Econometria II
Part 2: Col·linealitat

Riste Gjorgjiev
pareto.uab.es/∼rgjorgjiev/econ_cat

Lorenzo Burlon
idea.uab.es/lburlon/teaching_econometria2.html

Universitat Autònoma de Barcelona

Col·linealitat

L'objectiu d'aquest tema és poder contestar les següents preguntes

- Què és la Col·linealitat
- Quins són els seus efectes al MLG
- Com la podem detectar
- Què podem fer per evitar els efectes negatius

Mirem les següents sèries de dades *(corr.xls)*

X_2	X_3	X_3^*	X_4
10	50	52	7
15	75	75	4
18	90	97	20
24	120	129	19

Què podem dir sobre la correlació entre elles.

Coeficientes de correlacin, usando las observaciones 1 - 4

x2	x3	x3_	x4	
1.0000	1.0000	0.9973	0.7550	x2
	1.0000	0.9973	0.7550	x3
		1.0000	0.7978	x3_
			1.0000	x4

Com el podem calcular?

$$r_{ij} = \frac{cov(X_i, X_j)}{\sigma_X \sigma_Y}$$

Un Exemple

Utilitzant les dades *Mortalitat* dels EEUU sobre la taxa de mort per malaties del cor **(chd)**, el consum de tabac **(cig)**, begudes alcoloques **(spirits)**, cervesa **(beer)** i vi **(wine)**, en el format següent

- chd - tasa de mort per població de 100,000 habitants
- cig - lliures de consum de tabac per capita (339 cigarettes per libra de tabac)
- spirits - consum de begudes alcoloques per capita, (en galons taxats)
- beer - consum de cervesa per capita, (en galons taxats)
- wine - consum de vi per capita, (en galons taxats)

s'han obtingut els següents resultats

$$\widehat{\text{chd}} = \underset{(58.939)}{334.914} + \underset{(5.1560)}{5.41216}\,\text{cig} + \underset{(7.3730)}{36.8783}\,\text{spirits} - \underset{(1.2513)}{5.10365}\,\text{beer} + \underset{(12.735)}{13.9764}\,\text{wine} \tag{1}$$

$$\widehat{\text{chd}} = \underset{(56.624)}{353.581} + \underset{(4.7523)}{3.17560}\,\text{cig} + \underset{(7.2750)}{38.3481}\,\text{spirits} - \underset{(1.0102)}{4.28816}\,\text{beer} \tag{2}$$

$$\widehat{\text{chd}} = \underset{(67.210)}{243.310} + \underset{(6.1508)}{10.7535}\,\text{cig} + \underset{(8.0359)}{22.8012}\,\text{spirits} - \underset{(12.638)}{16.8689}\,\text{wine} \tag{3}$$

$$\widehat{\text{chd}} = \underset{(49.119)}{181.219} + \underset{(4.4371)}{16.5146}\,\text{cig} + \underset{(6.2079)}{15.8672}\,\text{spirits} \tag{4}$$

Un Exemple

veiem que

- el signe dels coeficients canvia segons el model
- la magnitud dels paràmetres varia massa
- les estimacions dels paràmetres són molt sensibles als models diferents

Aquests punts són uns efectes de la presencia de Col·linealitat

Col·linealitat, definició

Definició

L'existencia d'una relació lineal entre les variables explicatives (regressors) es diu col·linealitat

$X_1, X_2, \ldots, X_k$ - regressors d'un model. Diem que el model mostra presència de col·linealitat perfecte, si hi existeixen* $\lambda_1, \ldots, \lambda_k$ t.q.

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0$$

o presència de col·linealitat imperfecte, si a més existeix ν t.q.

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k + \nu = 0$$

La Col·linealitat Perfecte

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0$$

- L'efecte a la regressió
  - D'on pot aparèixer
    - errors en la construcció de la matriu dels regressors
    - especificació dels models amb variables fictícies (exemple)
  - Què passa amb la matriu $X'X$
- Què podem fer si el model és $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, on $X_{2i} = \alpha_1 + \alpha_2 X_{3i}$

Exemple

Fem un model per determinar l'efecte de les següents variables

- X_i - qualitat del pis
- variables fictícies per la localització del pis
  Barcelona (B), Girona (G), Lleida (L) i Tarragona (T)

al preu de lloger d'un pis Y_i

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 B + \beta_4 G + \beta_5 L + \beta_6 T + u_i$$

Quin és el problema en aquesta especificació?

La matriu $X'X$

En el cas k=3, tenim l'equació

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i,$$

on $X_{3i} = \lambda X_{2i}$ $(-\lambda X_{2i} + X_{3i} = 0)$ - Col·linealitat Perfecte
El valor estimat del paràmetre β_2, $\hat{\beta}_2$ és

$$\hat{\beta}_2 = \frac{(\sum Y_i X_{2i})(\sum X_{3i}^2) - (\sum Y_i X_{3i})(\sum X_{2i} X_{3i})}{(\sum X_{2i}^2)(\sum X_{3i}^2) - (\sum X_{2i} X_{3i})^2}$$

$\Rightarrow \hat{\beta}_2$ no és determinat

Què passa amb $\hat{\beta}$ en el cas quan k té un valor arbitrari?

Estimació amb Col·linealitat Perfecte

Considerem el mateix exemple

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i,$$

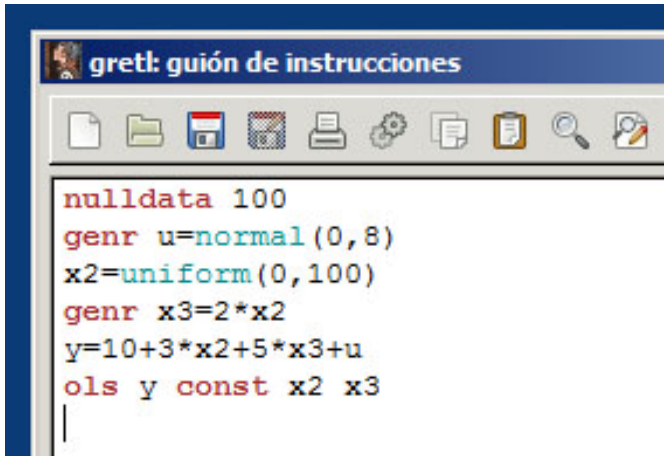
on $X_{3i} = \lambda X_{2i}$

Hi ha alguna cosa què podem fer? - Reescriure l'equació inicial

- un exemple amb Gretl

Generem dades amb col·linealitat perfecte i fem un MQO al model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \text{ on } X_{3i} = 2X_{2i}$$



Tenim els següents resultats

$$\widehat{\mathrm{y}} = \underset{(1.3975)}{8.80043} + \underset{(0.026752)}{13.0508}\ \mathrm{x2}$$

$$T = 100 \quad R^2 = 0.9996 \quad F(1, 98) = 2.3799\mathrm{e}{+005} \quad \hat{\sigma} = 7.6214$$

La Col·linealitat Imperfecte

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k + \nu = 0$$

- D'on pot aparèixer
  - moltes variables econòmiques tenen l'intenció de moure's al mateix temps
  - restriccions fisiques a la població utilitzada (despeses en llum $\rightarrow$ ingrés i grandària de la casa)
  - models sobre determinats

X_2	X_3	X_3^*
10	50	52
15	75	75
18	90	97
24	120	129

- Correlació entre les variables?
- estimació sota la col·linealitat imperfecte (k=3)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \text{ on } X_{3i} = \lambda X_{2i} + \nu_i$$

Els coeficients estimats són

$$\hat{\beta}_2 = \frac{(\sum Y_i X_{2i})(\sum X_{3i}^2) - (\sum Y_i X_{3i})(\sum X_{2i} X_{3i})}{(\sum X_{2i}^2)(\sum X_{3i}^2) - (\sum X_{2i} X_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum Y_i X_{3i})(\sum X_{2i}^2) - (\sum Y_i X_{2i})(\sum X_{2i} X_{3i})}{(\sum X_{2i}^2)(\sum X_{3i}^2) - (\sum X_{2i} X_{3i})^2}$$

Però què passa amb les variàncies:

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_{2i}^2 (1 - r_{23}^2)} \qquad var(\hat{\beta}_3) = \frac{\sigma^2}{\sum X_{3i}^2 (1 - r_{23}^2)}$$

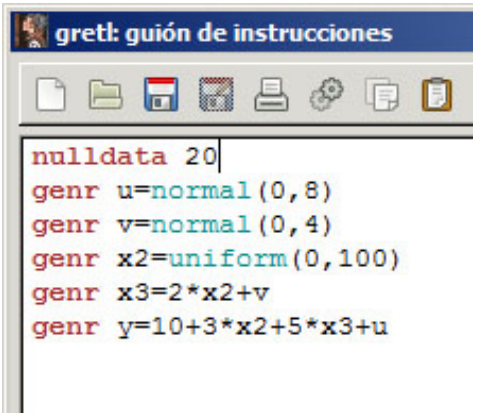
Si les variables X_2 i X_3, r_{23}^2 són més correlacionades, el seu coeficient de correlació està més a prop de 1. Implicacions?

Fem un exemple per demostrar les conseqüències de la col·linealitat.
Tenim el model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \text{ però}$$

$$X_{3i} = \alpha X_{2i} + v_i$$

Generem les dades amb gretl



Els resultats de l'estimació són

$$\underset{}{\widehat{y}} = \underset{(2.7815)}{15.6710} + \underset{(0.88412)}{3.01685}\,x2 + \underset{(0.44187)}{4.96338}\,x3$$

$$T = 20 \quad \bar{R}^2 = 0.9997 \quad F(2,17) = 29972. \quad \hat{\sigma} = 7.7634$$

(5)

Què passa si augmentem la mostra?

$$\underset{}{\widehat{y}} = \underset{(1.5862)}{10.8934} + \underset{(0.46297)}{2.74858}\,x2 + \underset{(0.23336)}{5.12017}\,x3$$

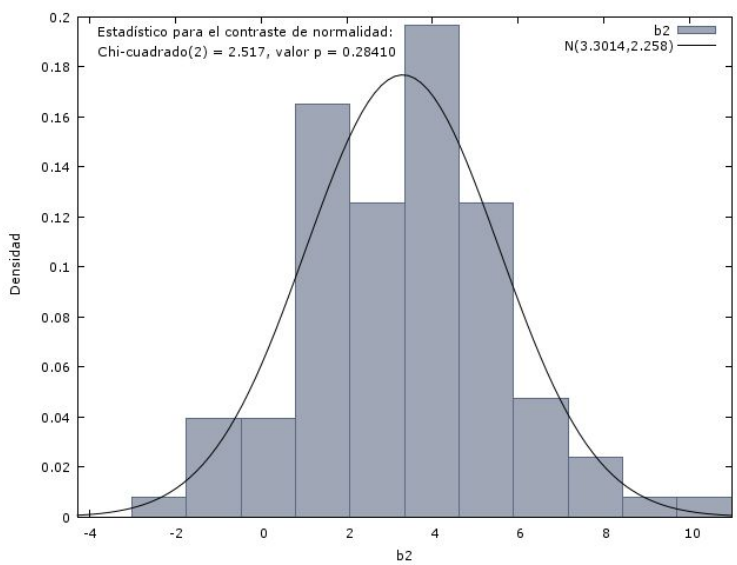
$$T = 100 \quad \bar{R}^2 = 0.9995 \quad F(2, 97) = 93948. \quad \hat{\sigma} = 8.5445 \tag{6}$$

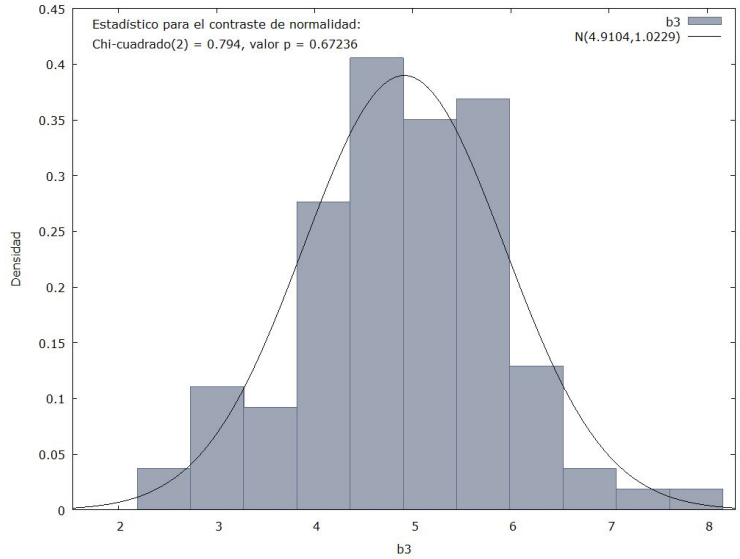
Fem un Monte Carlo per veure la distribució dels paràmetres

gretl: mC.inp

```
nulldata 100
smpl 1 30
genr x2=uniform(0,100)
set seed 123456
loop 100 --progressive
    genr u=normal(0,8)
    genr v=normal(0,4)
    genr x3=2*x2+v
    genr y=10+3*x2+5*x3+u
    ols y const x2 x3
    genr b2=$coeff(x2)
    genr b3=$coeff(x3)
    store coeff.gdt b2 b3
endloop
```

Aquí hi ha les distribucions dels paràmetres estimats





I si els coeficients no fossin correlacionats?

Una conseqüència de la Col·linealitat

L'efecte als errors estandard

Aquí veurem l'efecte d'un increment al coeficient de col·linealitat. Per tant utilitzarem els següent quadres

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

i fem una estimació a l'equació

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Els resultats utilitzant el primer quadre són

$$\underset{}{\widehat{\mathrm{Y}}} = \underset{(0.77368)}{1.19390} + \underset{(0.18481)}{0.446341}\,\mathrm{X2} + \underset{(0.085066)}{0.00304878}\,\mathrm{X3}$$

$$T = 5 \quad R^2 = 0.8101 \quad F(2,2) = 4.2665 \quad \hat{\sigma} = 0.97437$$

i utilitzant el segon:

$$\underset{}{\widehat{\mathrm{Y}}} = \underset{(0.74802)}{1.21081} + \underset{(0.27206)}{0.401351}\,\mathrm{X2} + \underset{(0.12523)}{0.0270270}\,\mathrm{X33}$$

$$T = 5 \quad R^2 = 0.8143 \quad F(2,2) = 4.3857 \quad \hat{\sigma} = 0.96352$$

En els dos casos, notem un coeficient de determinació alt!

Són les variables independents significatives?
Els valors de la distribució t són:

$$t_{2,0.20} = 1.886,\ t_{2,0.10} = 2.9,\ t_{2,0.05} = 4.3$$

El valor de l'estadistica t per la variable X_2 és

- 2.4151 en el primer cas i
- 1.4752 en el segon

Quina és la raó per la pujada de l'ee?

- mirem al coeficient de correlació entre X_2 i X_3
  1. $r_{23} = 0.5523$
  2. $r_{23} = 0.8285$
- la resposta és $var(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_{2i}^2(1-r_{23}^2)}$

Una altra conseqüència de la Col·linealitat

Utilitzant les dades co1.xls, s'ha fet la següent estimació del consum Y sobre l'ingrés X_2 i la riquesa X_3

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$\begin{gathered}\widehat{\mathrm{Y}} = \underset{(6.7525)}{24.7747} + \underset{(0.82290)}{0.941537}\,\mathrm{X2} - \underset{(0.080664)}{0.0424345}\,\mathrm{X3} \\ T = 10 \quad \bar{R}^2 = 0.9531 \quad F(2,7) = 92.402 \quad \hat{\sigma} = 6.8080\end{gathered} \tag{7}$$

Inconsistència entre les estadístiques F i t
$F > 19.3532 = F_{0.05}(2,7)$ però $t = \frac{0.941537}{0.82290} = 1.14 < t_{8,0.025}$,
$t = \frac{0.0424345}{0.080664} = 0.52 < t_{8,0.025}$

Per comprobar la correlació entre X_2 i X_3 fem la regressió

$$\widehat{\mathrm{X3}} = \underset{(29.476)}{7.54545} + \underset{(0.16426)}{10.1909}\,\mathrm{X2}$$
$$T = 10 \quad \bar{R}^2 = 0.9977 \quad F(1,8) = 3849.0 \quad \hat{\sigma} = 29.840 \tag{8}$$

Doncs, per veure l'efecte de l'ingrés sobre el consum tenim l'equació

$$\widehat{\mathrm{Y}} = \underset{(6.4138)}{24.4545} + \underset{(0.035743)}{0.509091}\,\mathrm{X2}$$
$$T = 10 \quad \bar{R}^2 = 0.9573 \quad F(1,8) = 202.87 \quad \hat{\sigma} = 6.4930 \tag{9}$$

És ara X_2 una variable signicativa?
Si féssim l'última regressió amb X_3, tindríem uns resultats semblants

Col·linealitat en un model amb més que dues variables

El Criteri de Klein

Utilitzant les dades de

- Y - número de persones què treballen
- X_1 - index de deflació del PIB
- X_2 - PIB en millons de dòlars
- X_3 - número d'aturats
- X_4 - número de persones inscrites a les forces armades
- X_5 - població no institucionalitzada, > 14 anys
- X_6 - any

s'ha fet la següent regressió:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + u$$

Modelo 1: MCO, usando las observaciones 1947–1962 ($T = 16$)
Variable dependiente: Y

	Coeficiente	Desv. Tpica	Estadstico t	Valor p
const	77270.1	22506.7	3.4332	0.0075
X1	1.50619	8.49149	0.1774	0.8631
X2	−0.0358192	0.0334910	−1.0695	0.3127
X3	−2.02023	0.488400	−4.1364	0.0025
X4	−1.03323	0.214274	−4.8220	0.0009
X5	−0.0511041	0.226073	−0.2261	0.8262
X6	1829.15	455.478	4.0159	0.0030

R^2	0.995479		
$F(6, 9)$	330.2853	Valor p (de F)	4.98e–10

Hi ha indicació de presencia de col·linealitat?

Aquí hi ha els coefficient de correlació entre les variables

Coeficients de correlació

X1	X2	X3	X4	X5	X6	
1.0000	0.9916	0.6206	0.4647	0.9792	0.9911	X1
	1.0000	0.6043	0.4464	0.9911	0.9953	X2
		1.0000	−0.1774	0.6866	0.6683	X3
			1.0000	0.3644	0.4172	X4
				1.0000	0.9940	X5
					1.0000	X6

El criteri de Klein:

Si hi ha un valor dels R^2 auxiliars que és més gran què R^2 global, potser hi ha un nivell de collinealitat que provoqui problemes

R^2 auxiliars?
-són els coeficients de determinació de regressar una de les variables independents sobre les altres

- Ex: $R^2_{1,23456}$ és el coeficient de determinació de regressar X_1 sobre X_2, X_3, X_4, X_5, X_6

El quadre amb aquests coeficients és

$R^2_{1,23456}$	0.9926
$R^2_{1,23456}$	0.9994
$R^2_{1,23456}$	0.9702
$R^2_{1,23456}$	0.7213
$R^2_{1,23456}$	0.9970
$R^2_{1,23456}$	0.9986

i el R^2 global és 995479

Segons la regla de Klein, hi ha gran possibilitat de col·linealitat. Què fem?

- Y - número de persones què treballen
- X_1 - index de deflació del PIB
- X_2 - PIB en millons de dòlars
- X_3 - número d'aturats
- X_4 - número de persones inscrites a les forces armades
- X_5 - població no institucionalitzada, > 14 anys
- X_6 - any

1. 

Segons la regla de Klein, hi ha gran possibilitat de col·linealitat. Què fem?

- Y - número de persones què treballen
- X_1 - index de deflació del PIB
- X_2 - PIB en millons de dòlars
- X_3 - número d'aturats
- X_4 - número de persones inscrites a les forces armades
- X_5 - població no institucionalitzada, > 14 anys
- X_6 - any

1. 
2. expressar el PIB en PIBR

Segons la regla de Klein, hi ha gran possibilitat de col·linealitat. Què fem?

- Y - número de persones què treballen
- X_1 - index de deflació del PIB
- X_2 - PIB en millons de dòlars
- X_3 - número d'aturats
- X_4 - número de persones inscrites a les forces armades
- X_5 - població no institucionalitzada, > 14 anys
- X_6 - any

1. 
2. expressar el PIB en PIBR
3. X_5 creix amb el temps, doncs eliminem el temps

Segons la regla de Klein, hi ha gran possibilitat de col·linealitat. Què fem?

- Y - número de persones què treballen
- X_1 - index de deflació del PIB
- X_2 - PIB en millons de dòlars
- X_3 - número d'aturats
- X_4 - número de persones inscrites a les forces armades
- X_5 - població no institucionalitzada, > 14 anys
- X_6 - any

1. 
2. expressar el PIB en PIBR
3. X_5 creix amb el temps, doncs eliminem el temps
4. eliminar també la X_3

Hem de definir la nova variable $PIBR = \frac{PIB}{DEF} = \frac{X2}{X1}$ i tornar fer l'estimació:

Modelo 2: MCO, usando las observaciones 1947–1962 ($T = 16$)
Variable dependiente: Y

	Coeficiente	Desv. Tpica	Estadstico t	Valor p
const	65720.4	10624.8	6.1856	0.0000
PIBR	97.3650	17.9155	5.4347	0.0002
X4	−0.687966	0.322238	−2.1350	0.0541
X5	−0.299537	0.141761	−2.1130	0.0562

Media de la vble. dep.	65317.00	D.T. de la vble. dep.	3511.968
Suma de cuad. residuos	3440470	D.T. de la regresin	535.4492
R^2	0.981404	R^2 corregido	0.976755
$F(3, 12)$	211.0972	Valor p (de F)	1.20e–10

Algunes problemes ara?

- aquests resultats són unes conseqüències tipiques de la col·linealitat

Per evitar-la

- fer res!
- no incloure variables per les que hi ha informació de ser relacionades
- fer regressions entre els regressors i
- eliminar una o més variables correlacionades
- transformar les variables. Ex: en forma de diferències (veurem més detalls a l'últim capítol)

Un exercici

Utilitzant unes dades anuales dels EEUU, s'han obtingut:

$$\widehat{\log \mathrm{Y}} = \underset{(1.38)}{2.81} - \underset{(0.34)}{0.53} \log \mathrm{K} + \underset{(0.14)}{0.91} \log \mathrm{L} + \underset{(0.021)}{0.047}\, \mathrm{t}$$
$$R^2 = 0.97 \quad F(1,8) = 189.8 \tag{10}$$

Y - l'index de producció, K - l'index de capital i L - l'index de treball, t - temps. Utilitzant la mateixa informació s'ha fet la regressió:

$$\widehat{\log(\mathrm{Y}/\mathrm{L})} = \underset{(0.03)}{-0.11} + \underset{(0.15)}{0.11} \log (\mathrm{K}/\mathrm{L}) + \underset{(0.006)}{0.006}\, \mathrm{t}$$
$$R^2 = 0.65 \quad F(1,8) = 19.5 \tag{11}$$

Preguntes:

1. Existeix col·linealitat en la primera equació?
2. Qual és el signe *a priori* de log K? Coinsideix el resultat amb l'expectativa
3. Hi ha una justificació per la forma de l'equació (10)?
4. Quina és la lògica de l'estimació (11)?
5. Si va haver col·linealitat en el model (10), està reduïda en la (11)?