability), which is unobserved.
We also an additional variable: $NEAR$ is a dummy variable that takes value 1 if the individual lived close to a university, and zero otherwise, for which we know that $C\,(NEAR, \varepsilon_1)$.

a) Estimate model (S1) by OLS, using the subsample of white young men (i.e., restricting to $WHITE = 1$).

b) Suppose that $ABIL$ is actually a relevant variable, and $C\,(ABIL, ED) \neq 0$, while it is uncorrelated with the remaining explanatory variables in (E1). Will the OLS estimate of β_1 in (E1) consistently estimate the causal effect of education on wages? Justify.

c) Consider the use of $NEAR$ as an instrument. Estimate the auxiliary equation
$$ED = \pi_0 + \pi_1 EX + \pi_2 EX^2 + \pi_3 NEAR + v$$
Given the assunptions and the estimations above, can we assert that $NEAR$ is a valid instrument for ED? Justify.

d) Estimate model (S1) by 2SLS, using the subsample of white young men, and $NEAR$ as instrument for ED.

e) Is ED is exogenous? Justify. Given the results, choose the appropriate estimate of the causal effect of education on wages and interpret it.

f) Consider now the full sample of both white and non-white young men. To account for ethnical differences, we consider the equation
$$\ln W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 EX^2 + \beta_4 WHITE + \beta_5\,(ED \times WHITE) + \varepsilon_2 \quad (2)$$
Propose instruments for the two potentially endogenous variables, ED and $(ED \times WHITE)$. Estimate their corresponding first-stage equations and check the validity of the instruments.

g) Estimate (E2) by OLS and 2SLS and compare the results. Are ED and $(ED \times WHITE)$ exogenous? What estimates would you choose? Justify.

4. In accordance with Friedman's Permanent Income Theory,
$$Y_i^* = \alpha + \beta X_i^* \quad (3)$$

where Y_i^* is 'permanent'consumption and X_i^* is 'permanentíncome. Instead of observing the 'permanentes'variables, we observe

$$Y_i = Y_i^* + u_i$$
$$X_i = X_i^* + v_i$$

where Y_i, X_i measure with error (u_i v_i) the corresponding variables of interest Y_i^*, X_i^*. Using the observed variables, we can write the consumption function as

$$\begin{aligned} Y_i &= \alpha + \beta\left(X_i - v_i\right) + u_i \\ &= \alpha + \beta X_i + \left(u_i - \beta v_i\right) \end{aligned}$$

a) Suppose that $E(u_i) = E(v_i) = 0$, $Var(u_i) = \sigma_u^2$, $Var(v_i) = \sigma_v^2$, $Cov(Y_i^* u_i) = 0$, $Cov(X_i^* v_i) = 0$, $Cov(u_i X_i^*) = Cov(v_i Y_i^*) = Cov(v_i u_i) = 0$. Show that the OLS estimator of β, $\hat{\beta}$, converges in probability to

$$\frac{\beta}{1 + (\sigma_v^2/\sigma_{X^*}^2)}$$

b) Comment the sign of the asymptotic bias of $\hat{\beta}$.

5. The argument that inflation boosts economic growth has been neglected by empirical cross-section studies. Such studies regress the real income (GDP) growth on the inflation rate. Nevertheless, the variables measuring inflation and real income, X_i and Y_i, are subject to measurement error. Assume that, actually, there exists a exact relationship between real income growth, Y_i^*, and the true inflation rate, X_i^*. Also, it is assumed that the nominal income growth, $W_i^* = X_i^* + Y_i^*$ is correctly measured, so it can be used to measure real income growth. Then

$$\begin{aligned} Y_i &= W_i^* - X_i \\ X_i &= X_i^* + \varepsilon_i \qquad \varepsilon_i \sim iid(0, \sigma_\varepsilon^2) \end{aligned}$$

a) Derive the probabilistic limit of the OLS estimator of Y on X.

b) Given the result above, what can be said about the future empirical results neglecting teha tinflation boosts growth?

6. Consider the following specification for demand and supply of wine in a country

$$\begin{aligned} q_i^D &= \alpha_1 p_i + \alpha_2 y_i + u_{i1} \\ q_i^S &= \beta p_i + u_{i2} \\ q_i^D &= q_i^S = q_i \end{aligned}$$

where, for municipality i, q_i is the wine consumption per household, p_i is the relative price of wine and y_i is the income per household. In the system above, price and demanded quantity are simultaneously determined, while the variable y_i is exogenous. **All the variables are in logs**. For a sample of size 1000, we have obtained the following statistics:

$$\begin{array}{lll} \sum_i p_i^2 = 42 & \sum_i p_i q_i = 5 & \sum_i p_i y_i = 12 \\ \sum_i y_i^2 = 10 & \sum_i y_i q_i = 3 & \\ \sum_i^2 q_i^2 = 11 & & \end{array}$$

Suppose that we are concerned with estimating a supply equation.

a) Compute the OLS estimator of β e and interpret the coefficient. Is the estimator consistent? Justify.

b) Is y_i a valid instrument for p_i?. Justify. Compute the instrumental-variables estimator of β using such instrument.

7. A company selling sport goods is interested on evaluating the impact of clientsíncomes on its sales. For such purpose, it undertakes a survey for a sample of individuals who buy sport goods and considers the following specification:

$$\begin{aligned} expend = {} & \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + \beta_4 gender + \beta_5 bach + \beta_6 south \\ & + \beta_7 weight + \beta_8 weight \times gender + \beta_9 south \times gender + \varepsilon \end{aligned} \tag{C}$$

where for each individual,

expend is his/her annual expenditure in sport goods (in thousand euros),
inc is his/her annual income (in thousand euros),
age is his/her age in years,
gender is a binary variable which takes value 1 if the individual is a woman and 0 otherwise,
bach denotes the individual's marital status, taking value 1 if bachelor and 0 otherwise,
south is a binary variable which takes value 1 if the individual lives in the south and 0 otherwise,
weight is his/her weight in kilograms.

Furthermore, income can be correlated with unobserved characteristics which, in turn, might affect expenditure in sport goods. If that is the case, then $C(inc, \varepsilon) \neq 0$. The remaining covariates are not expected to be correlated with the error. In addition to the aforementioned covariates, there is information about the years of education of the individual (*educ*) and the years of education of his/her father (*educp*), which are also assumed to be uncorrelated with any unobserved characteristic affecting the expenditure in sport goods (ε). It is expected that individuals who are more educated or with more educater parents will have higher income. The following estimates have been obtained with Gretl (notice that some results may have been omitted):

OUTPUT 1: OLS estimates using the 935 observations 1–935
Dependent variable: *expend*

	Coefficient	Std. Error	t-ratio	p-value
const	37,7486	31.5318	1,1972	0.2316
inc	0,6671	0.6138	1,0869	0.2774
age	0,3401	1.9090	0,1782	0.8586
*age*2	−0,0036	0.0014		
gender	−2,2428	0.6031	−3,7190	0.001
bach	0,7775			0.002
south	−0,1803	0.5536	−0,3257	0.7447
weight	−0,1025	0.0512	-2.0023	0.0455
weight×*gender*	−0,0323	0.1506	−0,2144	0.8303
south×*gender*	−0,0692	1.4793	−0,0468	0.9627

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.047795	S.E. of regression	0.007188
R^2	0.019495	Adjusted R^2	0.009955
$F(9, 925)$	2.043539	P-value(F)	0.032059
Log-likelihood	3292.838	Akaike criterion	−6565.675
Schwarz criterion	−6517.270	Hannan–Quinn	−6547.218

NOTA: El R^2 de una estimación similar a la anterior omitiendo *gender*, *weight*×*gender* y *south*×*gender* es 0,009495

OUTPUT 2: OLS estimates using the 935 observations 1–935
Dependent variable: *inc*

	Coefficient	Std. Error	t-ratio	p-value
const	1,2033	1.8148	0,6631	0.5075
age	−0,0577	0.1103	−0,5233	0.6010
*age*2	0,0006	0.0017	0,3597	0.7192
gender	0,1523	0.0995	1,5305	0.1263
bach	−0,1739	0.0436	−3,9930	0.0001
south	0,0589	0.0315	1,8688	0.0620
weight	−0,0035	0.0029	−1,2100	0.2267
weight×*gender*	−0,0165	0.0102	−1,6209	0.1055
south×*gender*	0,0977	0.0966	1,0111	0.3123
educp	0,0138	0.0047	2,9464	0.0033
educ	0,0470	0.0068	6,9315	0.0000

Mean dependent var	0.975920	S.D. dependent var	0.405896
Sum squared resid	99.17078	S.E. of regression	0.368579
R^2	0.186567	Adjusted R^2	0.175424
$F(10, 730)$	16.74310	P-value(F)	1.79e–27
Log-likelihood	−306.2997	Akaike criterion	634.5994
Schwarz criterion	685.2874	Hannan–Quinn	654.1416

NOTA: El R^2 de una estimación similar a la anterior omitiendo *educp* y *educ* es 0,1717

OUTPUT 3: OLS estimates using the 935 observations 1–935
Dependent variable: *expend*

	Coefficient	Std. Error	t-ratio	p-value
const	48,0705	35.4399	1,3564	0.1754
inc	−0,4594	0.2069	−2,2199	0.0267
age	−0,4097	2.1563	−0,1900	0.8494
*age*2	0,055	0.0324	1,6959	0.0899
gender	−0,5713	1.9915	−0,2869	0.7743
bach	−0,2548	0.9057	−0,2813	0.7786
south	0,1037	0.6348	0,1634	0.8703
weight	−0,1139	0.0568	−2,0047	0.0454
weight×*gender*	−0,2221	0.2004	−1,1082	0.2681
south×*gender*	1,6185	1.8847	0,8587	0.3908
resid×*inc*	6,3914	2.1915	2,9164	0.0036

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.037615	S.E. of regression	0.007178
R^2	0.028753	Adjusted R^2	0.015448
$F(10, 730)$	2.161070	P-value(F)	0.018389
Log-likelihood	−2506.447	Akaike criterion	5034.894
Schwarz criterion	5085.582	Hannan–Quinn	5054.436

OUTPUT 4: TSLS,using the 935 observations 1–935
Dependent variable: *expend*
Instrumented: *inc*
Instruments: const *age age*2 *gender bach south weight weight*×*gender south*×*gender feduc educ*

	Coefficient	Std. Error	z	p-value
const	48,0705	37.2741	1,2896	0.1972
inc	4,5944	2.1767	2,1107	0.0348
age	−0,4097	2.2679	−0,1806	0.8566
*age*2	0,0055	0.0341	0,1620	0.8713
gender	−0,5713	2.0946	−0,2727	0.7850
bach	−0,2548	0.9526	−0,2675	0.7891
south	0,1037	0.6676	0,1553	0.8766
weight	−0,1139	0.0598	−1,9060	0.0566
weight×*gender*	−0,2221	0.2107	−1,0537	0.2920
south×*gender*	1,6185	1.9822	0,8165	0.4142

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.041666	S.E. of regression	0.007549
R^2	0.000006	Adjusted R^2	-0.012306
$F(9, 731)$	1.546437	P-value(F)	0.127654

a) In the baseline model, ¿can we assert that

$$E\left(expend|inc, age, gender, bach, south, weight\right) = L\left(expend|inc, age, gender, bach, south, weight\right)?$$

Justify.

b) Given the available evidence and the assumptions above, can be establish that *educp* and *educ* are valid instruments for *inc*? Justify your answer, indicating the evidence in which you base it.

c) Can we assert that *inc* is endogenous? Justify.

d) We want to test whether the expenditure in sport goods is independent of gender. Write the null and the alternative hypotheses in terms of the model parameters. Explain how you would build the test statistic and, if there is the information required, implement the test. Otherwise, explain what information is missing.