Universidad Carlos III de Madrid
Econometrics I
Final Exam, Convocatoria Extraordinaria, Course 2009-2010.

September 6, 2010

Total Time: 2 Hours.

Important Remark: Note that some information contained in the outputs below is redundant.

Question 1 One of the most widely used production function is CES (Constant Elasticity of Substitution) function, because it nests as special cases other production functions commonly used in the empirical literature such as the Cobb-Douglas or Leontieff. The expression of the CES function is:

$$Y = \gamma \left(\delta K^{\rho} + (1-\delta) L^{\rho}\right)^{v/\rho},$$

where Y, K and L denote output, capital and labor, γ is the efficiency parameter, δ is the rate at which the two factors enter the production function, ρ is the parameter that defines the elasticity of substitution, and v is the parameter measuring the returns to scale, so that $v = 1$, $v > 1$, $v < 1$ indicate, respectively, constant returns, increasing returns and decreasing returns to scale.

One researcher has specified an econometric model to estimate the above technology using data on 25 manufacturing companies based on first-order linear approximation of the CES function (expressed in logarithms):

$$\begin{aligned} \log Y &= \log\gamma + \delta v \log K + (1-\delta) v \log L - 1/2\rho \log(1-\delta) v [\log(K/L)]^2 + \epsilon \\ &= \beta_0 + \beta_1 \log K + \beta_2 \log L + \beta_3 [\log(K/L)]^2 + \epsilon. \end{aligned}$$

(a) Test the hypothesis of constant returns to scale in the CES technology. You should clearly establish the null hypothesis and the method of the test.

(b) Investigate the significance of the parameter β_3 in the OLS estimation.

(c) Consider now that the true model to represent technology companies in a given sector is

$$\log(Y/L) = \beta_0 + \beta_1 \log(K/L) + \beta_2 [\log(K/L)]^2 + \epsilon,$$

where the error term satisfies the assumptions of linear regression model: $E(\epsilon|K,L) = 0, Var(\epsilon|K,L) = \sigma^2$, and where $\beta_2 > 0$ and $Cov(\log(K/L), [\log(K/L)]^2) > 0$.

If you omit $[\log(K/L)]^2$, get the sign and magnitude of the bias of the inconsistency (or asymptotic bias) of the estimator of β_1 in the simple regression of $\log(Y/L)$ on $\log(K/L)$. What would be the bias in the case of conditional heteroskedasticity?

Output 1. Dependent variable: $\log Y$
Method: Least Squares
Sample: 1 25
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.9602	2.2051	-0.89	0.384
$\log K$	0.6501	0.0303	21.47	0.000
$\log L$	0.5592	0.2075	2.69	0.013
$[\log(K/L)]^2$	0.0879	-	-	-

R-squared 0.9912
Adjusted R-squared 0.9900
S.E. of regression 0.0266
Sum squared resid 0.0148

Output 2. Dependent Variable: $\log(Y/L)$
Method: Least Squares
Sample: 1 25
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.0155	0.0084	1.84	0.079
$\log(K/L)$	0.6262	0.0144	43.51	0.000
$[\log(K/L)]^2$	0.0379	0.0323	1.17	0.253

R-squared 0.9921
Adjusted R-squared 0.9914
S.E. of regression 0.0264
Sum squared resid 0.0154

Output 3. Dependent Variable: $\log Y$
Method: Least Squares
Sample: 1 25
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.6394	1.1255	0.56	0.576
$\log K$	0.6130	0.0135	45.44	0.000
$\log L$	0.3214	0.1142	2.81	0.010

R-squared 0.9905
Adjusted R-squared 0.9896
S.E. of regression 0.0271
Sum squared resid 0.0161

Question 2 We want to understand what are the determinants of participation in physical activity. We have information for adults with ages ranging between 25 and 55 years. The dummy variable *sport* takes the value one if the individual participated in any physical activity during the previous week and zero otherwise. The explanatory variables that we have in our model are: dummy variable $female$ which takes the value one if the individual is female, and the continuous variables age (age), age squared ($age2$), and years of education ($yedu$).

(a) Interpret the coefficient of the variable $female$. Propose a model where the effect of education and age on the decision to participate in a physical activity depends on gender. Discuss whether this model is more general or not than estimating two different linear models for male and female.

(b) Obtain an expression for the conditional variance of *sport* in terms of the variables $female$, age and $yedu$. Test if there is conditional heteroskedasticity in the linear model using the information provided in Output 1.

(c) Using the estimated linear probability model, indicate whether the probability of participating in physical activity always decreases with age. Explain in detail how one can test that for individuals with 20 years old this probability decreases with age.

Output 1: OLS estimates using 4986 observations from 1-4993
Missing and incomplete observations that have been removed: 7
Dependent variable: *sport*
Standard Deviations robust to heteroscedasticity

Variable	Coefficient	Std. Dev	t-Statistic	Prob.
const	0.793	0.141	5.624	0.000
female	-0.276	0.0127	-21.725	0.000
age	-0.020	0.0074	-2.701	0.007
age2	0.0002	1.086e-04	1.841	0.066
yedu	0.0148	0.0017	8.768	0.000

Mean var. dependent	0,359807
Std. Dev. of var. dependent	0,479992
Sum squared resid	1021,94
Std. Dev of resid ($\hat{\sigma}$)	0,452955
$R-squared$	0,110198
$AdjustedR-squared$	0,109483
$F(4,4981)$	173,525
log-likelihood	-3123,6

Question 3 The following is a simultaneous equation model that we consider to examine whether the openness of an economy (*open*) leads to lower rates of inflation (inf),

$$\begin{aligned} inf &= \delta_{10} + \gamma_{12} open + \delta_{11} \log(pcinc) + u_1 \\ open &= \delta_{20} + \gamma_{21} inf + \delta_{21} \log(pcinc) + \delta_{22} \log(land) + u_2. \end{aligned}$$

It is assumed that (the logarithms of) *pcinc* (income per capita) and *land* (agricultural land) are exogenous throughout the year. Various estimates have been obtained using OLS and 2SLS and are provided below.

(a) Get the reduced form of the system. How can its parameters be estimated consistently?

(b) Study the identification of simultaneous equations using the provided estimation results. Are there any potentially overidentified equation?

(c) Test whether the variable *open* is an endogenous regressor in the first equation. Depending on the result, discuss which parameter estimates of γ_{12} and δ_{11} would be preferable.

(d) If you know that $\delta_{11} = 0$, how would it change your answer to question (b)? What if alternately one knows that $\delta_{21} = 0$ (but it ignores the value of δ_{11} or of any other parameter in the system)?

Output 1: OLS estimates using 114 observations 1-114
Dependent Variable: *inf*

Variable	Coefficient	Std. Dev.	t-Statistic	Prob.
const	25,1040	15,2052	1,6510	0,1016
open	-0,215070	0,0946289	-2,2728	0,0250
lpcinc	0,0175673	1,97527	0,0089	0,9929

Mean var. dependent	17,2640
Std. Dev. of var. dependent	23,9973
Sum squared resid	62127,5
Std. Dev of resid ($\hat{\sigma}$)	23,6581
$R-squared$	0,0452708
$AdjustedR-squared$	0,0280685
$F(2,111)$	2,63167
Prob. for $F()$	0,0764453

Output 2: OLS estimates using 114 observations 1-114
Dependent Variable: *open*

Variable	Coefficient	Std. Dev.	t-Statistic	Prob.
const	116,226	15,8808	7,3187	0,0000
inf	-0,0680353	0,0715556	-0,9508	0,3438
lpcinc	0,559501	1,49395	0,3745	0,7087
lland	-7,3933	0,834814	-8,8563	0,0000

Mean var. dependent	37,0789
Std. Dev. of var. dependent	23,7535
Sum squared resid	34865,3
Std. Dev of resid ($\hat{\sigma}$)	17,8033
$R - squared$	0,453162
$Adjusted R - squared$	0,438249
$F(3, 110)$	30,3855
Prob. for $F()$	$<$ 0,00001

Output 3: OLS estimates using 114 observations 1-114
Dependent Variable: *inf*

Variable	Coefficient	Std. Dev.	t-Statistic	Prob.
const	-12,615	21,0313	-0,5998	0,5498
lpcinc	0,191394	1,98158	0,0966	0,9232
lland	2,55380	1,08049	2,3635	0,0198

Mean var. dependent	17,2640
Std. Dev. of var. dependent	23,9973
Sum squared resid	61903,2
Std. Dev of resid ($\hat{\sigma}$)	23,6154
$R - squared$	0,0487174
$F(2, 111)$	2,84229
Prob. for $F()$	0,0625432

Output 4: OLS estimates using 114 observations 1-114
Dependent Variable: *open*

Variable	Coefficient	Std. Dev.	t-Statistic	Prob.
const	117,085	15,8483	7,3878	0,0000
lpcinc	0,546479	1,49324	0,3660	0,7151
lland	-7,5671	0,814216	-9,2937	0,0000

Mean var. dependent	37,0789
Std. Dev. of var. dependent	23,7535
Sum squared resid	35151,8
Std. Dev of resid ($\hat{\sigma}$)	17,7956
$R - squared$	0,448668
$F(2, 111)$	45,1654
Prob. for $F()$	$<$ 0,00001

Output 5: 2SLS estimates using 114 observations 1-114
Dependent Variable: *inf*
Instruments: *lland*

Variable	Coefficient	Std. Dev.	t-Statistic	Prob.
const	26,8993	15,4012	1,7466	0,0807
open	-0,337487	0,144121	-2,3417	0,0192
lpcinc	0,375823	2,01508	0,1865	0,8520

Mean var. dependent	17,2640
Std. Dev. of var. dependent	23,9973
Sum squared resid	63064,2
Std. Dev of resid ($\hat{\sigma}$)	23,8358
$F(2, 111)$	2,62498
Prob. for $F()$	0,0769352

Hausman Test –
Null Hypothesis: The OLS estimates are consistent
Asymptotic Test Statistic: χ_1^2 = 1,35333 with p-value = 0,244697

First-stage $F(1, 111)$ = 86,3734

CRITICAL VALUES: Z is normal with zero mean and variance one and χ_q^2 is a chi-square with q degrees of freedom; $\Pr(Z > Z_\alpha) = \alpha$; $\Pr\left(\chi_q^2 > \chi_{q;\alpha}^2\right) = \alpha$. Note that the distribution F can be approximated by the χ^2. That is, $\chi_q^2 \sim q \cdot F_{q,n}$ for n large; $\Pr\left(\chi_q^2 > \chi_{q;\alpha}^2\right) \simeq \Pr\left(q \cdot F_{q,n} > \chi_{q;\alpha}^2\right)$.

$Z_{0,025} = 1,96$	$Z_{0,05} = 1,645$	$Z_{0,01} = 2,326$	$Z_{0,005} = 2,576$	
$Z_{0,1} = 1,282$	$\chi^2_{3;0,01} = 11,34$	$\chi^2_{3;0,05} = 7,82$	$\chi^2_{5;0,05} = 11,07$	
$\chi^2_{2;0,05} = 5,99$	$\chi^2_{2;0,01} = 9,21$	$\chi^2_{6;0,05} = 12,59$	$\chi^2_{2;0,1} = 4,61$	$\chi^2_{1;0,05} = 3,84$
$\chi^2_{6;0,01} = 16,81$	$\chi^2_{4;0,05} = 9,49$	$\chi^2_{3;0,1} = 6,25$	$\chi^2_{4;0,01} = 13,28$	$\chi^2_{1;0,01} = 6,64$