Tema 4: **Cuestiones importantes en el Modelo Lineal General (MLG)**

Universidad Complutense de Madrid

2013

Colinealidad (I)

- La multicolinealidad es un problema que surge cuando las variables explicativas del modelo están altamente correlacionadas entre sí. Este es un problema complejo, porque en cualquier regresión las variables explicativas van a presentar algún grado de correlación.

- Matemáticamente, decimos que existe multicolinealidad cuando tenemos problemas a la hora de invertir la matriz $X^T X$
- Si $\left|X^T X\right| \approx 0 \longrightarrow$ existe multicolinealidad de grado
- Si $\left|X^T X\right| = 0 \longrightarrow$ existe multicolinealidad exacta

Colinealidad (II)

- Si la multicolinealidad es exacta, alguna variable explicativa es combinación lineal exacta de otras y el sistema de ecuaciones normales tiene infinitas soluciones. Fácil de detectar y de resolver (por ejemplo, eliminando algún regresor colineal con otro u otros).

- Si la multicolinealidad es de grado, alguna variable está altamente correlacionada con otra(s), pero el sistema de ecuaciones normales tiene una única solución. Más difícil de detectar y de resolver. Una pista para detectar este tipo de multicolinealidad es reconocer una serie de efectos perniciosos que tiene sobre los resultados de la estimación MCO.

Colinealidad (III)

(1) Las varianzas y covarianzas estimadas de los parámetros se hacen muy grandes conforme aumenta el grado de colinealidad. Es decir:

$$\hat{\text{var}}(\hat{\beta}_{MCO}) = \hat{\sigma}^2 (X^T X)^{-1} = \hat{\sigma}^2 \frac{Adj(X^T X)}{\left| X^T X \right|}$$

y al ser el determinante de la matriz $X^T X$ cercano a cero, esto infla las varianzas y covarianzas de los parámetros estimados.

Ello implica que la precisión de la estimación disminuye a medida que aumenta la colinealidad.

Colinealidad (IV)

(2)Los estadísticos *t* de significación individual estarán sesgados a la baja. Esto hará que tendamos a no rechazar la $H_0 : \beta_i = 0$ más frecuentemente de lo que se debiera si no existiese colinealidad alta:

$$t = \frac{\hat{\beta}_i}{\sqrt{\hat{\text{var}}(\hat{\beta}_i)}} \sim t_{n-k}$$

(3)El contraste de significación global de las pendientes del modelo no se verá afectado ante la presencia de multicolinealidad. Esto es así, porque el *R*-cuadrado no se ve afectado por el problema, ya que la bondad del ajuste seguirá siendo parecida ante la presencia de variables explicativas superfluas.

Colinealidad (V)

Importante: un síntoma claro de multicolinealidad de grado es que los parámetros no sean individualmente significativos, pero sí de manera conjunta. Esto es una contradicción estadística, salvo que exista un problema en los datos.

(4) Otro síntoma de multicolinealidad de grado es que ligeros cambios en las matrices de datos *X* e *Y* (por ejemplo, añadiendo o suprimiendo unas pocas observaciones) pueden llevar a grandes cambios en los parámetros estimados. Si la multicolinealidad es exacta no se arregla nada aumentando o disminuyendo la muestra con la que se trabaja.

Colinealidad (VI)

Detección: se pueden usar dos tipos de métodos:

(A) Métodos basados en la correlación muestral entre variables explicativas

(B) Métodos basados en medir el tamaño de la matriz $X^T X$

(A.1)Métodos basados en la correlación muestral entre variables explicativas: El más inmediato es calcular la correlación lineal simple existente entre pares de variables explicativas. Si hacemos esto para los *k* regresores del modelo, obtenemos una matriz *R* con la forma:

$$R = \begin{bmatrix} 1 & r_{12} & . & r_{1k} \\ r_{21} & 1 & . & r_{2k} \\ . & . & 1 & . \\ r_{k1} & r_{k2} & . & 1 \end{bmatrix}$$

Colinealidad (VII)

(A.2)Métodos basados en la correlación muestral entre variables explicativas: Otro método consiste en calcular los llamados “factores de inflación de varianza” o *VIF’s* definidos como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación de la regresión del *j*-ésimo regresor sobre el resto. El valor mínimo es 1 y un *VIF* > 10 puede indicar la existencia colinealidad.

Colinealidad (VIII)

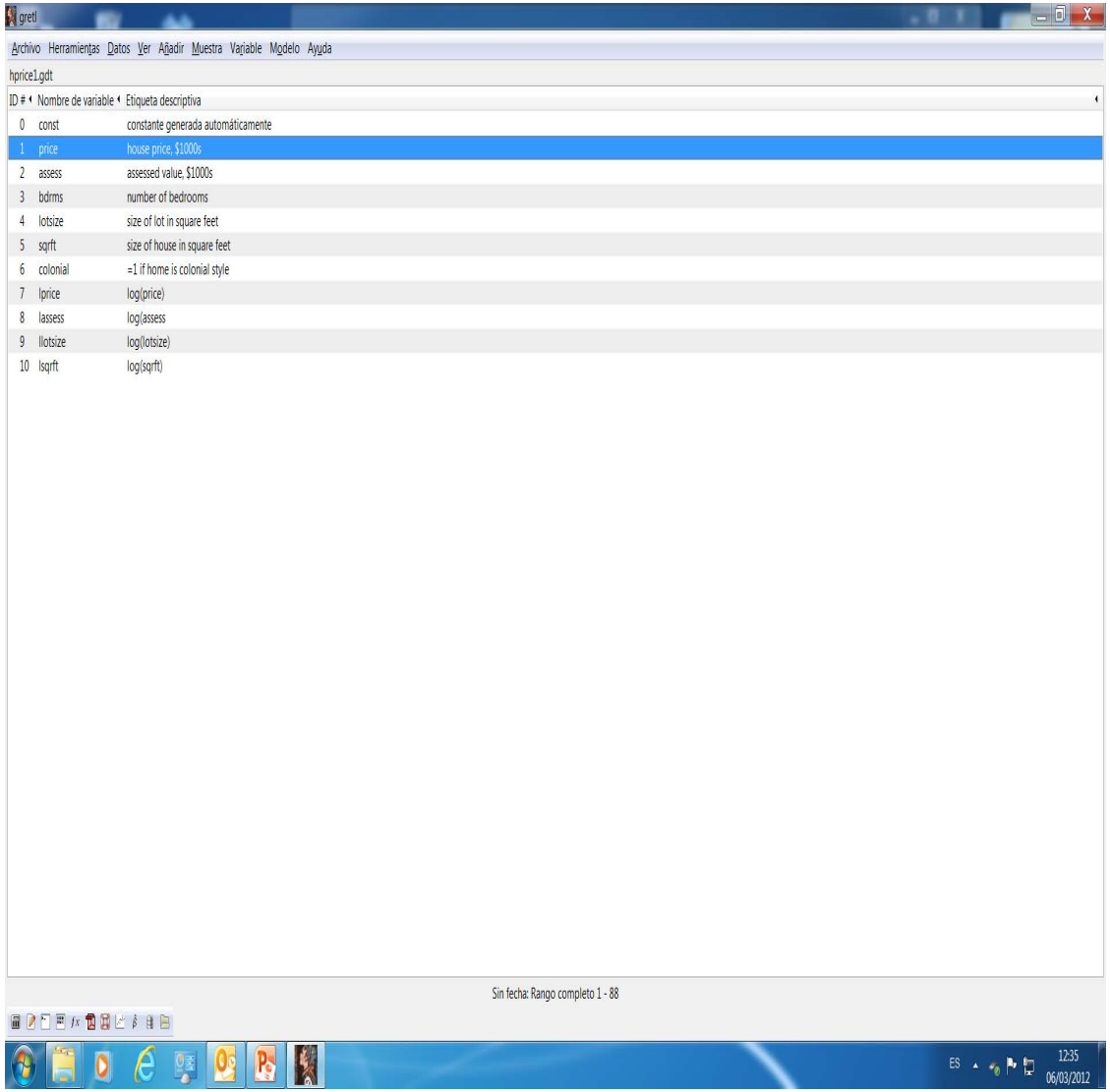
(B)Métodos basados en el tamaño de la matriz

Medir o examinar el tamaño relativo de los autovalores de la matriz $X^T X$ .De esta forma, se elimina el problema de las unidades de medida. En concreto, se calcula el número de condición de la matriz como la raíz cuadrada del cociente entre el autovalor más grande y el más pequeño.

$$\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = N^{o} \text{ de condicion}$$

Existen reglas heurísticas como que un número de condición mayor que 20 ó 25, ya sugiere la presencia de alta colinealidad entre los regresores del modelo.

Caso práctico de colinealidad (I)



Escogemos de Gretl (Archivo de datos de muestra en Wooldridge) el Data Set llamado hprice1. Contiene una sección cruzada del precio de 88 viviendas junto con algunas características físicas de las mismas (nº de habitaciones, tamaño en metros cuadrados de la casa, si es de tipo colonial o no) y su valor de tasación.

El fichero contiene además los logaritmos de las variables continuas (precio, valor de tasación, etc). No se puede tomar logaritmos a las variables ficticias (1 ó 0).

Caso práctico de colinealidad (II)

```
Modelo 2: MCO, usando las observaciones 1-88
Variable dependiente: price
             Coeficiente   Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------
const      -40,3045       21,3424          -1,888       0,0625    *
assess       0,909444      0,0584150       15,57        2,40e-026 ***
bdrms        9,74999       6,60392          1,476       0,1436
colonial     9,47922      10,5269           0,9005      0,3705
lotsize      0,000592640   0,000482637      1,228       0,2229

Media de la vble. dep.  293,5460   D.T. de la vble. dep.   102,7134
Suma de cuad. residuos  155249,8   D.T. de la regresión    43,24904
R-cuadrado              0,830856   R-cuadrado corregido    0,822704
F(4, 83)                101,9264   Valor p (de F)          3,33e-31
Log-verosimilitud      -453,7866   Criterio de Akaike      917,5731
Criterio de Schwarz     929,9598   Crit. de Hannan-Quinn   922,5634
```

Se observa que: (i) los coeficientes asociados a las características de la casa, no son estadísticamente significativos; (ii) el coeficiente asociado al valor de tasación (`assess`) no es distinto estadísticamente a uno; (iii) El R-cuadrado es alto, a pesar de no ser significativas las variables y (iv) el estadístico *F* de significación global indica que conjuntamente sí son significativas las variables.

Caso práctico de colinealidad (III)

```
Factores de inflación de varianza (VIF)
Mínimo valor posible = 1.0
Valores mayores que 10.0 pueden indicar un problema de colinealidad

         assess    4,539
          bdrms    1,556
        lotsize    1,175
          sqrft    4,527
       colonial    1,121

VIF(j) = 1/(1 - R(j)^2), donde R(j)^2 es el coeficiente de determinación en la
regresión de la variable j-ésima sobre las demás variables independientes

Propiedades de la matriz X'X:
 norma-1 = 1,8140033e+010
 Determinante = 2,8519724e+027
 Número de condición recíproca = 1,8419781e-010
```

Se observa que sobre todo son colineales la tasación (`asses`) con el resto y los metros cuadrados (`sqrft`) con el resto. El nº de condición de la matriz `X'X` es muy alto, indicando una colinealidad de grado importante.

Caso práctico de colinealidad (IV)

Métodos basados en la correlación muestral entre las variables explicativas

Coeficientes de correlación, usando las observaciones 1 - 88
valor crítico al 5% (a dos colas) = 0,2096 para n = 88

assess	bdrms	lotsize	sqrft	colonial	
1,0000	0,4825	0,3281	0,8656	0,0829	assess
	1,0000	0,1363	0,5315	0,3046	bdrms
		1,0000	0,1838	0,0140	lotsize
			1,0000	0,0654	sqrft
				1,0000	colonial

Se observa que los coeficientes de correlación muestral más altos entre pares de variables se corresponden con (tasación – metros cuadrados) y (nº de dormitorios – metros cuadrados)

No obstante, hay otros coeficientes de correlación estadísticamente distintos de cero, por ejemplo, la tasación con el nº de dormitorios.

Variables ficticias (I)

- Una variable ficticia (o *dummy*) es una variable artificial construida por el investigador que suele tomar valor 1 ó 0 y que tiene distintas utilidades en un modelo econométrico. Los usos más importantes son:
  - Para llevar a cabo los llamados contrastes de cambio estructural.
  - Para captar en los datos estacionalidad determinista.
  - A veces, disponemos de información cualitativa acerca de un conjunto de individuos, que sólo puede representarse a través de dummies. Por ejemplo, el sexo, la raza o el nivel de estudios de un individuo son características del mismo que requieren del uso de este tipo de variables.

Variables ficticias (II)

Supongamos que queremos explicar las diferencias salariales de un conjunto de individuos con distintos niveles de estudios. En un principio, definimos tantas variables ficticias como niveles de estudio, es decir:

$$E_{i1} = \begin{Bmatrix} 1 \text{ si el individuo tiene estudios primarios} \\ 0 \text{ resto de los casos} \end{Bmatrix}$$

$$E_{i2} = \begin{Bmatrix} 1 \text{ si el individuo tiene estudios secundarios} \\ 0 \text{ resto de los casos} \end{Bmatrix}$$

$$E_{i3} = \begin{Bmatrix} 1 \text{ si el individuo tiene estudios superiores} \\ 0 \text{ resto de los casos} \end{Bmatrix}$$

Variables ficticias (III)

- Denotando por W_i al salario del individuo *i*-ésimo, una primera especificación de esta función de salarios que sólo depende del nivel de estudios es:

$$W_i = \beta_1 E_{i1} + \beta_2 E_{i2} + \beta_3 E_{i3} + \varepsilon_i$$

- La interpretación que tienen los coeficientes asociados a las *dummies* es sencilla. De la ecuación anterior, podemos obtener el salario esperado (medio) de un individuo con estudios primarios, con estudios secundarios y con estudios superiores. Es decir:

$$E\left(W_i \,/\, E_{i1} = 1, E_{i2} = 0, E_{i3} = 0\right) = \beta_1$$

Variables ficticias (IV)

- El salario esperado de un individuo con estudios secundarios

$$E(W_i / E_{i1} = 0, E_{i2} = 1, E_{i3} = 0) = \beta_2$$

- El salario esperado de un individuo con estudios superiores

$$E(W_i / E_{i1} = 0, E_{i2} = 0, E_{i3} = 1) = \beta_3$$

- La diferencia $\beta_2 - \beta_3$ se interpreta como la diferencia en el salario esperado (medio) de un individuo con estudios secundarios con respecto al de un individuo con estudios superiores.

Variables ficticias (V)

- Si en la función de salarios anterior, incluyéramos un término constante, tendríamos un problema de multicolinealidad exacta, ya que para cualquier individuo se cumple que:

$$E_{i1} + E_{i2} + E_{i3} = 1, \forall i$$

- Lo que evita este problema es incluir término constante pero eliminando una de las variables ficticias, por ejemplo, la primera. El modelo sería entonces:

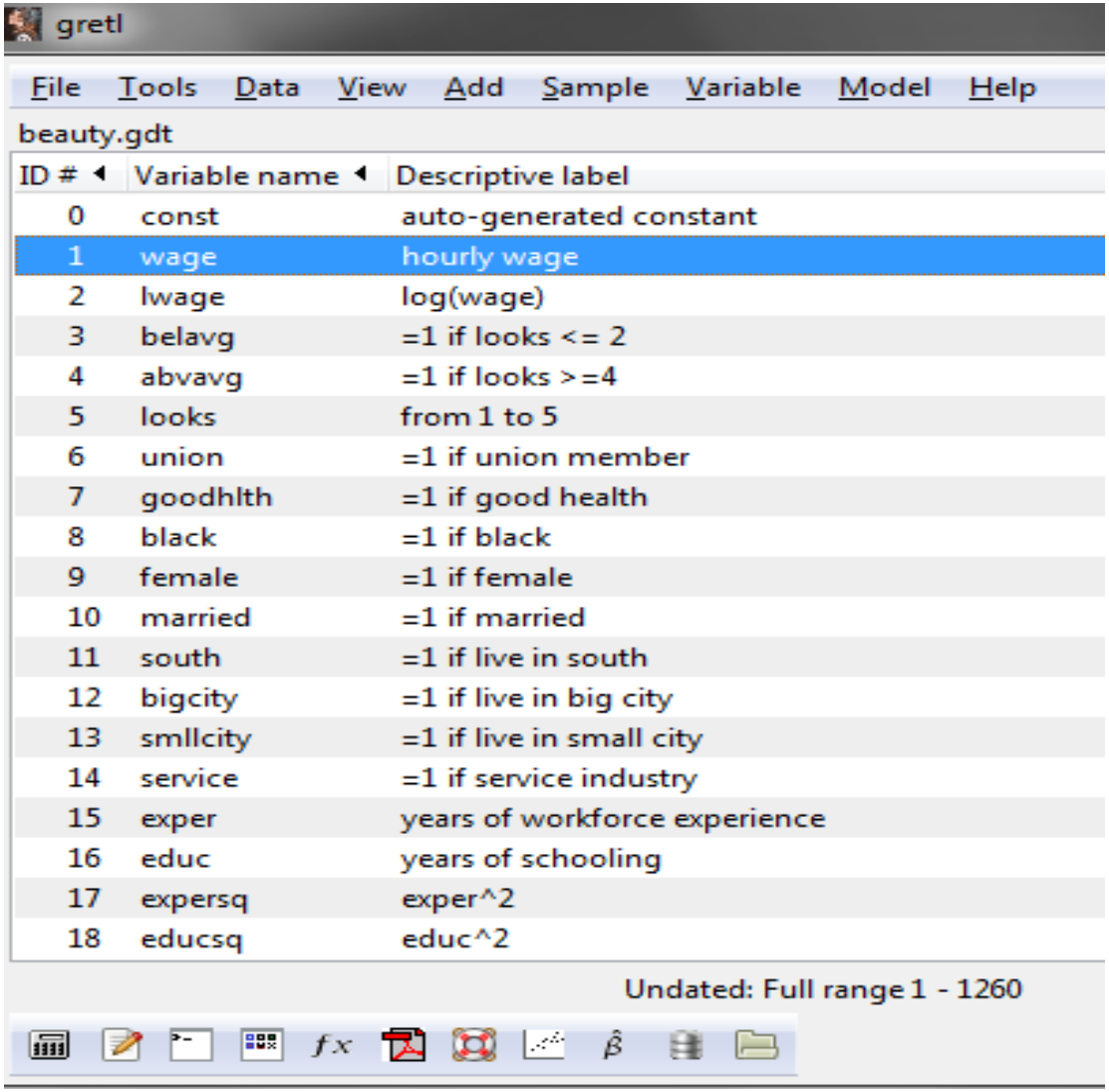
$$W_i = \alpha_1 + \alpha_2 E_{i2} + \alpha_3 E_{i3} + \varepsilon_i$$

- ¿Cuál es ahora la interpretación de los coeficientes?

Variables ficticias (VI)

- El salario esperado para un individuo con estudios primarios es ahora α_1
- El salario esperado para un individuo con estudios secundarios es $\alpha_1 + \alpha_2$
- El salario esperado para un individuo con estudios superiores es la suma de coeficientes $\alpha_1 + \alpha_3$
- Por tanto, la diferencia esperada en el salario de un individuo con estudios secundarios con respecto a uno con estudios primarios es α_2 y se espera que esta diferencia salarial sea positiva.

Uso de variables ficticias (I)



ID #	Variable name	Descriptive label
0	const	auto-generated constant
1	wage	hourly wage
2	lwage	log(wage)
3	belavg	=1 if looks <= 2
4	abvavg	=1 if looks >=4
5	looks	from 1 to 5
6	union	=1 if union member
7	goodhlth	=1 if good health
8	black	=1 if black
9	female	=1 if female
10	married	=1 if married
11	south	=1 if live in south
12	bigcity	=1 if live in big city
13	smllcity	=1 if live in small city
14	service	=1 if service industry
15	exper	years of workforce experience
16	educ	years of schooling
17	expersq	exper^2
18	educsq	educ^2

Estos datos de muestra en Gretl son del texto de Wooldridge. Recogen el salario por hora en dólares de 1260 trabajadores.

Además de factores comunes en la determinación del salario (educación, experiencia, sexo, raza, pertenencia a un sindicato, etc), lo más curioso de esta práctica es comprobar si la "belleza" física influye en el salario.
La belleza se recoge en la variable looks [con valores desde 1 (mala presencia) hasta 5 (buena presencia)].

Uso de variables ficticias (II)

```
Model 1: OLS, using observations 1-1260
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  male         7.36882      0.154242     47.77    6.20e-285 ***
  female       4.29936      0.212042     20.28    2.57e-079 ***

Mean dependent var   6.306690   S.D. dependent var   4.660639
Sum squared resid    24661.05   S.E. of regression   4.427570
R-squared            0.098232   Adjusted R-squared   0.097515
F(1, 1258)           137.0369   P-value(F)           4.05e-30
Log-likelihood      -3661.554   Akaike criterion     7327.108
Schwarz criterion    7337.386   Hannan-Quinn         7330.970
```

En este primer modelo estimado, se observa que el salario medio de un hombre (male = 1 y female = 0) en esta muestra es de 7.36882 dólares por hora.

Sin embargo, el salario medio de una mujer en esta muestra (es decir, cuando male = 0 y female = 1) es sólo de 4.29936 dólares por hora.

El salario medio de un trabajador en esta muestra (sin distinguir entre hombres y mujeres es de 6.30669 dólares por hora, es decir, la media de la v. dependiente)

Uso de variables ficticias (III)

```
Model 2: OLS, using observations 1-1260
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        7.36882      0.154242     47.77    6.20e-285 ***
  female      -3.06947      0.262207    -11.71    4.05e-030 ***

Mean dependent var   6.306690   S.D. dependent var   4.660639
Sum squared resid    24661.05   S.E. of regression   4.427570
R-squared            0.098232   Adjusted R-squared   0.097515
F(1, 1258)           137.0369   P-value(F)           4.05e-30
Log-likelihood      -3661.554   Akaike criterion     7327.108
Schwarz criterion    7337.386   Hannan-Quinn         7330.970
```

Este modelo es equivalente al anterior, al incluir término constante y eliminar una de las variables ficticias. Cambia la interpretación de los coeficientes estimados, pero no los resultados del ajuste MCO.

En este modelo, un hombre (female = 0) gana 7.37 dólares por hora en media y una mujer (female = 1) gana 7.37 - 3.07 = 4.30 dólares por hora.

Uso de variables ficticias (IV)

```
Model 3: OLS, using observations 1-1260
Dependent variable: wage
Omitted due to exact collinearity: female

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        4.29936      0.212042     20.28    2.57e-079 ***
  male         3.06947      0.262207     11.71    4.05e-030 ***

Mean dependent var   6.306690   S.D. dependent var   4.660639
Sum squared resid    24661.05   S.E. of regression   4.427570
R-squared            0.098232   Adjusted R-squared   0.097515
F(1, 1258)           137.0369   P-value(F)           4.05e-30
Log-likelihood      -3661.554   Akaike criterion     7327.108
Schwarz criterion    7337.386   Hannan-Quinn         7330.970
```

Al incluir término constante con las dos ficticias excluyentes (male y female), generamos multicolinealidad exacta. La solución es eliminar una de ellas.

Gretl da un mensaje advirtiendo del problema y elimina "`female`"

Uso de variables ficticias (V)

Se pueden combinar variables cualitativas con regresores continuos. Por ejemplo, la discriminación salarial por sexo puede deberse a que se infraestime la educación y la experiencia profesional de las mujeres.

Para contrastar este hecho, es necesario definir las variables semi-continuas siguientes:

femexp = female*exp

femeduc = female*educ

donde exp es el número de años de experiencia laboral y educ el número de años de educación.

Uso de variables ficticias (VI)

```
Model 4: OLS, using observations 1-1260
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -0.920837     0.813507     -1.132    0.2579
  exper       0.102006     0.0123301     8.273    3.30e-016 ***
  educ        0.495225     0.0562609     8.802    4.36e-018 ***
  female     -0.219410     1.35928      -0.1614   0.8718
  femexp     -0.0587202    0.0230312    -2.550    0.0109    **
  femeduc    -0.111982     0.0975468    -1.148    0.2512

Mean dependent var   6.306690   S.D. dependent var   4.660639
Sum squared resid    22099.70   S.E. of regression   4.198019
R-squared            0.191891   Adjusted R-squared   0.188669
F(5, 1254)           59.55431   P-value(F)           9.68e-56
Log-likelihood      -3592.468   Akaike criterion     7196.935
Schwarz criterion    7227.768   Hannan-Quinn         7208.522

Excluding the constant, p-value was highest for variable 9 (female)
```

Estos resultados implican que 1 año adicional de experiencia aumenta el salario en 0.102006 dólares por hora para un hombre.

No obstante, si eres mujer (female = 1), el efecto neto de un año más de experiencia es menor: 0.102006 - 0.0587202 = 0.04 aproximadamente.

Contrastes de cambio estructural (I)

- Supongamos que en una función de consumo con datos anuales, se desea contrastar si el consumo autónomo ha cambiado o no a raíz de una crisis que se produjo en un año concreto (denotado por n_1). Para ello, especificamos el siguiente modelo:

$$C_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \beta PIB_t + \varepsilon_t \quad , \quad \text{t=1,2..., } n_1, ..., n$$

- donde

$$D_{1t} = \begin{cases} 1 \text{ si } 1 \le t \le n_1 \\ 0 \text{ si } n_1 < t \le n \end{cases} \qquad D_{2t} = \begin{cases} 0 \text{ si } 1 \le t \le n_1 \\ 1 \text{ si } n_1 < t \le n \end{cases}$$

Contrastes de cambio estructural (II)

- La variable D_1 se activa en los años previos a la crisis y D_2 toma valor uno en los años posteriores a la crisis.
- Por tanto, del modelo anterior, tenemos que en los años antes de la crisis, la función de consumo es:

$$C_t = \alpha_1 + \beta PIB_t + \varepsilon_t \quad , \quad \text{t=1,2..., } n_1$$

- La función de consumo en los años posteriores a la crisis es

$$C_t = \alpha_2 + \beta PIB_t + \varepsilon_t \quad , \quad \text{t=} n_1 + 1, ..., n$$

- El contraste de igualdad del consumo autónomo antes y después de la crisis, se puede llevar a cabo usando un único modelo.

Contrastes de cambio estructural (III)

Es:

$$C_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \beta PIB_t + \varepsilon_t \quad , \quad \text{t=1,2..., } n_1, ..., n$$

donde la hipótesis nula a contrastar es $H_0 : \alpha_1 = \alpha_2$

frente a $H_1 : \alpha_1 \neq \alpha_2$ Se puede usar el estadístico *t* de Student o bien, el estadístico *F* para el contraste.

Si se rechaza la nula, decimos que la crisis ha cambiado el consumo autónomo, es decir, que ha habido un cambio estructural en este parámetro.

Si no se rechaza, decimos que no ha habido un cambio estructural en este parámetro de la función de consumo antes y después de la crisis.

Contrastes de cambio estructural (IV)

- Otro contraste de cambio estructural que nos podemos plantear, es comprobar si la crisis ha alterado la función de consumo en todos sus parámetros. Podemos escribir un único modelo para toda la muestra:

$$C_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \beta_1 (PIB_t \times D_{1t}) + \beta_2 (PIB_t \times D_{2t}) + \varepsilon_t \quad , \forall t$$

- Comprobamos que la función de consumo en los años anteriores a la crisis es:

$$C_t = \alpha_1 + \beta_1 PIB_t + \varepsilon_t \quad , \quad t=1,2\ldots, n_1$$

- y después de la crisis es:

$$C_t = \alpha_2 + \beta_2 PIB_t + \varepsilon_t \quad , \quad t=n_1+1,\ldots n$$

Contrastes de cambio estructural (V)

El contraste de cambio estructural global puede llevarse a cabo usando el modelo:

$$C_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \beta_1 (PIB_t \times D_{1t}) + \beta_2 (PIB_t \times D_{2t}) + \varepsilon_t \ , \ \forall t$$

y la hipótesis nula a contrastar es $H_0 : \alpha_1 = \alpha_2 , \beta_1 = \beta_2$

donde al haber dos restricciones, habrá que usar necesariamente el estadístico *F* para llevar a cabo el contraste. Si se rechaza la nula, decimos que ha habido un cambio estructural global en la función de consumo. Si no se rechaza la hipótesis nula, decimos que la crisis no ha cambiado estructuralmente esta función de consumo.

Contrastes de cambio estructural (VI)

Es evidente que se puede prescindir de una ficticia, ya que se cumple que $D_{1t} + D_{2t} = 1, \quad \forall t$

Si eliminamos la primera en el modelo anterior usando que $D_{1t} = 1 - D_{2t}$ se tiene:

$$C_t = \alpha_1(1 - D_{2t}) + \alpha_2 D_{2t} + \beta_1 PIB_t(1 - D_{2t}) + \beta_2 PIB_t D_{2t} + \varepsilon_t$$

y reagrupando términos, el modelo a estimar es:

$$C_t = \alpha_1 + (\alpha_2 - \alpha_1) D_{2t} + \beta_1 PIB_t + (\beta_2 - \beta_1) PIB_t D_{2t} + \varepsilon_t$$

Contrastes de cambio estructural (VII)

- Denotando por

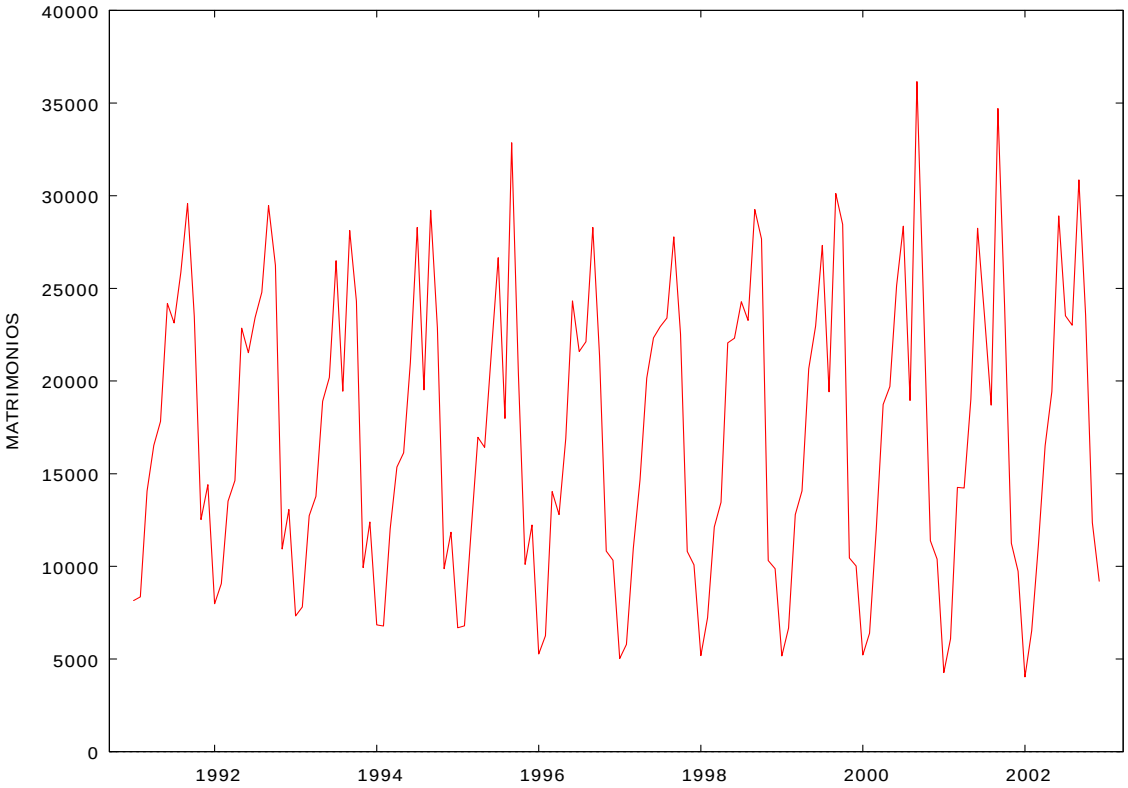
$$\gamma_1 = \alpha_2 - \alpha_1 \qquad \gamma_2 = \beta_2 - \beta_1$$

- El contraste de cambio estructural global (en la constante y la pendiente del modelo) se reduce a contrastar la hipótesis nula de que

$$H_0 : \gamma_1 = 0, \gamma_2 = 0 \qquad H_1 : \gamma_1 \neq 0, \gamma_2 \neq 0$$

- Si se rechaza la nula $\longrightarrow$ existe cambio estructural
- Si no se rechaza la nula $\longrightarrow$ no existe cambio

Tratamiento de la estacionalidad (I)



La estacionalidad en los datos es un comportamiento que se repite cada año, ligado a la estación y el clima. Cuando llega el verano, aumenta el nº matrimonios, el consumo de helados, baja el IPI al cerrar las fábricas por vacaciones, etc.

En este gráfico se muestra la evolución del número de matrimonios celebrados en España desde Enero de 1991 hasta Diciembre de 2002 (es decir, un total de 144 observaciones)

Se observan dos características: (a) una ligera tendencia a decrecer y (b) una marcada estacionalidad. Más matrimonios en los meses de primavera-verano que en otoño-invierno (el mes más alto en esta muestra es septiembre)

Tratamiento de la estacionalidad (II)

Modelo : MCO, usando las observaciones 1991:01-2002:12 (T = 144)

Variable dependiente: MATRIMONIOS (número)

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
ENERO	5930,25	576,494	10,29	1,38e-018	***
FEBRERO	6980,33	576,494	12,11	3,68e-023	***
MARZO	12638,7	576,494	21,92	7,86e-046	***
ABRIL	15144,0	576,494	26,27	2,83e-054	***
MAYO	19173,9	576,494	33,26	5,00e-066	***
JUNIO	23564,7	576,494	40,88	8,35e-077	***
JULIO	24962,8	576,494	43,30	7,00e-080	***
AGOSTO	21374,2	576,494	37,08	1,17e-071	***
SEPTIEMBRE	30533,5	576,494	52,96	8,41e-091	***
OCTUBRE	24070,3	576,494	41,75	6,19e-078	***
NOVIEMBRE	10897,8	576,494	18,90	2,24e-039	***
DICIEMBRE	11132,8	576,494	19,31	2,82e-040	***

Media de la vble. dep.	17200,28	D.T. de la vble. dep.	7794,329
Suma de cuad. residuos	5,26e+08	D.T. de la regresión	1997,033
R-cuadrado	0,939403	R-cuadrado corregido	0,934353
F(11, 132)	186,0298	Valor p (de F)	1,17e-74
Log-verosimilitud	-1292,378	Criterio de Akaike	2608,757
Criterio de Schwarz	2644,395	Crit. de Hannan-Quinn	2623,238
rho	-0,197149	Durbin-Watson	2,374971

Se puede eliminar una de las doce dummies, por ejemplo DICIEMBRE e incluir término constante. ¿Cómo cambiaría la interpretación de los parámetros estimados? ¿Y el R-cuadrado?

El nº de matrimonios en media en Enero es de 5.930, frente a los 30.533 que se celebran en Septiembre

El coeficiente asociado a cada MES, se interpreta como el nº medio de matrimonios en ese mes.

Sólo la estacionalidad explica más del 90% de la variabilidad de los matrimonios, siendo (R-cuadrado = 0.9394)
Todos los parámetros son significativos.

Introducción de términos polinómicos. Contrastes RESET (I)

En un Modelo Lineal General (MLG) como:

$$y_t = \beta_1 + \sum_{j=2}^{k} \beta_j x_{tj} + \varepsilon_t$$

puede que la variable endógena dependa de "forma no lineal" de las *x*'s. Para contrastar esta dependencia, se especifica un nuevo modelo incluyendo los "cuadrados" y "productos cruzados" de las *x*'s:

$$y_t = \beta_1 + \sum_{j=2}^{k} \beta_j x_{tj} + \sum_{j=2}^{k} \gamma_{jj} x_{tj}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{tj} x_{th} + \varepsilon_t$$

siendo la hipótesis nula que todos los coeficientes γ_{jj} y γ_{jh} sean iguales a cero. Hay $\frac{1}{2}k(k-1)$ restricciones en la nula.

Introducción de términos polinómicos. Contrastes RESET (II)

En el contraste anterior, si *k* es grande, podemos tener problemas de grados de libertad insuficientes. Por ello, este contraste se lleva a cabo de una forma más simple añadiendo un único término cuadrático en la regresión que es $\hat{y}_t^2$, sabiendo que $\hat{y}_t = x_t^T \hat{\beta}$ y que $\hat{\beta}$ es el estimador MCO del modelo original. Es decir:

$$y_t = x_t^T \beta + \gamma \hat{y}_t^2 + \varepsilon_t$$

Bajo la nula de que el modelo está correctamente especificado, se tiene que $H_0 : \gamma = 0$ y se puede llevar a cabo usando el ratio *t* de significación individual de ese parámetro. Este test se llama Contraste RESET de Ramsey y es fácil de calcular usando Gretl.

Introducción de términos polinómicos. Contrastes RESET (III)

Si se quieren contrastar no linealidades de orden más alto, el modelo libre se escribe como:

$$y_t = x_t^T \beta + \gamma_1 \hat{y}_t^2 + \gamma_2 \hat{y}_t^3 + \ldots + \gamma_p \hat{y}_t^p + \varepsilon_t$$

donde *p* es una potencia finita y se contrasta la siguiente hipótesis nula conjunta:

$$H_0 : \gamma_1 = 0, \gamma_2 = 0, \ldots, \gamma_p = 0$$

El estadístico de contraste a usar es necesariamente una *F* con grados de libertad (*p, n-k-p*).

Introducción de términos polinómicos. Contrastes RESET (IV)

Los contrastes RESET no son informativos. Es decir, cuando se rechaza la nula, no sabemos cúal es el modelo alternativo al inicial más adecuado. A veces, el modelo no lineal alternativo surge del sentido económico del modelo ó del sentido común. Por ejemplo, si en una función de salarios W_i , pensamos que a partir de un determinado nº de años de experiencia EXP_i , el salario ya no crece más (llega a un punto máximo), podemos modelizar este hecho especificando un modelo como:

$$W_i = \beta_1 + \beta_2 EXP_i + \beta_3 EXP_i^2 + \varepsilon_i$$

donde se cumple que $\dfrac{\partial W_i}{\partial EXP_i} = \beta_2 + 2\beta_3 EXP_i$

Una vez estimado el modelo, podemos igualar ésta última derivada a cero y calcular el nº de años de experiencia laboral que maximizan el salario.

Introducción de términos polinómicos. Contrastes RESET (V)

O bien, cuando pensamos que el efecto marginal que tiene la educación sobre el salario depende también de la experiencia de la persona, podemos encajar esta idea en un modelo como:

$$W_i = \beta_1 + \beta_2 EXP_i + \beta_3 EDUC_i + \beta_4 EXP_i \times EDUC_i + \varepsilon_i$$

donde derivando tenemos $\dfrac{\partial W_i}{\partial EDUC_i} = \beta_3 + \beta_4 EXP_i$

El regresor $EXP_i \times EDUC_i$ se llama término de interacción.