MÉTODOS ESTADÍSTICOS Y ECONOMÉTRICOS EN LA EMPRESA Y PARA FINANZAS

José Antonio Ordaz Sanz

María del Carmen Melgar Hiraldo

Carmen María Rubio Castaño

Departamento de Economía, Métodos Cuantitativos e Historia Económica

Universidad Pablo de Olavide

ISBN: 978-84-694-7251-4

TEMA 1

Introducción a las técnicas de Análisis Multivariante en el ámbito de la Economía y la Empresa

1.1. Métodos de Análisis Multivariante: definición y clasificación.-

El Análisis Multivariante comprende un conjunto de técnicas o métodos estadísticos cuya finalidad es analizar simultáneamente información relativa a varias variables para cada individuo o elemento estudiado. Algunos de estos métodos son puramente descriptivos de los datos muestrales, mientras que otros utilizan dichos datos muestrales para realizar inferencias acerca de parámetros poblacionales.

Entre los propósitos de estas técnicas, podemos citar, por ejemplo:

- Describir información de forma resumida.
- Agrupar observaciones o variables en subconjuntos homogéneos.
- Explorar la existencia de asociaciones entre variables.
- Explicar (o probar) comportamientos.

Existen diferentes clasificaciones de los métodos de Análisis Multivariante. Una de las más usuales distingue dos grandes grupos, según el objetivo del análisis: métodos de dependencia y métodos de interdependencia. Además, dentro de cada uno de estos grupos, la naturaleza de las variables juega un papel importante en la definición de los diversos métodos. Asimismo, cada método exige unas determinadas condiciones de aplicación para asegurar la fiabilidad de los resultados obtenidos.

Los métodos de dependencia suponen que las variables analizadas están divididas en dos grupos: las variables dependientes y las variables independientes. El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.

En cuanto a los métodos de interdependencia, éstos no distinguen entre variables dependientes e independientes, sino que tienen como objetivo identificar qué variables pueden estar relacionadas entre sí, cómo lo están y por qué.

A continuación se señala una relación de métodos, clasificados según el criterio mencionado.

 ISBN: 978-84-694-7251-4

MÉTODOS DE DEPENDENCIA

Variable(s) dependiente(s)	Variable(s) independiente(s)	
	Cuantitativa(s)	Cualitativa(s)
Cuantitativa(s)	- Regresión - Análisis factorial confirmatorio - Ecuaciones estructurales	- Regresión *dummy* - *t*-test - ANOVA - MANOVA
Cualitativa(s)	- Análisis discriminante (con 2 ó más grupos) - Probit - Logit	- Análisis discriminante *dummy* - Análisis conjunto (*conjoint*)

MÉTODOS DE INTERDEPENDENCIA

- Análisis factorial (AF)
  - AF de correlaciones (para variables cuantitativas)
  - AF de correspondencias (para variables cualitativas)
- Análisis de componentes principales
- Análisis *cluster* (o de conglomerados)

En este tema nos vamos a centrar en ofrecer las nociones fundamentales de tres técnicas de Análisis Multivariante que tienen un extenso número de aplicaciones en el ámbito de la Economía y la Empresa. En concreto, se trata del ANOVA, el Análisis discriminante y el Análisis *cluster* o de conglomerados.

Los modelos ANOVA (ANalysis Of VAriance) son técnicas de Análisis Multivariante de dependencia, que se utilizan para analizar datos procedentes de diseños con una o más variables independientes cualitativas (medidas en escalas nominales u ordinales) y una variable dependiente cuantitativa (medida con una escala de intervalo o de razón). En este contexto, las variables independientes se suelen denominar factores (y sus diferentes estados posibles o valores son *niveles* o *tratamientos*) y la variable dependiente se conoce como respuesta.

En cuanto al Análisis discriminante y al Análisis *cluster* o de conglomerados, ambos se utilizan para clasificar elementos en grupos o categorías. Sin embargo, mientras que el Análisis discriminante parte ya de grupos existentes y proporciona un medio para realizar futuras asignaciones de nuevos casos a partir de los valores de un conjunto de

 ISBN: 978-84-694-7251-4

variables independientes, el Análisis cluster no tiene preestablecidos los grupos, sino que el análisis de las variables es el que determina cómo agrupar en conglomerados a los elementos de la muestra.

A continuación, se introducirán brevemente los aspectos teóricos más importantes relativos a las técnicas referidas y se verá su aplicación práctica, mediante distintos ejemplos, a través del programa *PASW Statistics*.

1.2. El análisis de la varianza (ANOVA). ANOVA de un factor. Análisis de varianza factorial.-

El análisis de la varianza (ANOVA)

Los modelos ANOVA permiten, básicamente, comparar los valores medios que toma la variable dependiente en *J* poblaciones en las que los niveles de factores son distintos, con la finalidad de determinar si existen diferencias significativas según dichos niveles o si, por el contrario, la respuesta en cada población es independiente de los niveles de factores. Se trata, por tanto, de un contraste paramétrico que extiende al caso de *J* poblaciones el contraste de la igualdad de medias entre dos poblaciones independientes.

Algunos ejemplos de aplicación de estos modelos podrían ser los siguientes:

- Análisis de la duración media de varios tipos de lámparas.

  La variable dependiente indicaría la duración de una lámpara y el factor podría ser la potencia de la lámpara. Las poblaciones estarían formadas por lámparas con la misma potencia y habría tantas poblaciones como potencias distintas (éstas serían los niveles del factor considerado). Se trataría de contrastar la igualdad de las duraciones medias de las lámparas de cada población, para determinar si la potencia influye o no en dicha duración media.

- Estudio del efecto de varios tipos de fertilizantes sobre un cultivo.

  La variable dependiente podría ser la producción o el rendimiento de cada parcela; los factores serían, por ejemplo, el tipo de suelo de la parcela y el tipo de fertilizante utilizado. Cada población estaría formada por las parcelas que tienen un tipo concreto de suelo y en las que se utiliza un fertilizante determinado. El ANOVA se aplicaría para analizar la posible existencia de diferencias en las producciones medias de las parcelas según el tipo de suelo y el tipo de fertilizante utilizado.

Aunque existen muchos y muy diferentes modelos de ANOVA, puede obtenerse una clasificación bastante simple de los mismos atendiendo a tres criterios: el número de factores, el tipo de muestreo efectuado sobre los niveles de los factores y el tipo de aleatorización utilizada para seleccionar las muestras representativas de cada población y agrupar sus elementos (o *unidades experimentales*) en los distintos grupos que se desea comparar. Veamos esto con más detenimiento:

 ISBN: 978-84-694-7251-4

- Según el número de factores, se llama *ANOVA de un factor* al modelo en el que existe una única variable independiente; en cambio, si el modelo consta de más de un factor se habla de *modelo factorial* o *Análisis de Varianza Factorial.*

- En cuanto al muestreo de niveles, se refiere a la forma de establecer los niveles de cada factor. Esto depende, normalmente, de los intereses del investigador. Si se fijan únicamente aquellos niveles del factor que realmente interesa estudiar, estamos ante un modelo de ANOVA de *efectos fijos* (también llamado modelo I) mientras que si los niveles se seleccionan aleatoriamente de entre todos los posibles, se trata de un modelo ANOVA de *efectos aleatorios* (o modelo II).

- Las distinciones basadas en el tipo de aleatorización son equivalentes a las que se establecen al hablar de muestras independientes y muestras relacionadas. Como en todo experimento estadístico en el que no resulta posible trabajar con la población en su totalidad, se deben elegir muestras aleatorias y asignarse también aleatoriamente sus elementos a los diferentes niveles o tratamientos, para asegurar que no se cometan errores sistemáticos. Si las unidades experimentales reaccionan o responden a los tratamientos de la misma manera, se dice que son *homogéneas*. Por el contrario si responden de diferente manera a los tratamientos debido a sus diferencias intrínsecas, se dirán *heterogéneas*. Por otra parte, el tamaño de las muestras puede ser o no el mismo. Diremos que un diseño es *equilibrado o balanceado* si todas las muestras tienen el mismo tamaño y *no equilibrado o no balanceado* en caso contrario.

Como ya hemos indicado, el ANOVA trata de determinar si los niveles de factores pueden conllevar diferencias en la respuesta en los distintos grupos o poblaciones, contrastando la igualdad de medias de la variable dependiente en dichos grupos. Para ello, se basa en el estudio de la varianza.

Si dos poblaciones tienen la misma media y la misma varianza, la unión de ambas también tendrá la misma media y la misma varianza que las originales. Por tanto, es razonable pensar que, si se estimara la varianza poblacional a partir de muestras de las dos poblaciones iniciales, se obtendría un resultado similar al que se tendría si la estimación se efectuara con la unión de ambas muestras.

Sin embargo, si las dos poblaciones tienen distinta media (pero la misma varianza), al combinarlas cambiarían tanto la media de la nueva distribución como su varianza. En este caso, si se estimara la varianza poblacional a partir de una muestra extraída de las poblaciones iniciales, el resultado sería muy diferente de una estimación efectuada a partir de la muestra conjunta. Igual ocurriría si habláramos de más de dos poblaciones.

Este razonamiento es el punto de partida del ANOVA, que permite comparar las medias de varias poblaciones a partir del estudio de sus varianzas. En concreto, se analiza la relación entre las llamadas *medias cuadráticas inter-grupos* y las *medias cuadráticas intra-grupos*, que deben ser iguales si las medias de las poblaciones lo son.

 ISBN: 978-84-694-7251-4

Para poder aplicar esta técnica, deben verificarse previamente estas condiciones:

- Independencia: los individuos estudiados han de ser independientes entre sí.
- Aleatoriedad: las muestras o grupos objeto de estudio deben haberse obtenido de forma aleatoria.
- Normalidad: las muestras o grupos analizados deben seguir una distribución Normal.
- Homocedasticidad: debe haber igualdad de varianzas en las muestras o grupos estudiados.

Veremos a continuación cómo se plantea un problema con la técnica ANOVA, primero para el caso de un factor y luego para el caso factorial.

ANOVA de un factor

El análisis de la varianza de un factor se utiliza para comparar el valor medio de una variable dependiente cuantitativa en varios grupos, que se diferencian por los niveles del factor considerado.

En este apartado, se considerará un modelo de efectos fijos no equilibrado, en el que, por tanto, los tamaños muestrales no tienen por qué ser iguales.

Si denotamos por Y a la variable dependiente; J al número de muestras o grupos considerados (correspondientes cada uno a un nivel distinto del factor); $n_1, n_2, \ldots, n_J$ a los tamaños de cada una de las muestras; $n = \sum_{j=1}^{J} n_j$ al tamaño muestral total; e Y_{ij} al valor de la variable Y correspondiente a la observación i de la muestra j (para $j = 1,2,\ldots,J$ e $i = 1,2,\ldots,n_j$), la tabla siguiente resumiría los datos disponibles:

Muestras	Observaciones	Total	Medias
1	$Y_{11}\ Y_{21}\ \cdots\ Y_{i1}\ \cdots\ Y_{n_1 1}$	T_1	$\overline{Y}_1$
2	$Y_{12}\ Y_{22}\ \cdots\ Y_{i2}\ \cdots\ Y_{n_2 2}$	T_2	$\overline{Y}_2$
$\vdots$	$\vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots$	$\vdots$	$\vdots$
j	$Y_{1j}\ Y_{2j}\ \cdots\ Y_{ij}\ \cdots\ Y_{n_j j}$	T_j	$\overline{Y}_j$
$\vdots$	$\vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots$	$\vdots$	$\vdots$
J	$Y_{1J}\ Y_{2J}\ \cdots\ Y_{iJ}\ \cdots\ Y_{n_J J}$	T_J	$\overline{Y}_J$
		$T = \sum_{j=1}^{J} T_j$	$\overline{Y} = \frac{T}{n}$

La aplicación de la técnica ANOVA se basa en un contraste de hipótesis. La hipótesis nula que se contrasta en el ANOVA de un factor es que las medias poblacionales son iguales:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_J$$
$$H_1 : \text{En caso contrario}$$

Si se acepta la hipótesis nula, significará que los grupos no difieren en el valor medio de la variable dependiente y que, en consecuencia, dicho valor medio se podrá considerar independiente del factor.

Para contrastar dicha hipótesis, introducimos los conceptos de media cuadrática inter-grupos (CM_E) y de media cuadrática intra-grupos (CM_D), que vienen dados, respectivamente, por las expresiones:

$$CM_E = \frac{\sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2}{J-1} \quad \text{y} \quad CM_D = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2}{n-J}.$$

Los numeradores de cada una de estas medias cuadráticas se conocen como suma de cuadrados entre grupos, SC_E, y como suma de cuadrados dentro de grupos, SC_D. Por su parte, los denominadores son los llamados grados de libertad asociados a dichas sumas: $J-1$ y $n-J$, respectivamente.

El estadístico de prueba que utiliza ANOVA para contrastar la hipótesis nula planteada se construye a partir de los conceptos anteriores; concretamente, viene dado por:

$$F_{J-1,n-J} = \frac{CM_E}{CM_D} = \frac{\dfrac{\sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2}{J-1}}{\dfrac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2}{n-J}}.$$

Suponiendo cierta H_0, este estadístico sigue una distribución *F* de Snedecor con *J*-1 y *n-J* grados de libertad; por lo que dado un nivel de significación α, la región crítica vendrá determinada por los valores tales que $F > F_{J-1,n-J}^{1-\alpha}$, siendo $P[F \leq F_{J-1,n-J}^{1-\alpha}] = 1-\alpha$.

Ejemplo:

Consideremos cuatro compañías A, B, C y D, cuyas acciones cotizan en Bolsa y seleccionamos aleatoriamente las cotizaciones de esas acciones en diferentes instantes de tiempo. Así, para la compañía A se observa aleatoriamente la cotización en 5

 ISBN: 978-84-694-7251-4

instantes de tiempo, en la B se observa en 4 instantes, en la C en 6 y, por último, en la compañía D se observa la cotización de las acciones en 5 instantes de tiempo.

En la tabla siguiente se muestra la cotización en euros de las diferentes acciones en los instantes de tiempo seleccionados:

Factor	Observaciones	Tamaño (n_i)	Total	Medias
A	670 840 780 610 900	5	3.800	760
B	600 800 690 650	4	2.740	685
C	800 810 730 690 750 720	6	4.500	750
D	970 840 930 790 920	5	4.450	890
		$n = 20$	$T = 15.490$	$\overline{Y} = 774{,}5$

Suponiendo que se verifican las hipótesis de normalidad, aleatoriedad, independencia y homogeneidad de varianzas, se desea contrastar al nivel de significación del 1% si la cotización media de las acciones de cada una de las cuatro compañías se pueden considerar iguales.

Solución:

La hipótesis nula que se debe contrastar es:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$
$$H_1 : \text{En caso contrario}$$

La tabla ANOVA, en la que se indican las sumas de cuadrados, sus grados de libertad y las medias cuadráticas inter-grupos e intra-grupos, es en este ejemplo:

Fuente de variación	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F
Entre compañías	$SC_E = 103.395$	$J - 1 = 3$	$CM_E = \frac{SC_E}{J-1} = 34.465$	$F = \frac{CM_E}{CM_D} = 4{,}96$
Dentro de las compañías	$SC_D = 111.100$	$n - J = 16$	$CM_D = \frac{SC_D}{n-J} = 6.943{,}75$	
Total	$SC_T = 214.495$	$n - 1 = 19$		

La hipótesis nula H_0 se rechazará si $F > F_{3,16}^{0,99}$. A partir de las tablas estadísticas de la distribución *F* de Snedecor, determinamos que $F_{3,16}^{0,99} = 5{,}29$, por lo que aceptaríamos la hipótesis nula, al ser 4,96<5,29. Por tanto se puede concluir que, para un nivel de

 ISBN: 978-84-694-7251-4

significación del 1%, la cotización media coincide en las 4 compañías consideradas; esto es, que la evolución de las cotizaciones de la Bolsa es independiente de la compañía en que se analice.

Veamos a continuación cómo se resolvería este problema utilizando el paquete estadístico *PASW Statistics*.

En primer lugar, un estudio gráfico nos ayudará a formarnos una idea de la situación. A través de la sucesión de comandos *Gráficos / Generador de gráficos / Barras / Barras de error simple* y eligiendo el nivel de confianza en *Propiedades del elemento*, se llega a la representación de intervalos de confianza para las cotizaciones medias en cada empresa. Así, la *Figura 1* representa dichas medias y sus correspondientes intervalos de confianza al 99%. Gráficamente, no se observa gran diferencia en las cotizaciones medias y además los intervalos de confianza se solapan en buena medida, lo que podría indicar que los cuatro promedios no son significativamente distintos.

Sin embargo, los métodos gráficos no son, en general, definitivos. Por tanto, para poder determinar si existen diferencias y, en caso afirmativo, entre qué grupos existen, será necesario recurrir al análisis de varianza. Para llegar a la tabla ANOVA, deberemos pulsar *Analizar / Comparar medias / ANOVA de un factor* y a continuación indicamos la variable dependiente (cotización) y el factor (empresa). Podemos observar que el resultado, que se muestra en la *Figura 2*, coincide con el que se ha mostrado en la tabla de resultados anterior.

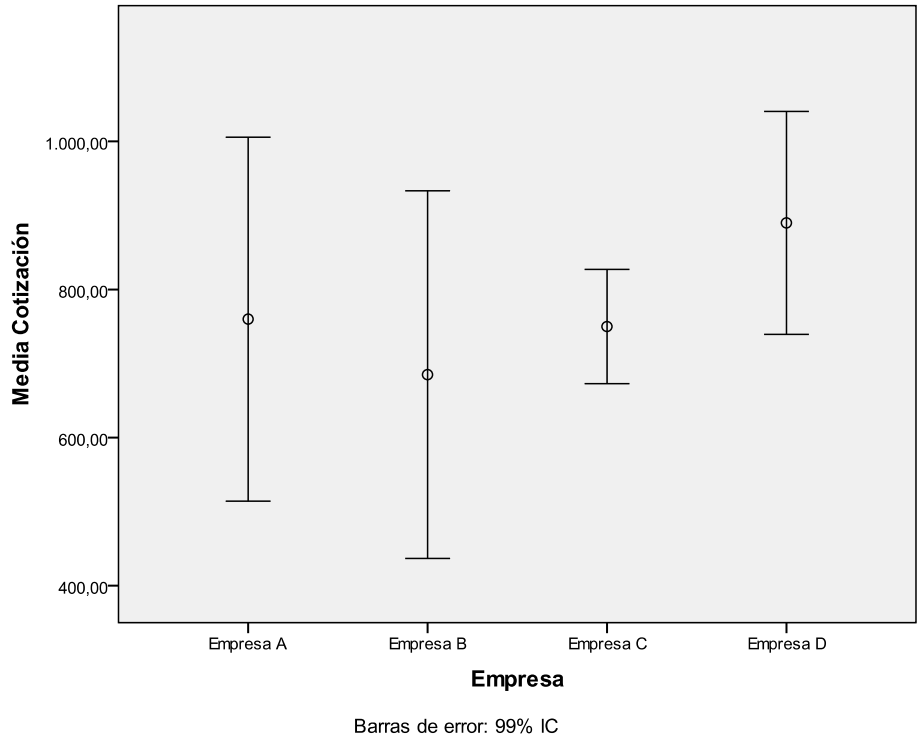


Figura 1

 ISBN: 978-84-694-7251-4

ANOVA

Colización

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	103395,000	3	34465,000	4,963	,013
Intra-grupos	111100,000	16	6943,750		
Total	214495,000	19			

Figura 2

Además, la tabla ANOVA que proporciona *PASW Statistics* nos da el *p*-valor asociado al estadístico de prueba, lo que facilita la toma de decisión en relación a la aceptación o rechazo de la hipótesis nula. Como sabemos, al ser el *p*-valor superior al nivel de significación elegido (0,013>0,01) aceptaríamos la hipótesis nula con lo que la cotización media será independiente de la empresa. Si embargo, si trabajáramos con un nivel de significación del 5%, la conclusión sería distinta, puesto que 0,013<0,05.

Una limitación importante del método que acabamos de desarrollar es que únicamente permite contrastar la hipótesis general de que los *J* promedios comparados son iguales. Sin embargo, en el caso de que se rechace esa hipótesis y por tanto las medias no sean iguales, no se podrá precisar cuáles son las muestras que tienen medias distintas. Para resolver esta cuestión, se deben utilizar otros contrastes, conocidos como *comparaciones múltiples post-hoc* o *comparaciones a posteriori*. Los métodos de este tipo que ofrece *PASW Statistics* son muy diversos y cada uno de ellos necesita de unas condiciones iniciales para su aplicación. Desarrollaremos a continuación el método de Scheffé, que tiene menos restricciones para su aplicación que los demás.

En general, este método consiste en formular un contraste sobre una combinación lineal de cualquier número de medias poblacionales. En el caso particular que nos interesa de comparación de medias, las hipótesis que se formulan para los distintos valores de j son las siguientes:

$$H_0 : \mu_{j_1} - \mu_{j_2} = 0$$
$$H_1 : \text{En caso contrario}$$

y, para ello, se utiliza el estadístico de prueba: $F_{J-1,n-J} = \dfrac{\left(\overline{Y}_{j_1} - \overline{Y}_{j_2}\right)^2}{(J-1)\dfrac{SC_D}{n-J}\left(\dfrac{1}{n_{j_1}} + \dfrac{1}{n_{j_2}}\right)}$, que,

suponiendo cierta H_0, sigue una distribución *F* de Snedecor con *J*-1 y *n*-*J* grados de libertad, por lo que dado un nivel de significación α, la región crítica vendrá determinada por los valores tales que $F > F^{1-\alpha}_{J-1,n-J}$, siendo $P[F \le F^{1-\alpha}_{J-1,n-J}] = 1-\alpha$.

 ISBN: 978-84-694-7251-4

Ejemplo:

El departamento de marketing de una empresa desea estudiar la repercusión de sus campañas publicitarias en las ventas de uno de sus productos. Se realizaron tres campañas diferentes, cada una en una provincia de la misma Comunidad Autónoma.

Las tres campañas tenían diferentes características en cuanto al medio de comunicación utilizado. La campaña A se centraba en la prensa escrita, la B en las emisoras de radio y la C en anuncios en vallas publicitarias. Durante los tres primeros meses, las cifras de ventas (en cientos de unidades) en cinco tiendas fueron las siguientes:

Medio de comunicación	Ventas (en cientos de unidades)
A (prensa)	30 20 35 42 60
B (radio)	85 73 92 86 75
C (vallas)	40 28 39 41 50

Se desea contrastar, a un nivel de confianza del 95%, si existen diferencias significativas en las cifras medias de ventas según el tipo de campaña publicitaria utilizada y, si es así, determinar entre qué tipos de campaña se dan tales diferencias.

Solución:

En este caso, podemos empezar viendo el gráfico de barras de error (*Figura 3*).

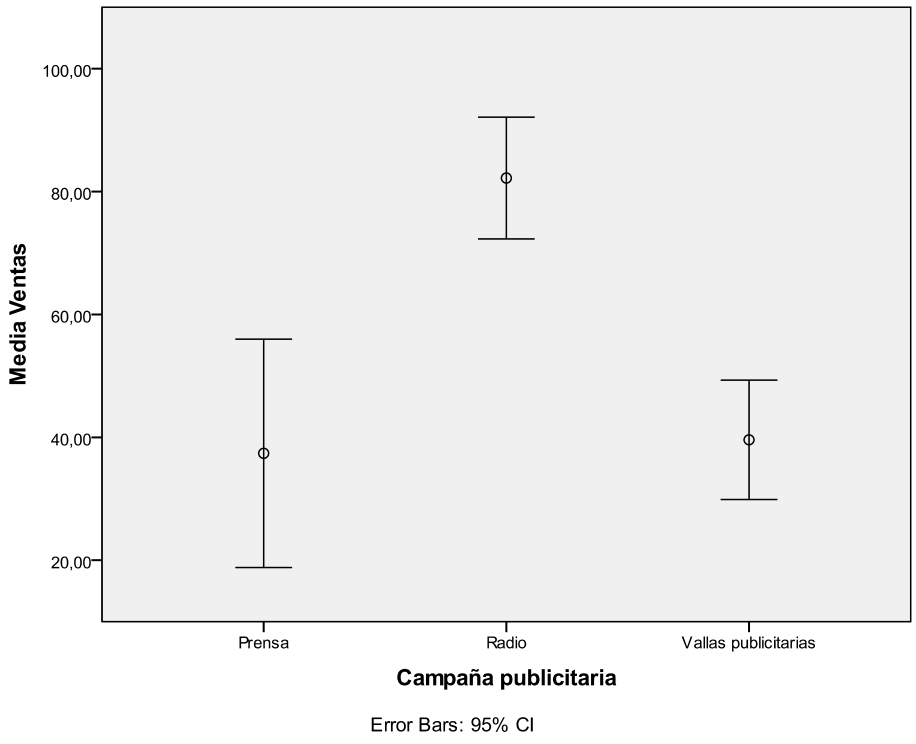


Figura 3

 ISBN: 978-84-694-7251-4

Dicho gráfico parece indicar que existen diferencias significativas entre las ventas medias en cada nivel del factor, puesto que los intervalos de los 3 no se solapan. Además, parece que las diferencias se darán cuando la campaña elegida es la "radio", pues los otros casos sí se solapan.

Antes de contrastar la hipótesis de igualdad de medias, comprobaremos si se verifican las hipótesis de aplicación del ANOVA de un factor; en concreto, la normalidad y la homoscedasticidad, puesto que los otros dos supuestos (independencia y aleatoriedad) hacen referencia a la elección de las muestras.

- Normalidad. Ésta se puede estudiar a través del test de Shapiro-Wilk (dado que el tamaño muestral es inferior a 50), que se obtiene a través de *Analizar / Estadísticos descriptivos / Explorar*, indicando la variable dependiente (ventas) y el factor (campaña publicitaria) y pulsando seguidamente en *Gráficos*, donde se elige la opción *Gráficos con prueba de normalidad*[1]. El resultado es el que nos muestra la *Figura 4*.

Pruebas de normalidad[b]

	Campaña	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Ventas	Prensa	,179	5	,200*	,971	5	,881
	Radio	,237	5	,200*	,917	5	,509
	Vallas publicitarias	,269	5	,200*	,930	5	,600

a. Corrección de la significación de Lilliefors

*. Este es un límite inferior de la significación verdadera.

b. No hay ningún caso válido para Ventas cuando Campaña = ,000. No se pueden calcular los estadísticos para este nivel.

Figura 4

En este contraste, la hipótesis nula plantea que los datos proceden de poblaciones normales. En las tres muestras (correspondientes a los tres tipos de campaña publicitaria: "prensa", "radio" y "vallas publicitarias") se acepta la hipótesis nula, dado que los *p*-valores toman, respectivamente, los valores 0,881; 0,509 y 0,600 que son mayores que 0,05, que es el nivel de significación con el que estamos trabajando.[2]

- Homoscedasticidad. En este caso, aplicamos el test de Levene, que establece como hipótesis nula la igualdad de varianzas en las distintas poblaciones. La forma de operar con *PASW Statistics* aquí es: *Analizar / Estadísticos descriptivos / Explorar*,

[1] Al elegir esta opción, junto a una serie de gráficos denominados "Gráficos Q-Q normales", *PASW Statistics* nos ofrece una tabla donde se recogen los resultados analíticos de las pruebas de normalidad. Dado que esta tabla es la que fundamentalmente nos interesa, es lo único que mostramos en la *Figura 4*, obviando los referidos gráficos.

[2] Obsérvese en la *Figura 4* que al efectuar el contraste con *PASW Statistics*, también obtenemos el resultado del test de Kolmogorov-Smirnov, que se aplica para tamaños muestrales superiores a 50.

 ISBN: 978-84-694-7251-4

indicando la variable dependiente (ventas) y el factor (campaña publicitaria) y pulsando seguidamente en *Gráficos*, donde se elige, dentro del apartado *Dispersión por nivel con prueba de Levene*, la opción *Estimación de potencia*.

Como muestra la *Figura 5*, la prueba de homogeneidad de la varianza arroja *p*-valores mayores que el nivel de significación fijado del 5%, lo que lleva a aceptar la hipótesis nula y por tanto la igualdad de varianzas (generalmente se suele utilizar la prueba efectuada *basándose en la media*).

Prueba de homogeneidad de la varianza[a]

		Estadístico de Levene	gl1	gl2	Sig.
Ventas	Basándose en la media	1,225	2	12	,328
	Basándose en la mediana.	,831	2	12	,459
	Basándose en la mediana y con gl corregido	,831	2	8,915	,467
	Basándose en la media recortada	1,186	2	12	,339

a. No hay ningún caso válido para Ventas cuando Campaña = ,000. No se pueden calcular los estadísticos para este nivel.

Figura 5

Una vez comprobadas las hipótesis necesarias para llevar a cabo el ANOVA, podemos aplicarlo. La *Figura 6* nos muestra el resultado. Para un nivel de significación del 5%, el *p*-valor resultante (0,000) nos lleva a rechazar la hipótesis nula de igualdad de medias. Concluimos, por tanto, que el tipo de campaña publicitaria utilizado repercute en las ventas medias.

ANOVA

Ventas

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	6377,733	2	3188,867	27,427	,000
Intra-grupos	1395,200	12	116,267		
Total	7772,933	14			

Figura 6

Para saber entre qué tipos de campañas publicitarias se encuentran las diferencias, llevamos a cabo un contraste de comparaciones múltiples, pulsando *Analizar / Comparar medias / ANOVA de un factor / Post hoc* y eligiendo la opción *Scheffé* en el cuadro *Asumiendo varianzas iguales*.

El resultado obtenido se muestra en la *Figura 7*, donde se ofrecen las distintas combinaciones de pares de tipos de campañas publicitarias, con el *p*-valor asociado al

 ISBN: 978-84-694-7251-4

contraste que plantea, como hipótesis nula, la nulidad de diferencia de las medias correspondientes. Se observa que existen diferencias significativas de medias entre las campañas de "radio" y "prensa" y las de "radio" y "vallas publicitarias"; en ambos casos, el *p*-valor es de 0,000, esto es, menor que el nivel de significación del 5% (además, las diferencias significativas aparecen marcadas con * por *PASW Statistics*). No sucede así, sin embargo, en el caso de la "prensa" y las "vallas publicitarias".

Comparaciones múltiples

Ventas
Scheffé

(I) Campaña	(J) Campaña	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Prensa	Radio	-44,80000*	6,81958	,000	-63,8101	-25,7899
	Vallas publicitarias	-2,20000	6,81958	,950	-21,2101	16,8101
Radio	Prensa	44,80000*	6,81958	,000	25,7899	63,8101
	Vallas publicitarias	42,60000*	6,81958	,000	23,5899	61,6101
Vallas publicitarias	Prensa	2,20000	6,81958	,950	-16,8101	21,2101
	Radio	-42,60000*	6,81958	,000	-61,6101	-23,5899

*. La diferencia de medias es significativa al nivel 0.05.

Figura 7

Otro modo alternativo de llegar a esta conclusión es comprobar si el intervalo de confianza para la diferencia de medias contiene o no al cero. Si es así, se acepta la hipótesis nula y consecuentemente no hay diferencias significativas; por el contrario, si no, se rechaza la hipótesis nula y sí se evidencian diferencias significativas.

Además, el programa *PASW Statistics* también proporciona una clasificación de los grupos considerados en subconjuntos homogéneos en cuanto a la media de la variable dependiente (*Figura 8*). Así, en nuestro ejemplo se observa que las campañas de "prensa" y "vallas publicitarias" pertenecen al mismo subconjunto (sus medias pueden considerarse iguales al nivel de significación del 5%), mientras que la campaña de "radio" forma un segundo subconjunto.

Ventas

Scheffé[a]

Campaña	N	Subconjunto para alfa = 0.05	
		1	2
Prensa	5	37,4000	
Vallas publicitarias	5	39,6000	
Radio	5		82,2000
Sig.		,950	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 5,000.

Figura 8

 ISBN: 978-84-694-7251-4

Análisis de varianza factorial

El análisis de varianza factorial permite estudiar la influencia de dos o más factores sobre la variable dependiente. En estos experimentos factoriales se pueden considerar por separado los efectos individuales de los factores y además se puede estudiar su interacción, que se introduce en el modelo de forma multiplicativa. La existencia de interacción indica por tanto que el efecto de los factores sobre la variable respuesta no es totalmente aditivo.

En un análisis de varianza factorial existe una hipótesis nula por cada factor y por cada posible combinación de factores. La hipótesis nula referida a un factor individual afirma que las medias de las poblaciones definidas por los niveles del factor son iguales; la referida al efecto de una interacción entre factores afirma que tal efecto es nulo.

Para contrastar cada una de estas hipótesis, el ANOVA factorial se sirve de estadísticos basados en la lógica ya expuesta para el ANOVA de un factor, y que siguen distribuciones de probabilidad *F* de Snedecor.

En un ANOVA factorial se trabaja con tantas poblaciones (que se suponen normales y homoscedásticas) como combinaciones haya de todos los niveles de los factores involucrados. También se asume que las observaciones han sido aleatoriamente seleccionadas (una muestra en cada población), siendo por tanto independientes entre sí.

Ejemplo:

Una subdelegación del Ministerio de Educación y Ciencia está interesada en estudiar la cantidad anual pagada por los padres de alumnos de Enseñanza Primaria en los colegios privados pertenecientes al territorio de su ámbito de competencia. Para realizar el estudio se clasificaron los colegios privados de este territorio por bloques, según su localización geográfica y según el número de alumnos por aula que los colegios afirmaban tener (considerando ésta última variable como categórica: 25 alumnos o más, o bien, menos de 25 alumnos). En cada una de las combinaciones obtenidas se seleccionó una muestra aleatoria de 3 colegios y se recogió información correspondiente a la cantidad anual (en cientos de euros) que el colegio recibía por cada alumno de Enseñanza Primaria. Los datos obtenidos se muestran en la tabla siguiente:

Alumnos \ Zona	Zona Norte	Zona Centro	Zona Sur
Menos de 25 alumnos por aula	32	25,4	50
	45,5	37,2	20,9
	28,95	23	27
25 ó más alumnos por aula	21,6	26,5	15
	25	17,2	24
	19	22	18

 ISBN: 978-84-694-7251-4

Suponiendo que se verifican las hipótesis de normalidad, independencia y homoscedasticidad, se desea saber si los colegios privados pertenecientes a esta subdelegación presentan diferencias significativas en las cantidades cobradas a los alumnos de Enseñanzas Primarias, según su localización y número de alumnos por aula.

Solución:

En este caso, se trata de un modelo con 2 factores: localización y número de alumnos por aula (por tramos). El primero de estos factores tiene 3 niveles: zona norte, zona centro y zona sur; mientras que el segundo tiene 2 niveles: menos de 25 alumnos por aula y 25 ó más alumnos por aula. De la combinación de los niveles, se obtienen 6 poblaciones que vamos a suponer normales, independientes y homoscedásticas, con las que se trabajará.

Para llevar a cabo un ANOVA factorial con *PASW Statistics* se utilizarán las especificaciones del procedimiento *Univariante* al que se llega a través de: *Analizar / Modelo Lineal General / Univariante*.

La primera de las tablas (*Figura 9*) ofrece datos generales del problema: nombre de las variables independientes (factores), sus niveles y el tamaño de cada grupo resultante.

Factores inter-sujetos

		Etiqueta del valor	N
Numalumnos	1	Menos de 25 alumnos por aula	9
	2	25 o más alumnos por aula	9
Zona	1	Zona Norte	6
	2	Zona Centro	6
	3	Zona Sur	6

Figura 9

La tabla resumen del ANOVA (*Figura 10*) contiene información similar a la que proporcionaba la tabla del modelo de un único factor: las fuentes de variación, las sumas de cuadrados, los grados de libertad, las medias cuadráticas, los estadísticos de prueba y los *p*-valores asociados a cada uno de estos estadísticos, que nos permite finalmente obtener la conclusión del contraste llevado a cabo.

Las filas correspondientes a *Numalumnos* (número de alumnos) y *Zona* recogen los efectos principales, es decir, los efectos individuales de los dos factores incluidos en el modelo: "número de alumnos por aula" y "zona en la que se encuentra el colegio". Los *p*-valores indican que, mientras los grupos definidos por la variable *número de alumnos* pagan unas cantidades medias significativamente diferentes (el *p*-valor = 0,014 < 0,05

que aparece nos lleva a rechazar la hipótesis nula de igualdad de medias), las cantidades medias pagadas en los grupos definidos por la variable *zona* no parecen diferir (el *p*-valor = 0,753 > 0,05, por lo que se acepta la hipótesis nula de igualdad de medias).

La siguiente fila (*Numalumnos*zona*) contiene información sobre el efecto *interacción* entre ambas variables. El estadístico *F* correspondiente a este efecto tiene asociado un nivel crítico de 0,714 > 0,05, lo que indica que el efecto de la interacción no es significativo.

Pruebas de los efectos inter-sujetos

Variable dependiente:Cantidad anual pagada

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	663,904[a]	5	132,781	1,885	,171
Intersección	12706,837	1	12706,837	180,437	,000
Numalumnos	574,040	1	574,040	8,151	,014
Zona	40,980	2	20,490	,291	,753
Numalumnos * Zona	48,884	2	24,442	,347	,714
Error	845,072	12	70,423		
Total	14215,813	18			
Total corregida	1508,976	17			

a. R cuadrado = ,440 (R cuadrado corregida = ,207)

Figura 10

Finalmente, es interesante observar el coeficiente que se ofrece en una nota al pie de la tabla: $R^2 = 0{,}44$. Dicho coeficiente se obtiene dividiendo la suma de cuadrados del Modelo corregido entre la suma de cuadrados Total corregida, e indica que los tres efectos incluidos en el modelo (número de alumnos, zona y su interacción, el producto de ambos: número de alumnos*zona) son capaces de predecir el 44% de la cantidad pagada.

1.3. Análisis discriminante.-

El Análisis discriminante es una técnica de Análisis Multivariante que pertenece al grupo de los métodos de dependencia. Como todos éstos, estudia la relación entre varias variables que se clasifican unas como dependientes y otras como independientes.

Partiendo de un conjunto de elementos que pertenecen a diferentes grupos previamente establecidos, se trata de analizar la información relativa a una serie de variables independientes con un doble fin:

- Explicativo: Determinar la contribución de cada variable independiente a la clasificación correcta de cada elemento.

 ISBN: 978-84-694-7251-4

- Predictivo: Determinar el grupo al que pertenece un nuevo elemento para el que se conocen los valores que toman las variables independientes.

La pertenencia de los elementos objeto de estudio a un grupo u a otro se introduce en el análisis a través de una variable cualitativa que toma tantos valores como grupos existentes. Esta variable juega el papel de variable dependiente. Las variables independientes suelen llamarse en este análisis, variables discriminantes o clasificadoras.

De acuerdo con el razonamiento que se indicará seguidamente, la información inicialmente disponible se sintetiza en las llamadas funciones discriminantes, que no son más que combinaciones lineales de las variables clasificadoras.

Antes de ello, es importante señalar que el Análisis discriminante, como todas las técnicas estadísticas, tiene que cumplir unos requisitos para su aplicación. En concreto, se trata de los siguientes:

- Los grupos deben ser mutuamente excluyentes y deben existir al menos 2. Cuando hay más de 2 grupos, se habla de Análisis discriminante múltiple.
- Para cada grupo, son necesarios más de 2 elementos o casos.
- El número de variables discriminantes a emplear no puede ser superior al número de casos menos 2.
- La variable dependiente que define los grupos ha de ser nominal.
- No puede haber relacionales lineales (multicolinealidad) entre las variables discriminantes.
- El número de funciones discriminantes que se pueden obtener viene dado por el mínimo entre J-1 y m, siendo J el número de grupos y m el número de variables clasificadoras empleadas.
- Las matrices de varianzas-covarianzas de cada grupo han de ser iguales (homoscedasticidad).
- Las variables discriminantes han de seguir una distribución normal multivariante.

Hay autores que consideran que las tres últimas hipótesis se deben contemplar de forma laxa; si no se verifican, los resultados pueden estar condicionados, pero no se invalida su calidad. Es decir, es preferible su verificación, pero no imposibilitan la aplicación del Análisis discriminante.

Existen varios procedimientos para calcular las funciones discriminantes y, a partir de ellas, asignar a los elementos entre los distintos grupos. Uno de los más utilizados es el método de Fisher, que describiremos brevemente para el caso de 2 grupos y m variables clasificadoras. Para el caso general, la idea subyacente es similar.

Como ya se ha indicado, se trata de crear, a partir de *m* variables clasificadoras que denotaremos por X_1, X_2, …, X_m, una función D, que será combinación lineal de dichas variables:

$$\boxed{D = a_1X_1 + a_2X_2 + \ldots + a_mX_m}$$

El objetivo que se persigue es que los valores de esta función se diferencien lo más posibles de un grupo a otro y sean muy parecidas para los elementos de un mismo grupo. Habrá que encontrar entonces los valores de los coeficientes a_1, a_2, …, a_m para que esto se cumpla. De este modo, se pretende reducir la dimensionalidad de las *m* variables independientes a una única dimensión, la de la combinación lineal D. Una vez creada esta función discriminante, se calculará su valor para los nuevos elementos (*puntuación discriminante*) y éstos se clasificarán en el grupo que corresponda según la puntuación obtenida.

El planteamiento del método de Fisher para hallar los coeficientes de la función discriminante consiste en maximizar la variación de la función D entre grupos, tratando al mismo tiempo, para evitar errores, de que la variación dentro de cada grupo sea lo menor posible. Atendiendo a este razonamiento, se debe maximizar la ratio:

$$\boxed{\lambda = \frac{\text{variación inter - grupos}}{\text{variación intra - grupos}}}.$$

Las variaciones inter-grupos e intra-grupos se calculan a partir de las correspondientes sumas de los cuadrados de las desviaciones de las puntuaciones con respecto a las de los *centroides*, es decir a las puntuaciones discriminantes correspondientes a valores de las variables independientes iguales a las medias de cada grupo; esto es:

$$\overline{D}_1 = a_1\overline{X_1^{(1)}} + a_2\overline{X_2^{(1)}} + \ldots + a_m\overline{X_m^{(1)}}; \qquad \overline{D}_2 = a_1\overline{X_1^{(2)}} + a_2\overline{X_2^{(2)}} + \ldots + a_m\overline{X_m^{(2)}}.$$

Así, λ puede expresarse en función de los coeficientes desconocidos a_1, a_2, …, a_m; se tratará entonces de maximizar esa función de dos variables para obtener los coeficientes de la función discriminante. La solución resultante indica que a_1, a_2, …, a_m son las coordenadas de un autovector asociado al mayor autovalor λ de cierta matriz cuyos elementos dependen únicamente de los valores observados de las variables independientes X_1, X_2, …, X_m.

Si existieran más grupos, sería necesario definir *r* funciones discriminantes (siendo *r* el mínimo entre *J*-1 y el número de variables clasificadoras *m*). En tal caso, se elegirían los autovectores asociados a los *r* mayores autovalores de la matriz.

Una vez definida la función discriminante, se debe fijar un criterio para clasificar a los nuevos elementos. Uno de ellos consiste en calcular el punto de corte discriminante (*PCD*), que no es más que la media de las puntuaciones discriminantes medias de cada grupo:

$$PCD = \frac{\overline{D^{(1)}} + \overline{D^{(2)}}}{2}.$$

y aplicar el siguiente criterio para clasificar un elemento *i*:

- Si $D_i < PCD$, se clasifica al elemento en el grupo 1.
- Si $D_i > PCD$, se clasifica al elemento en el grupo 2.

Cuando existen más de 2 grupos y, por tanto, más funciones discriminantes, el criterio para clasificar los nuevos elementos no resulta, desde el punto de vista teórico, tan evidente. Sin embargo, el procedimiento que usa el programa *PASW Statistics*, que es el que utilizaremos para desarrollar este tipo de análisis, es bastante sencillo: consiste en calcular una función de clasificación para cada grupo y asignar los elementos al grupo para el que esta función tome el valor más elevado. Estas funciones de clasificación son combinaciones lineales de las variables clasificadoras y constan, además, de un término independiente.

Veremos este procedimiento a continuación, resolviendo con *PASW Statistics* dos ejemplos: en el primero de ellos se considerarán dos grupos únicamente, mientras que en el segundo ejemplo se trabajará con tres grupos.

Ejemplo:

En un banco se tiene información acerca de 16 clientes que solicitaron préstamos instantáneos por valor de 6.000 euros cada uno. Al cabo de 3 años desde la concesión de dicho crédito había 8 clientes, de ese grupo de 16, que fueron clasificados como fallidos, mientras que los otros 8 clientes resultaron no fallidos o cumplidores, ya que reintegraron el préstamo. Para cada uno de los clientes se dispone de información sobre su patrimonio neto y su deuda pendiente correspondientes al momento de la solicitud, ambas variables medidas en miles de euros. Todo ello aparece en la siguiente tabla:

Fallidos			No fallidos		
Cliente	Patrimonio neto	Deuda pendiente	Cliente	Patrimonio neto	Deuda pendiente
1	7,80	24,60	9	31,20	6,00
2	22,20	41,40	10	58,80	25,20
3	30,00	18,00	11	54,00	28,80
4	35,40	39,00	12	72,00	12,00
5	42,60	32,40	13	37,80	31,20
6	24,00	16,20	14	52,20	6,60
7	47,40	45,60	15	66,60	24,60
8	30,60	22,80	16	59,40	9,60

 ISBN: 978-84-694-7251-4

En la mesa del director del banco hay ahora dos nuevas solicitudes de préstamo instantáneo. El primer solicitante dispone de un patrimonio neto de 60,6 (miles de euros), con deudas pendientes por valor de 40,8 (miles de euros). Para el segundo solicitante estos valores son de 58,2 y 13,2 (miles de euros) respectivamente

Se pide, mediante la aplicación del Análisis discriminante, construir una función discriminante a partir de las variables "patrimonio neto" y "deuda pendiente", que permita clasificar, con el menor error posible, a los nuevos clientes en el grupo de fallidos, o bien en el de no fallidos.

Solución:

Partiendo de las variables clasificadoras "patrimonio neto" y "deuda pendiente", se estimará 1 función discriminante.

Junto a las dos variables citadas, en *PASW Statistics* se debe crear una variable más que indique el grupo al que pertenece cada elemento. Esta variable la vamos a denominar aquí "Grupo" y le asignaremos el valor 1 para los clientes fallidos y el 2 para los no fallidos. Una vez introducidos todos los datos, si se pulsa *Analizar / Clasificar / Discriminante*, se obtendrá el siguiente cuadro de diálogo, en el que se ha elegido como variable de agrupación la variable "Grupo", que es la que indica a qué grupo pertenece cada individuo:

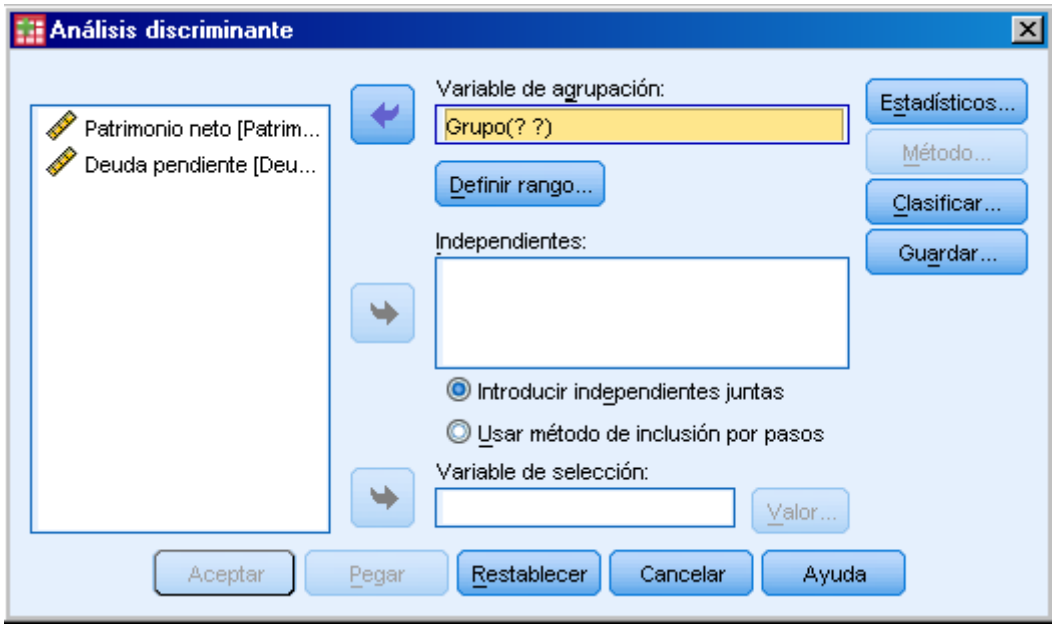


Figura 11

Como puede verse en la *Figura 11*, tras el nombre de la variable de agrupación aparecen, entre paréntesis, dos signos de interrogación. Se deberá pulsar en *Definir rango* e indicar los valores mínimo y máximo de los grupos que deseamos analizar, que son 1 y 2, respectivamente. A continuación, deberemos seleccionar las dos variables

 ISBN: 978-84-694-7251-4

independientes e introducirlas en el cuadro en blanco habilitado para ellas (*Independientes*). Seguidamente pulsaremos *Aceptar*, dejando las opciones que vienen predeterminadas por *PASW Statistics*.[3]

Las *Figuras 12* a *15* muestran algunos de los cuadros que se obtienen en el visor de resultados y que resultan de interés para nuestro propósito.

En primer lugar, se presentan algunos datos puramente descriptivos. Así, la *Figura 12* contiene un resumen de los casos, clasificándolos en válidos y perdidos. Por su parte, la *Figura 13* recoge el número de casos existentes en cada grupo. Es importante observar si existe mucha diferencia en el tamaño de los grupos, porque esto podría afectar a la clasificación; si así fuera, *PASW Statistics* nos da la opción de tenerlo en cuenta.

Resumen del procesamiento para el análisis de casos

Casos no ponderados		N	Porcentaje
Válidos		16	100,0
Excluidos	Códigos de grupo para perdidos o fuera de rango	0	,0
	Perdida al menos una variable discriminante	0	,0
	Perdidos o fuera de rango ambos, el código de grupo y al menos una de las variables discriminantes.	0	,0
	Total excluidos	0	,0
Casos Totales		16	100,0

Figura 12

Estadísticos de grupo

Grupo		N válido (según lista)	
		No ponderados	Ponderados
Fallidos	Patrimonio neto	8	8,000
	Deuda pendiente	8	8,000
No fallidos	Patrimonio neto	8	8,000
	Deuda pendiente	8	8,000
Total	Patrimonio neto	16	16,000
	Deuda pendiente	16	16,000

Figura 13

[3] En particular, se ha seleccionado *Introducir independientes juntas*, lo que significa que todas las variables independientes serán consideradas en el proceso discriminante. Si se hubiera elegido *Usar método de inclusión por pasos*, se irían seleccionando las variables independientes de mayor a menor poder discriminante y siempre que tuvieran un mínimo de poder discriminante. En lo que respecta a las opciones que pueden elegirse en *Estadísticos*, *Método*, *Clasificar* y *Guardar*, más adelante se explorarán algunas de ellas.

 ISBN: 978-84-694-7251-4

La tabla *Coeficientes estandarizados de las funciones discriminantes canónicas* (*Figura 14*) contiene la versión tipificada de los coeficientes de las funciones canónicas discriminantes. No se trata de la función discriminante que se utiliza para clasificar a los individuos en un grupo u otro (ésta se conoce como *función discriminante canónica no tipificada*). Estos coeficientes, al estar tipificados, son independientes de la métrica original de las variables independientes y permiten determinar el peso relativo de cada variable en la función discriminante (fijándonos en su valor absoluto), así como el sentido de su efecto (observando su signo).

Así, en este ejemplo puede concluirse que la variable "patrimonio neto" tiene mayor relevancia que la "deuda pendiente" a la hora de predecir el grupo de pertenencia de los individuos, puesto que su coeficiente en valor absoluto es más elevado (0,922 frente a 0,686). Y en cuanto a la interpretación exacta de los signos, es preciso conocer el signo de las puntuaciones de los centroides de cada grupo, es decir el signo de la función discriminante correspondiente a los valores medios de cada variable clasificadora (*Figura 15*). Según esto, en el presente ejemplo los signos indican que el grupo de los clientes "fallidos" se encuentra localizado, en promedio, en las puntuaciones negativas de la función, mientras que los clientes "no fallidos" se hallan en las positivas.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función
	1
Patrimonio neto	,922
Deuda pendiente	-,686

Figura 14

Funciones en los centroides de los grupos

Grupo	Función
	1
Fallidos	-1,225
No fallidos	1,225

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Figura 15

En general, se dirá que si la variable clasificadora toma un valor por encima de la media, el individuo se clasificará en el grupo en el que el signo de la puntuación del centroide coincida con el signo del coeficiente de la variable. De este modo, se puede

 ISBN: 978-84-694-7251-4

afirmar que un patrimonio neto por encima de la media[4] hace más probable la obtención de una puntuación discriminante positiva (al ser positivo el coeficiente) y, de esta manera, se ajustará más al patrón de los clientes "no fallidos" (ya que para éstos la puntuación del centroide es positiva). Por el contrario, una deuda pendiente por encima de la media propiciará una puntuación discriminante negativa (puesto que el coeficiente asociado a esta variable es negativo) y esto llevará a clasificar al individuo entre los "fallidos" (cuyo centroide tiene puntuación negativa).

Seguidamente se van a mostrar algunas opciones de *PASW Statistics* que se pueden elegir dentro del Análisis discriminante y que nos aportarán información determinante para clasificar nuevos individuos en los grupos existentes, así como para estudiar la fiabilidad de los resultados.

Como se recordará, en el cuadro de diálogo que se obtenía tras pulsar *Analizar / Clasificar / Discriminante* aparecían, entre otros, los botones *Estadísticos* y *Clasificar* (*Figura 11*).

Si dentro de *Estadísticos* se eligen como estadísticos descriptivos *ANOVAs univariados* y *M de Box* y como coeficientes de la función *De Fisher* y *No tipificados*, se obtendrán, además de los resultados ya descritos, los que se muestran en las *Figuras 16* a *19*.

La *Figura 16* proporciona los resultados de la aplicación de ANOVA a cada variable clasificadora, de manera que se puede contrastar, para cada una de ellas, la igualdad de medias en los dos grupos. En lo que se refiere al "patrimonio neto", su *p*-valor asociado nos lleva a rechazar la hipótesis nula, lo que significará que el patrimonio neto medio es distinto para fallidos y no fallidos; la conclusión sería la misma en lo que respecta a la "deuda pendiente", para un nivel de significación mínimo del 4,4%. Este hecho constituye un indicio de que las dos variables tienen poder discriminante y por tanto deben introducirse como tales en el análisis. Por el contrario, si no se observaran diferencias de medias entre los grupos para alguna de las variables clasificadoras, quizás no sería necesario incluirla en el modelo.

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
Patrimonio neto	,510	13,433	1	14	,003
Deuda pendiente	,740	4,910	1	14	,044

Figura 16

En cuanto a la prueba *M de Box*, se utiliza para contrastar la hipótesis nula de igualdad de las matrices de varianzas-covarianzas de los grupos que, como ya se comentó, es uno de los requisitos para la aplicación del Análisis discriminante. Dicho contraste se lleva a

[4] Los valores medios de las variables pueden fácilmente conocerse en *PASW Statistics* llevando a cabo un análisis descriptivo de las mismas.

cabo utilizando el estadístico *M de Box* (0,951) que muestra la *Figura 17*. Su *p*-valor asociado vale 0,849, lo que lleva a aceptar la hipótesis nula de que las matrices de varianzas-covarianzas son iguales.

Resultados de la prueba

M de Box		,951
F	Aprox.	,268
	gl1	3
	gl2	35280,000
	Sig.	,849

Contrasta la hipótesis nula de que las matrices de covarianzas poblacionales son iguales.

Figura 17

A continuación, podemos observar los coeficientes de la función de clasificación para cada grupo (*Figura 18*) que también se suelen denominar *funciones discriminantes lineales de Fisher*. Estos coeficientes se emplean únicamente para clasificar a los nuevos individuos en alguno de los grupos ya existentes. Para ello, se calcula el valor de las dos funciones (una por grupo) y el individuo se clasificará en el grupo para el que se obtenga una mayor puntuación.

Coeficientes de la función de clasificación

	Grupo	
	Fallidos	No fallidos
Patrimonio neto	,130	,302
Deuda pendiente	,216	,061
(Constante)	-5,876	-9,396

Funciones discriminantes lineales de Fisher

Figura 18

De acuerdo con todo lo expuesto hasta ahora, procedamos a clasificar a los nuevos solicitantes de préstamo. Recordemos que el primero disponía de un patrimonio neto de 60,6 (miles de euros) y tenía deudas pendientes por valor de 40,8 (miles de euros); por su parte, para el segundo solicitante estos valores eran de 58,2 y 13,2 respectivamente.

La función de clasificación para el grupo de "fallidos" sería:

$$0{,}130 * \text{Patrimonio neto} + 0{,}216 * \text{Deuda pendiente} - 5{,}876 .$$

Según esto, para el solicitante 1 esta función valdría: 10,8148; y para el solicitante 2 sería: 4,5412.

En cuanto a la función de clasificación para los "no fallidos", ésta vendría dada por:

 ISBN: 978-84-694-7251-4

$$0{,}302 * \text{Patrimonio neto} + 0{,}61 * \text{Deuda pendiente} - 9{,}396 .$$

La puntuación del solicitante 1 en este caso sería: 11,394; y para el solicitante 2: 8,9856.

Como podemos ver, ambos solicitantes obtienen mayores puntuaciones en la segunda función, por lo que los dos se clasificarán en el grupo de los clientes "no fallidos".

La última de las opciones elegidas en el cuadro *Estadísticos* nos da los coeficientes de la función canónica discriminante (*Figura 19*). Éstos son los coeficientes que el programa utiliza para clasificar a los individuos, calculando las puntuaciones y comparándolas con el punto medio de los centroides. Sin embargo, para nosotros no es relevante, puesto que no las utilizaremos para la clasificación y, además, al tratarse de coeficientes no tipificados, pueden estar afectados por las unidades de medidas de las variables independientes, lo que dificulta su interpretación.

Coeficientes de las funciones canónicas discriminantes

	Función
	1
Patrimonio neto	,070
Deuda pendiente	-,063
(Constante)	-1,437

Coeficientes no tipificados

Figura 19

Para terminar, vamos a examinar algunas de las opciones disponibles en la opción *Clasificar* del Análisis discriminante; en concreto *Probabilidades previas* (donde marcaremos *Todos los grupos iguales*[5]) y *Visualización* (aquí elegiremos *Resultados para cada caso* y *Tabla de resumen*).

La primera tabla resultante (*Figura 20*) indica simplemente el porcentaje de los casos totales que pertenecen al grupo "fallidos" y al grupo "no fallidos", bajo la denominación *Probabilidades previas*. Vendría a ser una referencia inicial, en el sentido de que si eligiésemos un cliente al azar y lo clasificásemos sistemáticamente como perteneciente al grupo de los "fallidos", acertaríamos en el 50% de los casos, ya que ése es el porcentaje de clientes de la muestra inicial que se encuentran en ese grupo (lo mismo ocurriría, en este caso, con los "no fallidos").

La aplicación del Análisis discriminante resultará tanto mejor en cuanto se incremente el porcentaje de aciertos.

[5] Marcamos esta opción porque así es en nuestro ejemplo. Si las muestras tuvieran tamaños distintos, habría que elegir *Calcular según tamaños de grupos*.

 ISBN: 978-84-694-7251-4

Probabilidades previas para los grupos

Grupo	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
Fallidos	,500	8	8,000
No fallidos	,500	8	8,000
Total	1,000	16	16,000

Figura 20

Los *Resultados para cada caso* se recogen en la *Figura 21*. Para cada cliente de la muestra inicial, se señala el grupo real al que pertenece, el pronosticado, si ha habido error en la predicción (se indica con **) y la probabilidad de que cada caso pertenezca a cada grupo condicionada a la distancia existente al centroide de cada grupo. Como se puede observar, ha habido únicamente un cliente mal clasificado: el número 13.

Estadísticos por casos

Número de caso		Grupo real	Grupo mayor					Segundo grupo mayor			Puntuaciones discriminantes
			Grupo pronosticado	P(D>d | G=g) p	P(D>d | G=g) gl	P(G=g | D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	Grupo	P(G=g | D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	Función 1
Original	1	1	1	,222	1	,998	1,491	2	,002	13,479	-2,446
	2	1	1	,203	1	,998	1,617	2	,002	13,854	-2,497
	3	1	1	,447	1	,757	,578	2	,243	2,856	-,465
	4	1	1	,849	1	,970	,036	2	,030	6,972	-1,415
	5	1	1	,462	1	,769	,540	2	,231	2,942	-,490
	6	1	1	,651	1	,869	,204	2	,131	3,994	-,773
	7	1	1	,813	1	,919	,056	2	,081	4,901	-,989
	8	1	1	,618	1	,856	,249	2	,144	3,810	-,727
	9	2	2	,398	1	,717	,714	1	,283	2,577	,380
	10	2	2	,906	1	,938	,014	1	,062	5,439	1,107
	11	2	2	,494	1	,790	,468	1	,210	3,119	,541
	12	2	2	,099	1	,999	2,715	1	,001	16,795	2,873
	13	2	1**	,636	1	,863	,224	2	,137	3,909	-,752
	14	2	2	,551	1	,989	,355	1	,011	9,279	1,821
	15	2	2	,639	1	,985	,220	1	,015	8,523	1,694
	16	2	2	,361	1	,995	,833	1	,005	11,310	2,138

**. Caso mal clasificado

Figura 21

Por último, la *Tabla de resumen*, también llamada *Matriz de confusión* se muestra en la *Figura 22*. En ella pueden apreciarse los aciertos y errores obtenidos en la clasificación realizada con la función discriminante calculada. De los 8 clientes "fallidos", los 8 se han pronosticado en ese grupo (100% de aciertos), mientras que 1 de los "no fallidos" se ha clasificado erróneamente como "fallido" (87,5% de aciertos). En total, la clasificación ha acertado en: 8 + 7 = 15 casos, lo que representa un 93,8% del total y significa que el poder discriminante de las variables independientes consideradas resulta muy alto.

 ISBN: 978-84-694-7251-4

Resultados de la clasificación[a]

		Grupo	Grupo de pertenencia pronosticado		
			Fallidos	No fallidos	Total
Original	Recuento	Fallidos	8	0	8
		No fallidos	1	7	8
	%	Fallidos	100,0	,0	100,0
		No fallidos	12,5	87,5	100,0

a. Clasificados correctamente el 93,8% de los casos agrupados originales.

Figura 22

A continuación, se resolverá un nuevo ejemplo de Análisis discriminante con *PASW Statistics*, esta vez con tres grupos.

Ejemplo:

Un banco ordena un estudio que permita identificar con la mayor precisión posible aquellas solicitudes de préstamos que probablemente puedan llegar a convertirse en morosos o fallidos en el caso que se concedieran. Para ello, dispone de la información reflejada en la tabla que se ofrece más abajo, relativa a 25 clientes y a las variables que se definen seguidamente:

- *Categoría*: grado de cumplimiento del cliente en el reintegro del préstamo. Toma el valor 1 si el cliente es cumplidor; 2 si el cliente es moroso; 3 si el cliente es fallido.

- *Ingresos*: ingresos anuales del cliente, en miles de euros.

- *Patrneto*: patrimonio neto del cliente, en miles de euros.

- *Proviv*: variable dicotómica que toma el valor 1 si el cliente es propietario de la vivienda que habita; 0 en caso contrario.

- *Casado*: variable dicotómica que toma el valor 1 si el cliente está casado; 0 en caso contrario.

- *Salfij*: variable dicotómica que toma el valor 1 si el cliente es asalariado con contrato fijo; 0 en caso contrario.

Cliente	*Categoría*	*Ingresos*	*Patrneto*	*Proviv*	*Casado*	*Salfij*
1	1	32,7	336	1	1	0
2	1	18,6	204	1	0	1
3	1	24,6	138	0	1	1
4	1	37,2	270	1	0	1
5	1	23,7	114	1	1	1
6	1	7,5	132	1	1	1
7	1	29,4	90	0	1	1
8	1	53,4	228	1	1	1
9	1	20,1	324	0	1	1
10	1	31,2	480	1	1	0

 ISBN: 978-84-694-7251-4

11	1	17,1	108	1	1	1
12	1	39	132	1	1	1
13	1	45,6	216	1	1	1
14	2	26,1	234	1	1	0
15	2	8,1	48	0	1	1
16	2	12,6	114	0	0	1
17	2	8,7	150	1	0	1
18	2	38,4	24	0	1	1
19	2	22,8	114	1	1	0
20	2	14,7	60	0	1	1
21	3	19,8	42	0	1	0
22	3	5,1	72	0	1	0
23	3	7,2	30	1	1	1
24	3	11,1	36	1	0	0
25	3	15,9	150	0	0	0

Solución:

En este caso, se trata de aplicar el Análisis discriminante múltiple, ya que el banco ha clasificado a sus clientes en tres grupos. Habrá que construir funciones de clasificación que permitan clasificar, con los menores errores posibles, a los clientes en los diferentes grupos. Si se obtienen buenos resultados, estas funciones se podrán utilizar para analizar si se concede o no un préstamo a un futuro solicitante.

Como ya sabemos, en *Analizar / Clasificar / Discriminante* se obtiene un cuadro de diálogo en el que tenemos que seleccionar la variable de agrupación (cuyo rango es ahora *1 3*) y las variables independientes. Asimismo podemos elegir las opciones adecuadas para los resultados que deseamos analizar.

Para cada variable clasificadora contrastamos la igualdad de medias entre los grupos, para tratar de determinar si las variables serán realmente discriminantes. Los *ANOVAs* de la *Figura 23* nos indican que no se observan diferencias significativas entre los "cumplidores", "morosos" y "fallidos", en cuanto al hecho de ser propietario o no de la vivienda que habitan (*Proviv*) y de estar o no casado (*Casado*). Por tanto estas variables no deberían tener una gran influencia a la hora de clasificar a los clientes en uno u otro grupo.

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
Ingresos	,688	4,990	2	22	,016
Patrneto	,663	5,584	2	22	,011
Proviv	,870	1,639	2	22	,217
Casado	,948	,609	2	22	,553
Salfij	,721	4,262	2	22	,027

Figura 23

En este punto, podemos dar respuesta ya a la petición del banco calculando las funciones de clasificación para cada grupo. La *Figura 24* muestra los coeficientes de

 ISBN: 978-84-694-7251-4

cada una de ellas, para el grupo de clientes "cumplidores", para los "morosos" y para los "fallidos".

Coeficientes de la función de clasificación

	Categ		
	Cliente cumplidor	Cliente moroso	Cliente fallido
Ingresos	,201	,131	,071
Patrneto	,076	,050	,025
Proviv	5,074	3,347	2,562
Casado	9,363	7,054	4,873
Salfij	19,210	13,563	6,357
(Constante)	-25,768	-13,229	-5,467

Funciones discriminantes lineales de Fisher

Figura 24

De acuerdo con los coeficientes estimados, se obtienen las siguientes funciones:

- Clientes "cumplidores":

$$0{,}201 * \text{Ingresos} + 0{,}076 * \text{Patrneto} + 5{,}074 * \text{Proviv} + 9{,}363 * \text{Casado} + 19{,}210 * \text{Salfij} - 25{,}768$$

- Clientes "morosos":

$$0{,}131 * \text{Ingresos} + 0{,}050 * \text{Patrneto} + 3{,}347 * \text{Proviv} + 7{,}054 * \text{Casado} + 13{,}563 * \text{Salfij} - 13{,}229$$

- Clientes "fallidos":

$$0{,}071 * \text{Ingresos} + 0{,}025 * \text{Patrneto} + 2{,}562 * \text{Proviv} + 4{,}873 * \text{Casado} + 6{,}357 * \text{Salfij} - 5{,}467$$

Cuando el banco reciba una nueva solicitud de préstamo, podrá determinar a qué grupo puede pertenecer el cliente evaluando las tres funciones y asignándolo al grupo para el que se haya obtenido una mayor puntuación.

El poder predictivo de estas funciones de clasificación se puede valorar a través de la Tabla de resumen (*Figura 25*) que ofrece *PASW Statistics*.

Resultados de la clasificación[a]

		Categ	Grupo de pertenencia pronosticado			
			Cliente cumplidor	Cliente moroso	Cliente fallido	Total
Original	Recuento	Cliente cumplidor	12	1	0	13
		Cliente moroso	0	6	1	7
		Cliente fallido	0	1	4	5
	%	Cliente cumplidor	92,3	7,7	,0	100,0
		Cliente moroso	,0	85,7	14,3	100,0
		Cliente fallido	,0	20,0	80,0	100,0

a. Clasificados correctamente el 88,0% de los casos agrupados originales.

Figura 25

 ISBN: 978-84-694-7251-4

Como puede observarse, se han clasificado correctamente 12 + 6 + 4 clientes, o sea 22 de los 25 que conformaban la muestra inicial. Esto representa un 88,0% de aciertos, lo que nos lleva a afirmar que nuestro modelo es bastante bueno.

1.4. Análisis *cluster* o de conglomerados.-

A diferencia del Análisis discriminante, el Análisis *cluster* o de conglomerados es una técnica de Análisis multivariante de interdependencia. No distingue por tanto entre variables dependientes e independientes, sino que, dado un conjunto de variables (las variables de decisión), analizará la información contenida en ellas para clasificar a los elementos según su similitud en conglomerados, los cuales deben ser entre sí lo más distintos posible. Aquí no se parte de grupos previamente establecidos para la muestra, como se hace en el Análisis discriminante. Se trata de un análisis meramente descriptivo, que no tiene bases estadísticas sobre las que se puedan deducir inferencias para la población a partir de una muestra.

El Análisis cluster es ampliamente usado en diversas disciplinas. Por ejemplo, en el ámbito del mundo empresarial esta técnica es comúnmente usada en Marketing para, por ejemplo, dividir el mercado potencial de un nuevo producto en grupos, cada uno de los cuales estaría formado por consumidores homogéneos en base a una serie de características, facilitando así el diseño de políticas comerciales.

En la realización de un Análisis cluster se suelen distinguir tres etapas:

1) Elección de variables relevantes y su tratamiento.
2) Elección de la medida de proximidad entre elementos.
3) Criterio para agrupar elementos en conglomerados.

Las decisiones que se tomen en estas etapas determinarán la clasificación resultante, de forma que no es posible hablar de una clasificación idónea. A continuación, se describirá brevemente la tarea a realizar en cada etapa.

1) Elección de variables relevantes y su tratamiento.

La clasificación final dependerá de las variables de decisión que se introduzcan en el análisis, por lo que su elección es de vital importancia para la obtención de una correcta clasificación. Será necesario, por tanto, seleccionar las variables que sean útiles para el propósito planteado. En lo que se refiere al número de variables, si éste es excesivo, aumentarán los cálculos necesarios y podría complicarse la interpretación de los resultados. Para simplificar el número de variables existen distintas soluciones, como las técnicas de reducción de datos; es el caso, por ejemplo, del *Análisis de componentes principales*, que selecciona únicamente los primeros factores (los que explican un mayor porcentaje de la varianza) para su introducción como variables de entrada en el Análisis cluster.

 ISBN: 978-84-694-7251-4

La siguiente cuestión a abordar es la estandarización de los datos, proceso que consiste en transformar las unidades de medida de las variables para que éstas pasen a estar expresadas en unidades adimensionales. Así se contribuye a que el cálculo de "similitudes" sea más equilibrado. Sin embargo, esto no siempre es necesario, e incluso hay autores que son reacios a su utilización. El programa *PASW Statistics* ofrece distintas posibilidades de estandarización de los datos.

2) Elección de la medida de proximidad entre elementos.

El Análisis cluster intenta que los conglomerados sean exhaustivos, mutuamente excluyentes y lo más homogéneos posible, de manera que los casos que pueden ser considerados similares sean asignados a un mismo grupo, mientras que los considerados distintos estarán en grupos diferentes. Es necesario por tanto dar una definición de "similitud" de los casos o de "distancia" entre ellos.

Existe una gran variedad de medidas de distancia, basadas en los valores de las variables de decisión, estando condicionada la elección de una u otra por la escala de medida que adoptan dichas variables. Si los datos están estandarizados, en general una de las medidas más utilizada es la distancia euclídea al cuadrado.

Las medidas de distancia se recogerán en la denominada *matriz de semejanzas, proximidades o distancias*. Se trata de una matriz simétrica en la que cada elemento determina la distancia entre los pares de individuos correspondientes a la fila y columna donde se ubica dicho elemento.

3) Criterio para agrupar elementos en conglomerados.

El siguiente paso consiste en elegir las reglas que determinan el modo de agrupar los individuos en conglomerados. Las posibilidades que se pueden plantear son muy diversas y ninguna es manifiestamente mejor que las demás, por lo que el analista se verá obligado a emplear distintos métodos con el objeto de contrastar los resultados. En general, los métodos de agrupamiento se suelen dividir en dos grandes grupos: *métodos jerárquicos* y *métodos no-jerárquicos*.

Métodos jerárquicos: contemplan todas las agrupaciones posibles, incluyendo las más extremas de un solo conglomerado formado por todos los individuos y la de *n* conglomerados diferentes formados cada uno por un único individuo.

Existen dos tipos de técnicas jerárquicas: las aglomerativas y las divisivas. Las primeras parten de la existencia de un conglomerado distinto para cada observación, para posteriormente irlos fusionando hasta alcanzar a lo sumo un único grupo. Por su parte, en las técnicas divisivas la situación de partida es un único conglomerado que engloba a todas las observaciones y que progresivamente se va subdividiendo hasta que, a lo sumo, cada observación pertenece a un cluster diferente. Los métodos divisivos requieren demasiados cálculos, lo que motiva que los autores se inclinen habitualmente

por los métodos aglomerativos. De hecho, el programa *PASW Statistics* incluye únicamente métodos de este tipo.

Métodos no-jerárquicos: la característica fundamental que los distingue de los métodos jerárquicos es que solamente llevan a cabo agrupaciones de los individuos en un número concreto de conglomerados, que debe ser fijado de antemano con coherencia.

Una posibilidad para esto es realizar en primer lugar un Análisis cluster mediante procedimientos jerárquicos, que nos permite no sólo determinar el número de grupos o conglomerados *K* más adecuado, sino también la configuración de éstos que se tomarán como punto de partida.

A partir de aquí, la mayoría de los métodos no-jerárquicos establecen *K* entidades (estadísticos que representan, de alguna forma, a los elementos que conforman dichos conglomerados de partida), que servirán para ir definiendo la clasificación final de los elementos entre los distintos conglomerados fijados.

Cada tipo de método no-jerárquico procederá de un modo diferente para asignar los elementos a algún grupo. Así, por ejemplo, el *método de K-medias* implementado en *PASW Statistics* selecciona como entidades iniciales los centroides de cada uno de los *K* conglomerados conformados por el método jerárquico aplicado. Posteriormente, va asignando el resto de elementos al grupo cuyo centroide se encuentre más próximo. Se puede hacer una nueva estimación de los centroides a medida que se van incorporando nuevos elementos, o bien, cuando todos hayan sido asignados a los distintos grupos. Tras esta clasificación inicial, el método o algoritmo utilizado busca reasignaciones de los elementos entre los grupos que den lugar a una mejora en el criterio de agrupación considerado. De no poder realizar ningún cambio que mejore el resultado, el proceso se dará por concluido.

Ejemplo:

La Fundación La Caixa ha llevado a cabo un estudio[6] en el que caracteriza a las 50 provincias españolas mediante los siguientes índices sintéticos, expresados en una escala del 1 al 10:

Índice de renta

Índice de salud

Índice de servicios sanitarios

Índice de nivel educativo y cultural

Índice de oferta educativa, cultural y de ocio

Índice de empleo

[6] Datos del *Anuario Social de España 2004*. Colección Estudios Sociales. Fundación La Caixa.

Índice de condiciones de trabajo

Índice de vivienda y equipamiento del hogar

Índice de accesibilidad económica y seguridad vial

Índice de convivencia y participación social

Índice de seguridad ciudadana

Índice de entorno natural y clima

Se desea agrupar las provincias en una serie de conglomerados de acuerdo con su similitud en los índices señalados.

Solución:

Desarrollaremos este ejemplo utilizando algunas de las opciones que nos brinda *PASW Statistics*. En concreto, utilizaremos el método no jerárquico de *K*-medias.

Previamente, será necesario fijar el número de grupos que se quiere obtener y los valores iniciales que se tomarán como centroides y constituirán las entidades de partida. Para ello, se puede aplicar primero un método jerárquico, con el que se obtendrá el número de grupos adecuados y el valor inicial de los centroides de cada grupo. Después se podrá ya aplicar el método de *K*-medias y se obtendrá la clasificación final en conglomerados. Hay que señalar que, en ocasiones, debido al gran tamaño de algunas de las tablas de resultados obtenidas, no se incluirá la totalidad de éstas, aunque sí se hará mención a ellas.

En primer lugar, hay que tener en cuenta que, además de las variables que recogen los distintos índices sintéticos que se van a considerar en el estudio, es necesario crear una variable nominal en la que se identifique a cada elemento de la muestra, para que una vez formados los conglomerados se tenga claro qué provincias se han agrupado. Con este fin, se crea una variable de tipo "cadena" denominada *Provinci* en la que se han incluido los nombres de las 50 provincias consideradas.

Una vez incluidos todos los datos, a través de *Analizar / Clasificar / Conglomerados jerárquicos* se obtiene un cuadro de diálogo en el que se deben indicar tanto las variables de decisión como la creada para etiquetar los casos (*Figura 26*).

A continuación, se debe indicar la medida de proximidad entre casos con la que se desea trabajar, así como el método elegido para la formación de conglomerados. Esto se hará a través del botón *Método*. Nos hemos decantado por las opciones más habituales: la vinculación inter-grupos[7], como método para clasificar a los elementos en conglomerados, y la distancia euclídea al cuadrado, como medida de proximidad. En

[7] Este método se basa en valores medios. La distancia entre dos conglomerados se calcula tomando la media de las distancias entre cada elemento de uno y otro conglomerado. Los dos grupos que se encuentren a una menor distancia se fusionan para formar un nuevo cluster o conglomerado.

ese mismo cuadro, se nos da la opción de estandarizar las variables, pero en este caso no es necesario porque se trata de índices sintéticos que están todos expresados en una escala del 1 al 10.

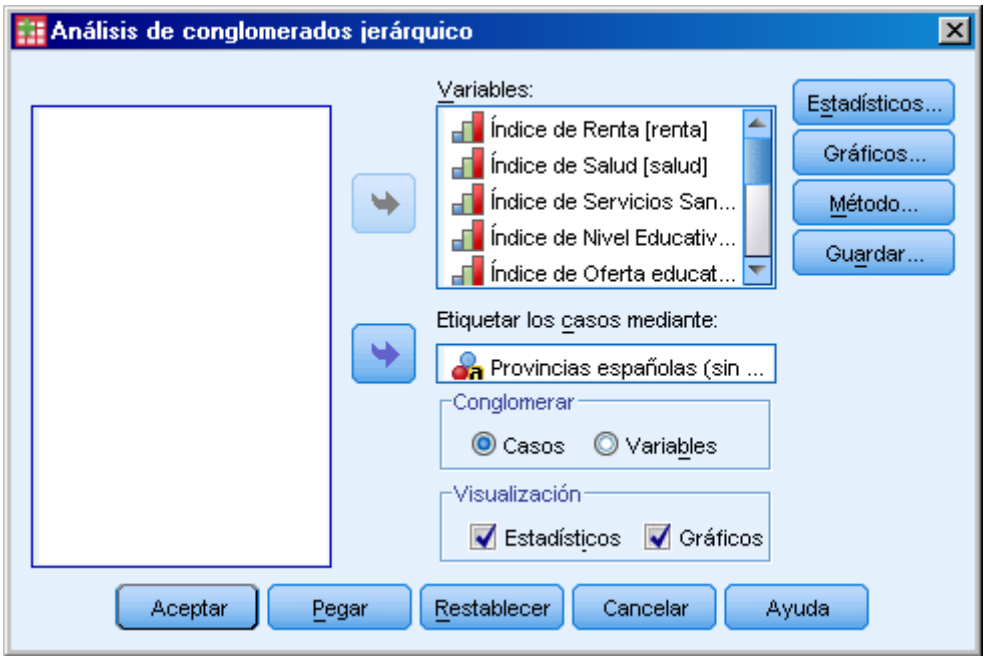


Figura 26

En cuanto a los resultados que queremos que nos devuelva *PASW Statistics*, hemos seleccionado en el botón *Estadísticos*: el *Historial de conglomeración*, la *Matriz de distancias* y un rango de soluciones de entre 3 y 5 conglomerados para el *Conglomerado de pertenencia* (con ello, le pedimos al programa que nos muestre el resultado que se obtendría si tuviésemos 3 conglomerados, 4 ó 5, para a partir de ahí decidir qué nos parece mejor; esto es ya decisión del investigador[8]).

Asimismo, escogemos la opción *Dendograma* en el botón *Gráficos*.

Además, en *Guardar* tenemos la posibilidad de crear nuevas variables en las que se incluirá el conglomerado asignado a cada provincia, para el número de conglomerados que fijemos. Para ello, se deberá proceder como muestra la *Figura 27*. Al haber decidido formar entre 3 y 5 conglomerados, se crearán tres variables con los nombres CLU3_1, CLU4_1 y CLU5_1, donde se guardarán los resultados en cada caso.

Una vez seleccionadas todas las opciones anteriores, se obtienen los resultados que comentaremos a continuación.

[8] Igualmente, también es decisión del investigador decidir dichos números de conglomerados iniciales, de 3 a 5, que desea tener a priori.

 ISBN: 978-84-694-7251-4

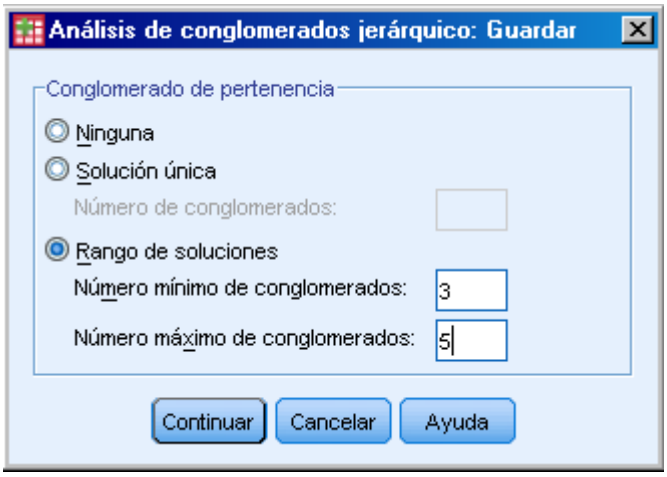


Figura 27

En primer lugar, se muestra un resumen de los casos, distinguiendo entre válidos y perdidos (*Figura 28*).

Seguidamente, aparece la *Matriz de distancias*, que no reproduciremos por su tamaño. Se trata de una tabla con 50 filas y 50 columnas (una por provincia), simétrica, en la que cada elemento indica la distancia (medida como el cuadrado de la distancia euclídea) entre las provincias correspondientes a la fila y la columna en la que se encuentra el elemento. Con esta medida de proximidad calculada a partir de los índices sintéticos, las provincias más cercanas son A Coruña y Pontevedra, cuya distancia es de 12.

Resumen del procesamiento de los casos[a]

Casos					
Válidos		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
50	100,0	0	,0	50	100,0

a. Vinculación promedio (Inter-grupos)

Figura 28

Teniendo en cuenta estas distancias entre provincias, se van formando los conglomerados uniendo las más "cercanas". El *Historial de conglomeración* (*Figura 29*) muestra las distintas etapas del proceso, indicando en cada una de ellas los elementos combinados y la distancia (*coeficientes*) entre ellos. Así, se observa que en la 1ª etapa se han unido las provincias 40 (A Coruña) y 43 (Pontevedra) que eran las más próximas. A continuación, se volvería a calcular la distancia entre todos los conglomerados existentes: el formado por A Coruña y Pontevedra, y los formados individualmente por cada una de las provincias restantes. Las más cercanas luego resultan ser las provincias 26 (Albacete) y 30 (Toledo), que se unen en un conglomerado en la 2ª etapa. Así se va procediendo sucesivamente hasta tener un único conglomerado con todos los elementos.

 ISBN: 978-84-694-7251-4

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	40	43	12,000	0	0	24
2	26	30	16,000	0	0	20
3	17	28	16,000	0	0	46
4	46	47	18,000	0	0	21
5	7	35	18,000	0	0	9
6	3	5	18,000	0	0	10
7	32	33	21,000	0	0	33
8	18	50	23,000	0	0	19
9	7	15	23,000	5	0	25
10	3	27	25,000	6	0	18
11	19	42	27,000	0	0	24
12	11	24	27,000	0	0	31
13	31	49	29,000	0	0	35
14	8	38	29,000	0	0	22
15	20	22	29,000	0	0	27
16	36	37	30,000	0	0	17
17	34	36	34,000	0	16	31
18	2	3	34,000	0	10	29
19	9	18	34,500	0	8	27
20	25	26	36,000	0	2	29
21	46	48	38,000	4	0	35
22	4	8	38,500	0	14	30
23	6	39	39,000	0	0	37
24	19	40	39,500	11	1	28
25	7	14	42,667	9	0	36
26	1	45	43,000	0	0	36
27	9	20	43,833	19	15	34
28	12	19	44,250	0	24	41
29	2	25	46,833	18	20	30
30	2	4	49,619	29	22	37
31	11	34	50,500	12	17	39
32	10	23	51,000	0	0	34
33	13	32	56,500	0	7	39
34	9	10	58,700	27	32	38
35	31	46	63,500	13	21	42
36	1	7	64,000	26	25	43
37	2	6	66,600	30	23	43
38	9	29	66,857	34	0	40
39	11	13	71,600	31	33	44
40	9	21	73,250	38	0	44
41	12	16	74,400	28	0	45
42	31	44	79,200	35	0	49
43	1	2	83,222	36	37	45
44	9	11	89,000	40	39	47
45	1	12	95,167	43	41	46
46	1	17	122,042	45	3	47
47	1	9	124,452	46	44	48
48	1	41	152,814	47	0	49
49	1	31	201,447	48	42	0

Figura 29

 ISBN: 978-84-694-7251-4

La tabla *Conglomerado de pertenencia* indica el conglomerado al que se ha asignado cada provincia, para cada número de conglomerados diferentes solicitado, es decir para los casos de 3 conglomerados, 4 ó 5. Estos valores, que pueden verse en la *Figura 30*, son también los que se guardan en las variables CLU3_1, CLU4_1 y CLU5_1.

El proceso iterativo de formación de conglomerados suele representarse gráficamente en el llamado *dendograma*, que puede verse en la *Figura 31*. Este gráfico muestra cómo las provincias (señaladas en el eje vertical) se agrupan en los sucesivos pasos, así como los niveles de distancia para los que las agrupaciones tienen lugar (en el eje horizontal).

La ventaja del *dendograma* es que nos permite ver de forma rápida a simple vista lo semejantes que son los elementos que constituyen un conglomerado, comparados con los elementos de otros conglomerados, por lo que nos ayuda en la tarea de determinar el número final de conglomerados más adecuado, dependiendo de nuestro objetivo. Esto, generalmente, será una decisión subjetiva del investigador, dependiendo del grado de proximidad que considere aceptable entre los elementos de un grupo. Normalmente se preferirá un número de grupos no demasiado elevado, pero también hay que tener en cuenta que, a medida que disminuye el número de grupos, aumenta la falta de similitud entre los elementos que componen esos grupos.

En el *dendograma* se observa que, en efecto, a medida que nos desplazamos hacia la derecha, disminuye el número de conglomerados, pero también aumenta la distancia entre sus elementos. En el caso de nuestro ejemplo, vamos a optar por los 5 conglomerados que aparecen determinados por la línea roja discontinua que hemos dibujado, aunque otras opciones podrían ser igualmente aceptables.

Una vez que se ha definido el número de grupos, el siguiente paso para poder aplicar el método no-jerárquico de *K*-medias consiste en calcular los centroides de los 5 grupos definidos. Estos centroides se tomarán como valores iniciales del proceso de iteración en el método de *K*-medias. Recordemos que el centroide de un grupo o conglomerado será un vector cuyas componentes son los valores medios de cada una de las variables independientes, para las provincias pertenecientes a ese grupo.

Para calcular estas medias, se pulsa *Analizar / Comparar medias / Medias* y, en el cuadro de diálogo resultante, se introducen todas las variables de las que se quiere calcular la media (en *Lista de dependientes*) y se indican asimismo los conglomerados que nos interesan (en *Lista de independientes*): los 5 con los que hemos decidido quedarnos. Estos conglomerados están incluidos en una variable cuya etiqueta es *Average Linkage (Between Groups)*. Pero debemos tener cuidado, porque hay tres variables con esa misma etiqueta, las correspondientes a los casos de 3, 4 y 5 conglomerados que queríamos analizar. Moviéndonos sobre ellas, aparece el nombre de la variable correspondiente. Nos interesa CLU5_1, que es la que guardaba los resultados para 5 conglomerados.

El informe resultante es el que se muestra en la *Figura 32*.

Conglomerado de pertenencia

Caso	5 conglomerados	4 conglomerados	3 conglomerados
1:Almería	1	1	1
2: Cádiz	1	1	1
3: Córdoba	1	1	1
4: Granada	1	1	1
5: Huelva	1	1	1
6: Jaén	1	1	1
7: Málaga	1	1	1
8: Sevilla	1	1	1
9: Huesca	2	2	1
10: Teruel	2	2	1
11: Zaragoz	2	2	1
12: Asturia	1	1	1
13: Balears	2	2	1
14: Palmas	1	1	1
15: Santa C	1	1	1
16: Cantabr	1	1	1
17: Ávila	3	1	1
18: Burgos	2	2	1
19: León	1	1	1
20: Palenci	2	2	1
21: Salaman	2	2	1
22: Segovia	2	2	1
23: Soria	2	2	1
24: Vallado	2	2	1
25: Zamora	1	1	1
26: Albacet	1	1	1
27: Ciudad	1	1	1
28: Cuenca	3	1	1
29: Guadala	2	2	1
30: Toledo	1	1	1
31: Barcelo	4	3	2
32: Girona	2	2	1
33: Lleida	2	2	1
34: Tarrago	2	2	1
35: Alicant	1	1	1
36: Castell	2	2	1
37: Valenci	2	2	1
38: Badajoz	1	1	1
39: Cáceres	1	1	1
40: Coruña	1	1	1
41: Lugo	5	4	3
42: Ourense	1	1	1
43: Ponteve	1	1	1
44: Madrid	4	3	2
45: Murcia	1	1	1
46: Navarra	4	3	2
47: Álava	4	3	2
48: Guipúzc	4	3	2
49: Vizcaya	4	3	2
50: Rioja	2	2	1

Figura 30

 ISBN: 978-84-694-7251-4

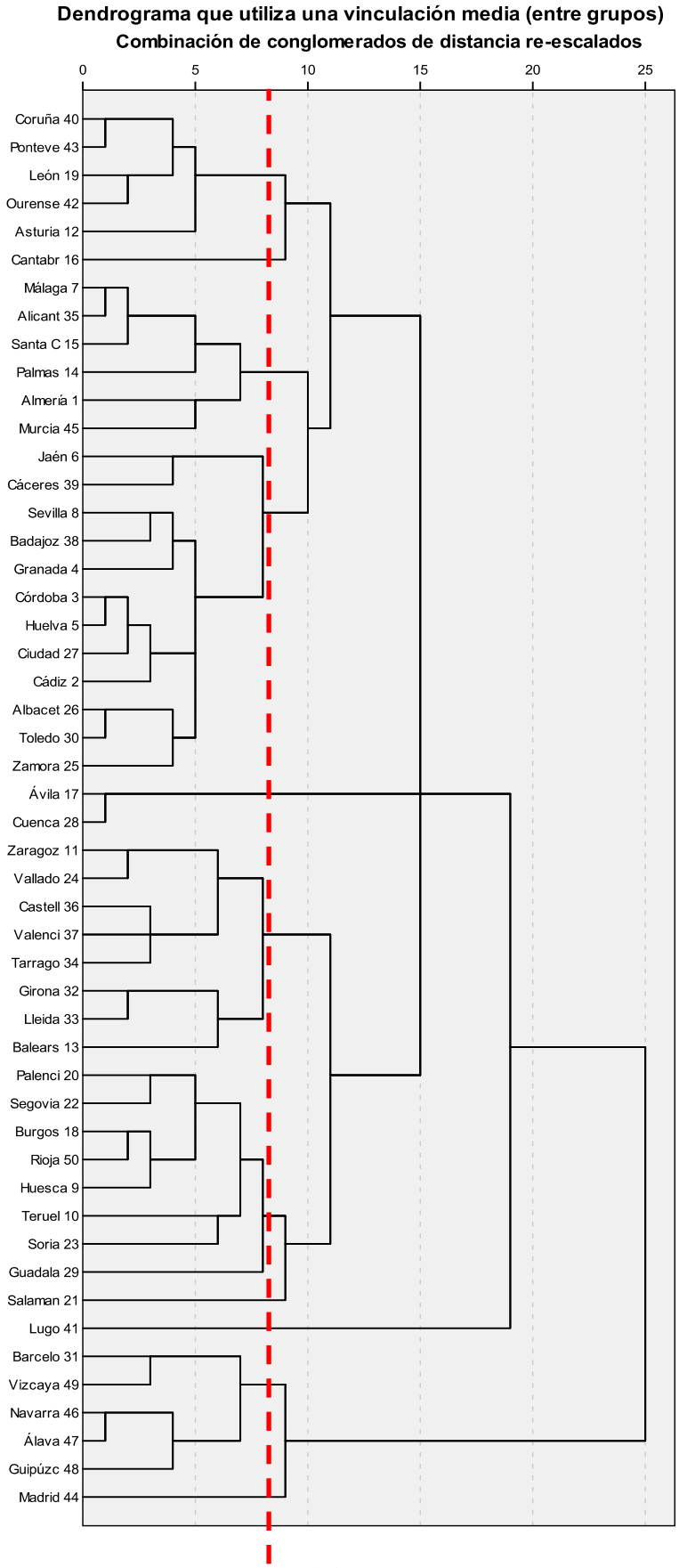


Figura 31

 ISBN: 978-84-694-7251-4

Informe

Media

Average Linkage (Between Groups)	Índice de Renta	Índice de Salud	Índice de Servicios Sanitarios	Índice de Nivel Educativo y Cultural	Índice de Oferta educativa, cultural y de Ocio	Índice de empleo	Índice de condiciones de trabajo	Índice de vivienda y equipamiento del hogar	Índice de accesibilidad económica y seguridad vial	Índice de convivencia y participación social	Índice de seguridad ciudadana	Índice de entorno natural y clima
1	3,29	5,29	4,21	4,46	3,63	4,00	5,21	4,79	6,04	5,50	6,29	5,54
2	6,94	5,65	6,06	6,24	7,71	7,12	5,06	5,94	4,53	6,29	6,00	4,71
3	4,50	8,00	1,50	1,50	7,50	3,50	5,50	2,00	1,50	9,00	7,00	5,50
4	9,17	6,17	9,50	9,67	5,67	8,17	5,83	9,17	8,00	2,83	2,33	6,50
5	3,00	1,00	5,00	1,00	3,00	6,00	10,00	1,00	3,00	6,00	9,00	7,00
Total	5,28	5,54	5,38	5,50	5,40	5,58	5,34	5,52	5,52	5,60	5,80	5,40

Figura 32

 ISBN: 978-84-694-7251-4

Las 12 medias que aparecen en cada fila de esta *Figura 32* son las componentes del centroide de cada grupo. Estos valores se deben copiar en un archivo de *PASW Statistics*, del que el programa los importará luego para tomarlos como valores iniciales del proceso de iteración del método no-jerárquico de *K*-medias. Dicho archivo lo hemos nombrado en este ejemplo *centroides.sav*. Se deben cumplir dos requisitos: la variable que identifica a los conglomerados debe denominarse *cluster_* y el resto de variables debe conservar el nombre del archivo inicial.

Una vez creado el archivo que contiene a los centroides, estamos en condiciones de ejecutar el análisis de conglomerados de *K*-medias. Para ello, en el archivo inicial pulsamos *Analizar / Clasificar / Conglomerado de K medias*, resultando el cuadro de la *Figura 33*. Deberemos introducir tanto las variables de decisión como la que usamos de etiqueta de las provincias. Indicaremos que el número de conglomerados es 5 y la ruta en la que se encuentra el archivo donde hemos guardado los centroides. Este se hace en *Centros de los conglomerados* / *Leer iniciales* / *Archivo de datos externo* / *Archivo*.

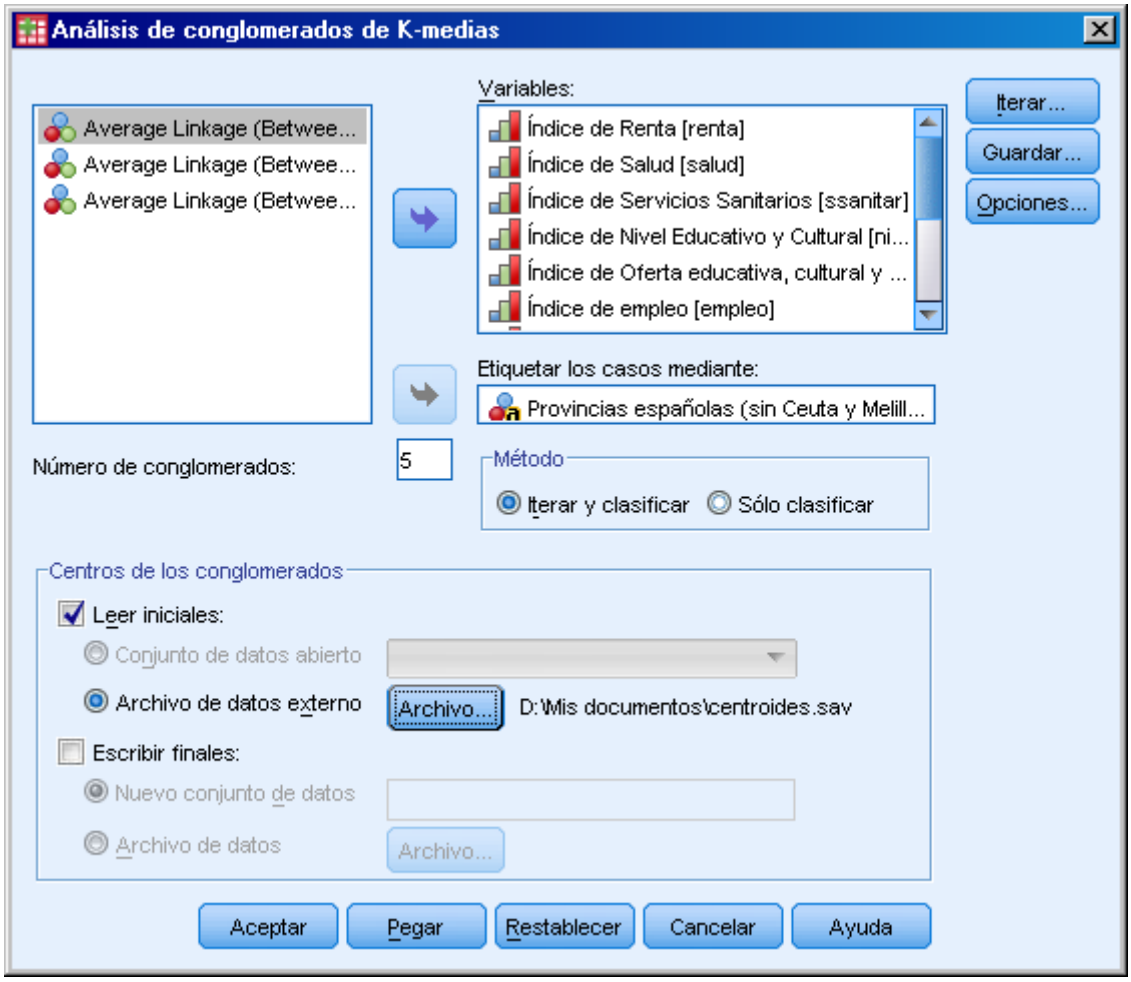


Figura 33

En *Iterar...* se pueden modificar opciones de cálculo, como el número máximo de iteraciones o el criterio de convergencia para detener el proceso iterativo en que se basa

 ISBN: 978-84-694-7251-4

el método de *K*-medias o la posibilidad de actualizar los centroides cada vez que se asigne una provincia a un conglomerado. Dejaremos las opciones que aparecen por defecto (10 iteraciones y que no se actualicen las medias (centroides)).

Guardar… permite crear dos variables que *PASW Statistics* denomina por defecto QCL_1 y QCL_2, en las que se almacenarán, respectivamente, el conglomerado al que estará asignada cada provincia una vez terminado el proceso de agrupación y la distancia de cada provincia al centroide de su conglomerado. Marcaremos ambas opciones.

También marcaremos todos los estadísticos optativos en el botón *Opciones*: *Centros de conglomerados iniciales*, *Tabla de ANOVA* e *Información del conglomerado para cada caso*.

A partir de los valores iniciales de los centroides, el proceso va asignando las provincias a aquel centroide que esté a una menor distancia. Cuando todas las provincias se hayan asignado, se recalculan los centroides de los conglomerados (dado que no marcamos la opción *Usar medias actualizadas*). Se sigue iterando hasta que ninguna reasignación de una provincia a otro grupo permita reducir la distancia entre las provincias dentro de cada conglomerado ni aumentar la distancia entre conglomerados. Conviene recordar que, a diferencia de los métodos jerárquicos, este procedimiento permite que un elemento asignado a un grupo en una iteración previa, pueda ser asignado a otro grupo en una iteración posterior.

El resultado del estudio, en cuanto al conglomerado o cluster al que finalmente pertenece cada provincia, así como la distancia entre ésta y el centroide de su cluster, puede verse en la *Figura 34*. Estos datos están también incluidos en las variables QCL_1 y QCL_2 que se crearon al efecto. Como ejemplo, cabe señalar que Barcelona, Madrid, Navarra, Álava, Guipúzcoa y Vizcaya se han agrupado en un único cluster.

La asignación resultante puede compararse con la obtenida mediante el procedimiento jerárquico; si se observa una elevada cantidad de reasignaciones, será indicativo de que el método jerárquico empleado no era el adecuado para los datos analizados.

Las *Figuras 35, 36* y *37* nos aportan datos complementarios del análisis.

Así, la *Figura 35* nos indica el tamaño de cada conglomerado.

La *Figura 36* muestra los centros de los conglomerados finales, es decir, los valores medios de cada variable de decisión en cada uno de los grupos de provincias configurados.

Y por su parte, la *Figura 37* señala la distancia entre los centros de los conglomerados, lo que puede servir para determinar la heterogeneidad entre grupos, aunque no para estudiar la homogeneidad interna de cada conglomerado. Recordemos que el objetivo del Análisis cluster es establecer grupos lo más homogéneos posible internamente pero los más heterogéneos posible entre sí.

 ISBN: 978-84-694-7251-4

Pertenencia a los conglomerados

Número de caso	Provincias españolas (sin Ceuta y Melilla)	Conglomerado	Distancia
1	Almería	1	5,682
2	Cádiz	1	5,932
3	Córdoba	1	4,259
4	Granada	1	5,477
5	Huelva	1	4,100
6	Jaén	1	6,133
7	Málaga	1	4,786
8	Sevilla	1	4,429
9	Huesca	2	4,538
10	Teruel	2	7,655
11	Zaragoz	2	5,574
12	Asturia	2	7,858
13	Balears	2	7,613
14	Palmas	1	7,895
15	Santa C	1	5,251
16	Cantabr	2	6,208
17	Ávila	3	2,000
18	Burgos	2	3,563
19	León	1	4,850
20	Palenci	2	5,904
21	Salaman	2	6,807
22	Segovia	2	5,425
23	Soria	2	7,364
24	Vallado	2	5,357
25	Zamora	1	5,568
26	Albacet	1	5,711
27	Ciudad	1	5,169
28	Cuenca	3	2,000
29	Guadala	2	6,876
30	Toledo	1	6,351
31	Barcelo	4	4,690
32	Girona	2	6,375
33	Lleida	2	5,837
34	Tarrago	2	5,012
35	Alicant	1	4,943
36	Castell	2	5,233
37	Valenci	2	6,658
38	Badajoz	1	4,816
39	Cáceres	1	6,665
40	Coruña	1	6,782
41	Lugo	5	3,122
42	Ourense	5	3,122
43	Ponteve	1	6,377
44	Madrid	4	6,403
45	Murcia	1	6,973
46	Navarra	4	3,651
47	Álava	4	4,655
48	Guipúzc	4	5,354
49	Vizcaya	4	4,830
50	Rioja	2	3,342

Figura 34

 ISBN: 978-84-694-7251-4

Número de casos en cada conglomerado

Conglomerado	1	21,000
	2	19,000
	3	2,000
	4	6,000
	5	2,000
Válidos		50,000
Perdidos		,000

Figura 35

Centros de los conglomerados finales

	Conglomerado				
	1	2	3	4	5
Índice de Renta	3	7	5	9	4
Índice de Salud	6	6	8	6	2
Índice de Servicios Sanitarios	4	6	2	10	5
Índice de Nivel Educativo y Cultural	4	6	2	10	2
Índice de Oferta educativa, cultural y de Ocio	4	7	8	6	3
Índice de empleo	4	7	4	8	6
Índice de condiciones de trabajo	5	5	6	6	8
Índice de vivienda y equipamiento del hogar	5	6	2	9	1
Índice de accesibilidad económica y seguridad vial	6	5	2	8	4
Índice de convivencia y participación social	6	6	9	3	6
Índice de seguridad ciudadana	6	6	7	2	9
Índice de entorno natural y clima	5	5	6	7	7

Figura 36

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3	4	5
1		7,389	8,950	12,984	7,828
2	7,389		10,178	9,119	10,574
3	8,950	10,178		18,317	10,665
4	12,984	9,119	18,317		16,681
5	7,828	10,574	10,665	16,681	

Figura 37

 ISBN: 978-84-694-7251-4

Una forma de analizar si la variabilidad entre conglomerados es mayor que dentro de los conglomerados es a través de la tabla ANOVA que se proporciona en la *Figura 38*.

ANOVA

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Índice de Renta	61,534	4	1,821	45	33,792	,000
Índice de Salud	11,778	4	3,362	45	3,503	,014
Índice de Servicios Sanitarios	50,966	4	2,531	45	20,133	,000
Índice de Nivel Educativo y Cultural	52,109	4	2,268	45	22,975	,000
Índice de Oferta educativa, cultural y de Ocio	45,423	4	2,851	45	15,931	,000
Índice de empleo	36,955	4	2,719	45	13,591	,000
Índice de condiciones de trabajo	3,661	4	4,013	45	,912	,465
Índice de vivienda y equipamiento del hogar	39,012	4	1,698	45	22,969	,000
Índice de accesibilidad económica y seguridad vial	25,858	4	3,179	45	8,135	,000
Índice de convivencia y participación social	18,481	4	2,713	45	6,813	,000
Índice de seguridad ciudadana	23,514	4	2,621	45	8,972	,000
Índice de entorno natural y clima	2,934	4	5,028	45	,584	,676

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

Figura 38

Así, para cada variable de decisión, se contrasta la igualdad de medias entre conglomerados, a través de un estadístico *F* que es el cociente de las medias cuadráticas inter-grupos e intra-grupos. Como es sabido, valores elevados del estadístico *F* reflejarán que la variabilidad entre los grupos es mucho mayor que la variabilidad dentro de cada grupo, por lo que preferiremos aquellas soluciones que lleven a mayores valores de *F*. De esta forma, los conglomerados o clusters elaborados son homogéneos especialmente en el factor *Índice de Renta* (para el que la *F* alcanza el valor más elevado). En la última columna de la tabla aparecen los *p*-valores asociados a cada contraste, pudiéndose observar que en todos los factores salvo en dos (*Índice de condiciones de trabajo* e *Índice de entorno natural y clima*) la variabilidad entre grupos supera a la variabilidad intra-grupos (en estos dos casos se acepta la hipótesis nula de igualdad de medias en los 5 conglomerados). No obstante, hay que ser prudente a la hora de extraer conclusiones en este sentido, puesto que como el mismo programa señala en una nota al pie de la tabla, este test debe usarse solamente con una finalidad

descriptiva. En cualquier caso, se puede emplear para valorar la relevancia de las variables empleadas y comparar las diferentes agrupaciones.

Para terminar, una cuestión adicional que se podría plantear es determinar cuáles son los perfiles comunes de las provincias que constituyen cada uno de los 5 conglomerados construidos. Esto es lo que se denomina "perfilar" los segmentos, y puede hacerse mediante la técnica multivariante del Análisis discriminante. Lo lógico sería emplear para ello nuevas variables, distintas a las empleadas en el anterior proceso de agrupación. Esta manera de proceder resulta muy habitual en este tipo de estudios.

 ISBN: 978-84-694-7251-4

TEMA 2

El modelo clásico de regresión lineal: especificación y estimación

La Econometría no es Estadística económica. Tampoco es lo que llamamos Teoría Económica (...). La Econometría tampoco debe ser considerada como sinónimo de aplicación de las Matemáticas a la Economía. La experiencia ha demostrado que cada uno de estos tres puntos de vista, el de la Estadística, la Teoría Económica y las Matemáticas, es necesario, pero por sí mismo no son condición suficiente para una comprensión real de las relaciones cuantitativas en la vida económica moderna. Es la unión de los tres aspectos lo que constituye una herramienta de análisis potente. Es esta unión lo que constituye la Econometría.[1]

Comienza con este tema el análisis del modelo econométrico, el eje central en torno al que se desarrollan los métodos econométricos.

En concreto, en el presente tema empezaremos viendo qué se entiende por un modelo econométrico. Después, iremos desarrollando las principales etapas que lo configuran: especificación, estimación, inferencia y predicción. Las dos primeras se verán en este mismo Tema, en tanto que las dos últimas se analizarán en el Tema siguiente. Todo este estudio se hará basándonos en el modelo clásico de regresión, caracterizado por cumplir una serie de supuestos y disfrutar con ello de un conjunto bien definido de propiedades.

Una vez conocido este modelo "perfecto", proseguiremos con el análisis del modelo cuando no se cumplen todos los requisitos o propiedades descritas previamente, viendo cuáles son las consecuencias principales de ello y tomando las medidas más oportunas para afrontarlas. Éste es el objetivo que nos plantearemos en el Tema 4.

Finalmente, en el Tema 6 abordaremos el estudio de un tipo de modelo muy importante en el mundo de la Empresa: los modelos de elección discreta. Estos modelos se caracterizan por ser su variable dependiente de tipo discreto y una de sus principales utilidades es su consideración en procesos de toma de decisiones.

2.1. Definición del modelo econométrico.-

Un modelo es una representación simplificada de la realidad, que debe ser plausible y manejable. Teniendo presente cuál es el objetivo de la Econometría, un modelo econométrico es un modelo que incluye las especificaciones necesarias para tratar de reflejar las relaciones empíricas del ámbito de la Economía.

[1] R. Frisch (*Econometrica*, vol. 1, nº 1, 1933).

 ISBN: 978-84-694-7251-4

La realidad económica es absolutamente inalcanzable en términos de modelización determinística. En todo modelo económico es preciso considerar un componente aleatorio que permita incluir:

a) todas aquellas variables relevantes que pudieran no estar especificadas inicialmente en el modelo (ya sea por ignorancia o consideración de teorías incompletas, por ausencia de datos disponibles o, simplemente, por sencillez en la definición del modelo);
b) variables de tipo cualitativo que por su naturaleza no sean cuantificables;
c) la aleatoriedad del comportamiento humano; y
d) posibles errores de medida en las variables utilizadas.

Así, el punto de partida de un modelo econométrico se fundamenta en dos componentes:

- el determinístico (formado por aquellas variables cuyas relaciones explícitamente deseamos considerar); y
- el aleatorio.

Analíticamente, pues, al tratar de estudiar el comportamiento de una cierta variable *Y*, tendríamos que el mismo estaría definido por una serie de variables conocidas $(X_1, X_2, \ldots, X_k)$ consideradas de forma explícita (y que se relacionan con *Y* según una determinada forma o función matemática) y por un elemento aleatorio, que denominamos perturbación, que comprendería "todo lo demás":

$$\boxed{Y = f(X_1, X_2, \cdots, X_k) + u}.$$

La definición completa de un modelo econométrico comprende una serie de etapas o fases, que se muestran en la *Figura 1*:

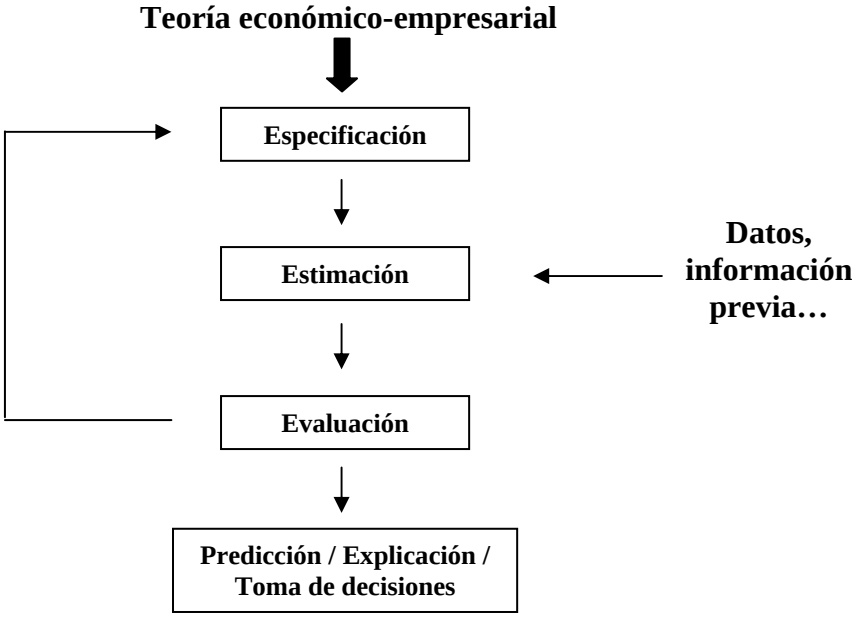


Figura 1

 ISBN: 978-84-694-7251-4

2.2. El modelo econométrico de regresión lineal: especificaciones simple y múltiple. Formulación matricial del modelo. Supuestos del modelo clásico.-

El modelo clásico de regresión lineal: especificaciones simple y múltiple

Como se acaba de ver, en términos generales, el modelo econométrico plantea que la variable dependiente (o explicada) está relacionada funcionalmente con una o varias variables independientes (o explicativas), pero asumiendo que la relación entre ambas partes no es exacta. Y para tener en cuenta la relación inexacta entre todas ellas, se añade una variable aleatoria estocástica o perturbación, que representa de alguna forma a todas las variables que no se tienen en cuenta de manera especifica en el modelo, los errores de medición y otros factores aleatorios.

La primera de las fases de definición de un modelo econométrico consiste en su especificación, esto es, en plantear qué variable económica queremos estudiar y qué variable o variables pueden determinar su comportamiento, eligiendo además la forma funcional matemática que las relaciona. A partir de ahí, seguidamente se procede a la siguiente fase, la estimación o concreción de dicha relación funcional.

Para dar contenido empírico al modelo, es decir, para estimar sus parámetros o coeficientes, utilizamos como herramienta el análisis de regresión. El análisis de regresión *trata del estudio de la dependencia o explicación del comportamiento de una variable respecto a una o varias variables independientes o explicativas, con el objetivo de estimar y/o predecir el valor esperado o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas*.

En este tema, vamos a empezar estudiando un caso particular de forma funcional para nuestro modelo: la función lineal. En concreto, la linealidad la interpretaremos tanto referida a las variables explicativas *X* como a los parámetros o coeficientes de regresión β.

La razón primera de esta elección radica en su sencillez matemática, pero se ve reforzada además por el hecho de que, empíricamente, el comportamiento de buena parte de la realidad económico-empresarial se puede modelizar de un modo razonablemente aceptable mediante relaciones lineales.

Así pues, nos centraremos en el modelo econométrico de regresión lineal:

$$\boxed{Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,N},$$

donde el subíndice *j* se refiere a las variables explicativas del modelo (hasta *k*) y el subíndice *i* a las distintas observaciones poblacionales de todas las variables del modelo (hasta *N*).

A la hora de especificar un modelo, un primer aspecto que debemos tener claro es si el "término independiente" está presente o no en el modelo. El término independiente no es más que la ordenada en el origen matemática del modelo; es decir, la variable a la que acompaña el parámetro β_1 en la ecuación anterior. Obsérvese que dicha variable, que podemos llamar X_1, se caracteriza porque adopta el valor 1 para todas las observaciones; esto es: $X_{1i} = 1, \ \forall i = 1, 2, ..., N$.

Lo más habitual es que los modelos se especifiquen de forma que incluyan término independiente, pero no tiene por qué ser así.

La distinción entre modelo de regresión simple y modelo de regresión múltiple hace referencia al número de variables explicativas que presenta un modelo. De este modo, cuando el comportamiento de una variable es explicado únicamente por otra variable, se habla de regresión simple. En cambio, cuando dicho comportamiento viene explicado en función de 2 ó más variables deterministas, se considera que la regresión es múltiple.

Según esto, por tanto, cuando se habla de un modelo de regresión lineal simple, su expresión (considerando la ordenada en el origen) es:

$$\boxed{Y_i = \beta_1 + \beta_2 X_{2i} + u_i \qquad \forall i = 1, 2, ..., N},$$

que, como podemos apreciar, no es sino un caso particular del modelo de regresión lineal en el que $k = 2$. (Por mayor simplicidad incluso, X_{2i} se podría escribir como X_i, ya que por ser la única variable X "visible", se podría prescindir de su subíndice "2").

En la mayoría de las ocasiones, nosotros nos referiremos al modelo de regresión múltiple, que es el caso general. Sin embargo, en otras, por simplicidad en las explicaciones (y sobre todo a nivel gráfico), haremos referencia al modelo de regresión simple.

Precisamente, la siguiente argumentación la haremos basándonos en un modelo de regresión lineal simple.

Pensemos que queremos estudiar una variable económica (variable dependiente) cuyo comportamiento creemos que depende, según una relación lineal, del de otra variable que actúa como independiente o explicativa de la primera. Si dispusiésemos de los valores de las N observaciones que conforman la población de tales variables, teóricamente podríamos representar una nube de puntos en la que podríamos "cruzar", en unos ejes cartesianos, los valores observados para ambas variables.

A cada valor de la variable explicativa X le podrían corresponder varios valores de la variable dependiente o explicada Y (si nos fijamos en la *Figura 2*, para un valor particular de X, por ejemplo, X_0, le podrían corresponder distintos valores de Y: Y_0, Y'_0, Y''_0). Si quisiéramos asociar a cada valor de la variable explicativa un único valor de la variable explicada, nos surgiría entonces la pregunta de cuál tomar. En este

 ISBN: 978-84-694-7251-4

punto, parece que lo lógico sería elegir un valor representativo de todos los posibles valores de *Y* que aparecen ligados a cada uno de los valores de *X*; este valor elegido sería el valor esperado o esperanza matemática de la variable *Y*, dado el valor de *X*: $E(Y \mid X_0)$.

Por tanto, el par de valores que asociaríamos sería: $(X_i, E(Y \mid X_i))$. De esta manera lo que modelizaríamos no sería el comportamiento de la variable dependiente, sino su comportamiento promedio o esperado; es decir, nuestro objetivo va a consistir en estimar el valor promedio de la variable dependiente, conocidos los valores de la variable explicativa: $E(Y \mid X_i) = f(X_i)$.

Si tomamos como ejemplo un modelo de regresión lineal simple (con ordenada en el origen), tendríamos entonces lo que se conoce como Recta de Regresión Poblacional (RRP):

$$\boxed{E(Y \mid X_i) = \beta_1 + \beta_2 X_i}.$$

Gráficamente (*Figura 2*), nuestro objeto de estudio son, por tanto, del conjunto de datos poblacionales, los puntos que conforman la RRP. Sobre esta recta se representan los valores medios de la variable *Y* para cada valor de la variable dependiente *X*.

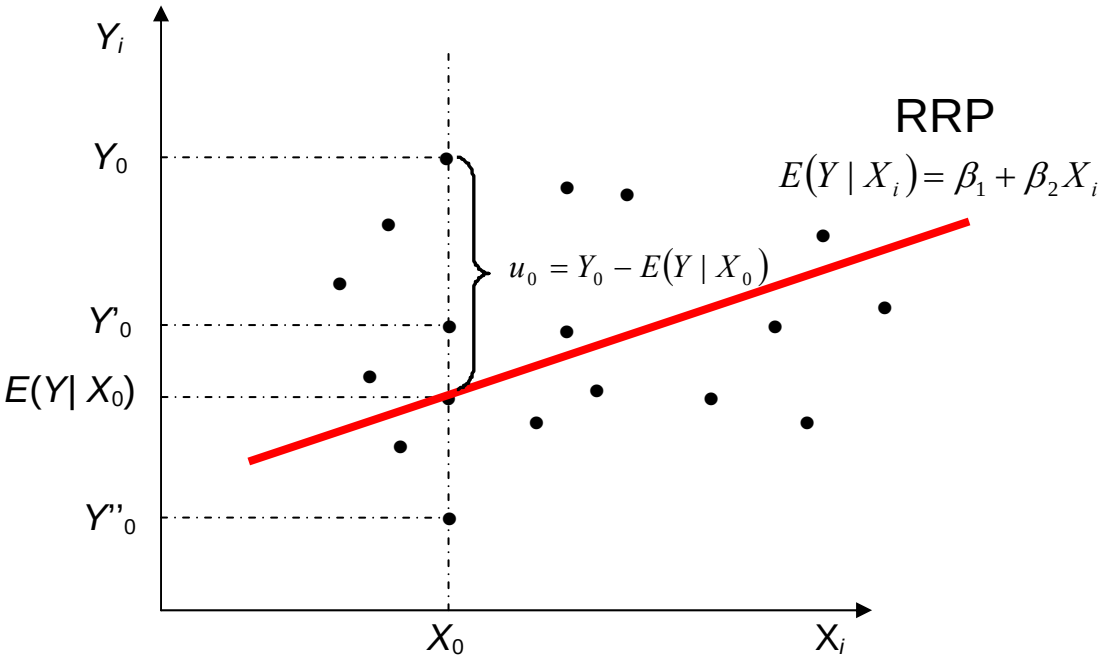


Figura 2

La RRP constituye un instrumento con el que intentamos estudiar la nube de puntos (o Modelo de Regresión Poblacional). En particular, muestra que, dado un determinado valor de *X*, los valores de la variable dependiente están agrupados alrededor de su esperanza condicional. La diferencia entre un valor concreto Y_0 de la variable explicada

"asociado" a un valor concreto X_0 de la variable explicativa y el valor promedio o esperado para dicho X_0, $E(Y \mid X_0)$, se correspondería conceptualmente con la parte del comportamiento no determinada por X_0, es decir, con la perturbación aleatoria: $Y_0 - E(Y \mid X_0) = u_0$; o lo que es lo mismo: $Y_0 = E(Y \mid X_0) + u_0$.

Generalizando para cualquier observación *i*, tendríamos por tanto que:

$$\boxed{u_i = Y_i - E(Y \mid X_i)}, \text{ o bien: } \boxed{Y_i = E(Y \mid X_i) + u_i}.$$

Obsérvese que, dado que el término de perturbación *u* es una variable aleatoria, la variable dependiente *Y* también lo va a ser, puesto que depende de la anterior. Su función de probabilidad va a ser la misma que la de la perturbación aleatoria. El siguiente paso sería determinar cuál es esta distribución de probabilidad.

El hecho de que la perturbación aleatoria represente la suma o influencia combinada (sobre la variable dependiente) de un elevado número desconocido de variables independientes que no se han incluido de forma explícita en el modelo de regresión, las cuales se suponen entre sí independientes e idénticamente distribuidas, así como con esperanza y varianza finitas, conduce a considerar, por el Teorema Central del Límite, que se comporte según una distribución normal de probabilidad. Así pues:

$$\boxed{u_i \to N}.$$

El supuesto de normalidad del término de perturbación aleatoria constituye un pilar fundamental sobre el que se asienta toda la teoría econométrica que vamos a desarrollar, en especial los aspectos inferenciales de la misma.

Prosiguiendo con el análisis de *u*, si nos centramos en el concepto que representa, no hay motivos para pensar que presente una tendencia o error sistemático hacia desviaciones siempre del mismo signo (positivas o negativas). Si tuviera tal desviación sistemática, dejaría de ser una componente aleatoria. Evidenciaría que el modelo se habría especificado incorrectamente. Por tanto, se supone que unas desviaciones serán positivas y otras negativas, de forma que estas diferencias poblacionales "a la larga" se verán compensadas; es decir, asumimos que:

$$\boxed{E(u_i \mid X_i) = 0}.$$

Este supuesto[2] se ve además reforzado por el hecho de que parece razonable pensar que lo deseable es que las desviaciones más frecuentes entre los valores observados de *Y* y sus promedios (que no es más que la definición de *u*) sean pequeñas. Si estamos ante

[2] Si bien el modo correcto de escribir esta expresión es: $E(u_i \mid X_i) = 0$, en la notación econométrica resulta habitual obviar que cuando estudiamos las variables estocásticas, éstas están condicionadas a los valores de las variables explicativas; de este modo, se suele escribir simplemente: $E(u_i) = 0$.

 ISBN: 978-84-694-7251-4

una distribución normal, recuérdese que la moda y la media (además de la mediana) coinciden en esta distribución, por lo que esta media tendería a ser cero.

Volviendo a la RRP, con ella lo que intentamos es analizar la nube de puntos poblacional, que lo habitual es que no sea conocida pues no se suele disponer de la totalidad de los valores poblacionales. En la práctica, sólo se tiene al alcance una muestra de valores de Y que se corresponden con valores fijos de X. En esta situación, nuestro objetivo final consiste en ajustar o estimar la RRP, esto es, obtener una estimación numérica de los valores de los parámetros β desconocidos, usando para ello la información proporcionada por observaciones muestrales de las variables del modelo. Se obtiene de este modo la denominada Recta de Regresión Muestral, (RRM), que no es sino una estimación de la RRP, una aproximación de la verdadera RRP, puesto que ésta no se puede estimar de manera precisa debido a las fluctuaciones muestrales.

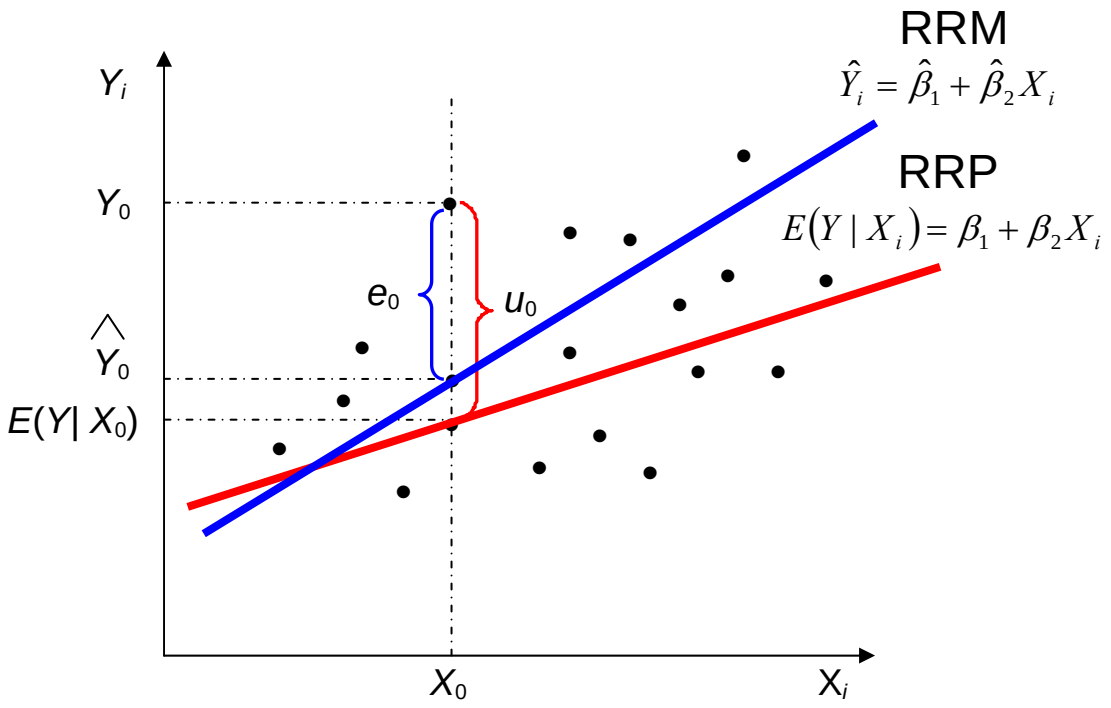


Figura 3

Es preciso reseñar que la RRM nos proporciona una estimación de $E(Y / X_i)$, no de Y_i; es decir, con la RRM lo que se obtiene es una estimación, a través de la inferencia estadística, del valor medio de Y y no de las observaciones reales[3] de Y.

[3] Obsérvese, sin embargo, cómo la notación parece estar mal escrita, pues se indica $\hat{Y}_i$, en lugar de $\widehat{E(Y / X_i)}$, que sería lo correcto. No obstante, ésta es la notación habitual en el ámbito econométrico y así seguiremos utilizándola, teniendo en todo caso presente cuál es el significado correcto de lo que estudiamos.

Las estimaciones obtenidas también dan lugar a desviaciones respecto a los valores reales de *Y*, registrándose el denominado error o residuo muestral, que se denota por e_i:

$$\boxed{e_i = Y_i - \hat{Y}_i}.$$

Conceptualmente, este error muestral es similar a la perturbación aleatoria u_i, pero no debe confundirse con ésta. Si se observa, el residuo es una estimación muestral de la perturbación aleatoria, que es poblacional: $e_i = \hat{u}_i$.

Para llevar a cabo la explicación de conceptos que hemos desarrollado hasta este punto, hemos recurrido en buena parte de nuestra exposición al caso del modelo de regresión simple (fundamentalmente en el aspecto gráfico), pero como bien podrá apreciarse, todo ello se puede extender fácilmente al caso general.[4] En lugar de Recta de Regresión Poblacional (RRP) y de Recta de Regresión Muestral (RRM), podríamos hablar de forma generalizada de Función de Regresión Poblacional (FRP) y de Función de Regresión Muestral (FRM).

Formulación matricial del modelo

Como ya sabemos, la formulación del modelo general de regresión lineal con *k* variables explicativas puede escribirse como:

$$\boxed{Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n}.$$

Si se escribe la ecuación que se obtiene para cada una de las *n* observaciones, tendremos:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + ... + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + ... + \beta_k X_{k2} + u_2 \\ &\cdots\cdots\cdots\cdots\cdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + u_n \end{aligned}$$

Si nos fijamos, podemos expresar el conjunto de todas estas ecuaciones de forma matricial, de modo que:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \text{ o abreviadamente: } \boxed{Y = X\beta + u},$$

[4] En el caso del modelo lineal simple, geométricamente hemos trabajado con una recta. Si el modelo tuviese dos variables explicativas, estaríamos ante un plano de regresión; y si tuviésemos más de dos, hablaríamos, en general, de un hiperplano de regresión.

donde la matriz Y es una matriz columna, de orden n x 1, cuyos elementos son los n valores de la variable dependiente Y; X es una matriz de orden n x k, estando conformadas sus columnas por los n valores de cada una de las k variables independientes del modelo; β es una matriz de orden k x 1, siendo sus elementos cada uno de los k parámetros que acompañan a cada una de las variables independientes del modelo; y u es una matriz de orden n x 1, referida las perturbaciones de cada una de las n observaciones del modelo.

Supuestos del modelo clásico

En el análisis de regresión, nuestro objetivo no va a ser sólo estimar los parámetros del modelo considerado, sino también hacer un ejercicio de inferencia sobre los verdaderos valores de tales parámetros o coeficientes. Para ello deben hacerse ciertos supuestos sobre los distintos componentes que conforman el modelo (algunos de los cuales ya se han referido). El cumplimiento de estos supuestos da lugar al llamado *modelo clásico* de regresión lineal.

Vamos a distinguir 3 tipos de supuestos: no estocásticos, estocásticos y los referidos a la distribución de probabilidad.

- ***Supuestos no estocásticos***

1. El modelo de regresión es lineal en los parámetros y en las variables explicativas. (Ya comentado).
2. Los valores de las variables explicativas X son fijos en muestreo repetido; es decir, las variables X se suponen no estocásticas. Así, suponiendo fijo el valor de X, se pueden observar los distintos niveles de la variable Y a la hora de obtener la muestra. Por ello, el análisis es de regresión condicional, esto es, condicionado a los valores dados de las variables X.
3. El número de observaciones n debe ser mayor o igual que el número de parámetros o coeficientes de regresión k a estimar (es decir, mayor o igual que el número de variables explicativas): $n \geq k$.
4. No existe multicolinealidad perfecta entre las variables explicativas del modelo, es decir, no hay relaciones lineales exactas entre las mismas. Matemáticamente, ello significa que teniendo en cuenta que la matriz X es de orden n x k, su rango debe ser k, con $n \geq k$.
5. El modelo de regresión está correctamente especificado. Esto supone asumir que todas las variables relevantes están incluidas en el modelo, que la forma funcional elegida es la correcta y que los supuestos que planteamos sobre las variables estocásticas (que veremos seguidamente) son ciertos. Este supuesto es lo suficientemente restrictivo como para cuestionar las conclusiones extraídas en el

momento que se detecta algún error en las cuestiones anteriores. Así, debemos tener presente que en todo momento los resultados basados en el análisis de regresión lineal están condicionados al modelo escogido, debiéndose pensar cuidadosamente su formulación.

- ***Supuestos estocásticos***

1. El valor medio o esperanza de la perturbación u_i es igual a cero para todas las observaciones *i*. (Ya razonado anteriormente).

   Esto significa que los factores explicativos incluidos en la perturbación no influyen de forma sistemática en el valor promedio de *Y*.

   Si se considera el vector o matriz columna de las perturbaciones asociadas a cada observación *i*, se expresaría de este modo:

$$E(u) = \theta_{n\times 1}, \text{ o lo que es lo mismo: } E\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

2. La varianza de la perturbación aleatoria es igual para todas las observaciones *i*; es decir, es algún número positivo constante que denotaremos por σ_u^2. Es lo que se denomina *homoscedasticidad*.

$$Var(u_i) = E[u_i - E(u_i)]^2 = E(u_i^2) - E(u_i)^2 = E(u_i^2) = \sigma_u^2$$

$$\forall i = 1,2,...n$$

   Este supuesto implica que la variabilidad en torno a la media de las observaciones de la variable dependiente es igual para cualquier observación o valor de la variable explicativa, es decir, que las varianzas condicionadas de *Y* son también homoscedásticas.

3. No existe autocorrelación entre las perturbaciones. Dados dos valores cualesquiera *i* y *j*, la correlación entre las correspondientes perturbaciones u_i y u_j, es cero.

$$Cov(u_i, u_j) = E[(u_i - E(u_i)) \cdot (u_j - E(u_j))] = E(u_i \cdot u_j) - E(u_i) \cdot E(u_j) = E(u_i \cdot u_j) = 0$$

$$\forall i \neq j$$

   Esto significa que, dados los valores de *X*, las desviaciones de dos valores cualesquiera de *Y* en relación a su media no muestran patrones sistemáticos. Expresado en forma sencilla, este supuesto implica que el término de perturbación relacionado con una observación no está influenciado por el término de perturbación de otra observación diferente.

 ISBN: 978-84-694-7251-4

Estos dos últimos supuestos sobre la perturbación aleatoria, homoscedasticidad y no autocorrelación, pueden expresarse matricialmente mediante su matriz (simétrica) de varianzas-covarianzas de la siguiente forma:

$$Var-Cov(u)=\begin{pmatrix} Var(u_1) & Cov(u_1,u_2) & Cov(u_1,u_3) & \cdots & Cov(u_1,u_n) \\ & Var(u_2) & Cov(u_2,u_3) & \cdots & Cov(u_2,u_n) \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ & & & & Var(u_n) \end{pmatrix}=\begin{pmatrix} \sigma_u^2 & 0 & 0 & \cdots & 0 \\ & \sigma_u^2 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ & & & & \sigma_u^2 \end{pmatrix}=\sigma_u^2 \cdot I_{n\times n}$$

4. Las covarianzas entre la perturbación y los valores de las variables explicativas *X* son cero.

$$Cov(u_i, X_{ji})=E\left[(u_i-E(u_i))\cdot(X_{ji}-E(X_{ji}))\right]=E(u_i\cdot X_{ji})-E(u_i)\cdot E(X_{ji})=E(u_i\cdot X_{ji})=0$$

$$\forall j=1,2,...,k \qquad \forall i=1,2,...,n$$

Este supuesto implica asumir que *X* y *u* tienen una influencia separada y aditiva sobre *Y*, y por tanto es posible determinar por separado sus efectos sobre la variable dependiente.

- ***Supuestos sobre la distribución de probabilidad***

1. La perturbación aleatoria *u* se distribuye según una distribución normal. (Ya comentado).

   Teniendo en cuenta este supuesto, así como los supuestos estocásticos sobre la media y la varianza y la covarianza de cada una de las u_i, tenemos que: $u_i \rightarrow N(0,\sigma_u^2),\ \forall i=1,2,\ldots,n$.

   Matricialmente, refiriéndonos al vector de perturbación aleatoria *u*, esto se expresaría:

$$\boxed{u \rightarrow N_n\left(\theta_{n\times 1};\sigma_u^2\cdot I_{n\times n}\right)}.$$

   Consecuencia: Dado que la variable *Y* depende de *u* (a través de la expresión $Y=X\beta+u$), y puesto que *u* es un vector aleatorio normal, *Y* resulta ser también un vector aleatorio normal.

   En particular, su media o esperanza matemática va a ser: $E(Y)=X\beta$.

   Este resultado se puede deducir fácilmente de:

   $E(Y)=E(X\beta)+E(u)=E(X\beta)=X\beta$, ya que $E(u)=\theta$ y *X* son no estocásticas.

   En cuanto a la matriz de varianzas-covarianzas de *Y*, tendremos que ésta coincide con la de la perturbación aleatoria *u*: $Var-Cov(Y)=\sigma_u^2\cdot I_{n\times n}$, ya que:

$Var(Y_i) = E[Y_i - E(Y_i)]^2 = E(u_i^2) = Var(u_i) = \sigma_u^2, \ \forall i$, y además las covarianzas entre distintas observaciones de *Y* son cero, como se puede ver si se deduce a partir de la definición de la expresión de la covarianza y teniendo en cuenta que la perturbación aleatoria es la diferencia entre las observaciones de *Y* y sus valores esperados:

$$Cov(Y_i, Y_j) = E[(Y_i - E(Y_i)) \cdot (Y_j - E(Y_j))] = E(u_i \cdot u_j) = 0, \ \forall i \neq j .$$

Así pues, en definitiva, la distribución de probabilidad del vector de todas las observaciones de *Y*, resulta ser:

$$\boxed{Y \to N_n(X\beta, \sigma_u^2 \cdot I_{n \times n})}.$$

2.3. Estimación por mínimos cuadrados ordinarios (MCO). Propiedades de los estimadores MCO. Interpretación de los coeficientes de regresión. Efecto marginal.-

Estimación por mínimos cuadrados ordinarios (MCO)

El objetivo en el análisis de regresión es estimar la FRP a partir de la información proporcionada por una muestra en la forma más precisa posible, es decir, llegar a tener una FRM.

En términos generales hay dos métodos principales de estimación: el método de los mínimos cuadrados ordinarios (MCO) y el método de máxima verosimilitud (MV).

El método de MCO, que estudiamos en este apartado, es el que más se emplea en el análisis de regresión, sobre todo por ser en gran medida intuitivo y matemáticamente más simple que el método de MV.

Considérese el modelo de regresión lineal general:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, 2, ..., n ,$$

que no resulta observable de forma directa. Mediante el análisis de regresión, realizamos una estimación del mismo, obteniendo finalmente la FRM:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, 2, ..., n .$$

Como ya vimos anteriormente de manera gráfica para el caso de un modelo de regresión lineal simple (*Figura 3*), esta aproximación va a registrar desviaciones respecto a los valores reales de *Y*, obteniéndose un error o residuo muestral:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} .$$

Si nos fijamos, partiendo de una nube de puntos, nuestro objetivo final sería encontrar aquella función lineal concreta, de entre las infinitas posibilidades existentes, que fuese

 ISBN: 978-84-694-7251-4

lo "más representativa" posible de dicha nube de puntos; esto es, aquélla que diese lugar a valores estimados de Y ($\hat{Y}_i$) que hiciera que los residuos fuesen los más pequeños posibles en su conjunto.

De este modo, nuestro objetivo se puede alcanzar mediante un problema de optimización matemática; en particular, de minimización de una función que sería la suma de los residuos al cuadrado[5] (*SCR*):

$$\underset{\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k}{Min}\ SCR \equiv \underset{\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k}{Min}\ \sum_{i=1}^{n} e_i^2 = \underset{\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k}{Min}\ \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 =$$

$$= \underset{\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k}{Min}\ \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki}\right)^2 .$$

Este método de estimación recibe el nombre de Mínimos Cuadrados Ordinarios (MCO).

Para obtener la solución óptima a través de este método, el llamado estimador mínimo-cuadrático, calculamos los puntos críticos de la función, es decir, aquéllos que anulan el gradiente (primeras derivadas) de *SCR* (condición necesaria de 1er orden) y posteriormente, según el signo de la matriz Hessiana (segundas derivadas) de *SCR*, determinamos si se trata de un máximo o un mínimo (condición suficiente de 2º orden).

Así pues, si calculamos el gradiente de nuestra función objetivo *SCR*, tendremos que:

$$\nabla SCR = \begin{pmatrix} \frac{\partial\, SCR}{\partial\, \hat{\beta}_1} \\ \frac{\partial\, SCR}{\partial\, \hat{\beta}_2} \\ \vdots \\ \frac{\partial\, SCR}{\partial\, \hat{\beta}_j} \\ \vdots \\ \frac{\partial\, SCR}{\partial\, \hat{\beta}_k} \end{pmatrix} = \theta .$$

Donde:

- $\frac{\partial\, SCR}{\partial\, \hat{\beta}_1} = \sum_{i=1}^{n} 2 \cdot \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki}\right) \cdot (-1) =$

[5] Al considerar la suma, estaríamos teniendo en cuenta el conjunto de todos los residuos. El hecho de tomar la suma del cuadrado de los residuos, en lugar de la suma de dichos residuos directamente, radica fundamentalmente en que, si se observa, los valores de los residuos serán en unos casos positivos y en otros negativos. Al tomar la suma de todos ellos, las desviaciones de un signo se podrían compensar con las del otro signo y acabar finalmente anulándose, desvirtuándose entonces nuestro objetivo. Esto, sin embargo, no ocurrirá si tomamos el cuadrado de los residuos.

$$= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) = 0 ;$$

es decir: $\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right) = \boxed{\sum_{i=1}^{n} e_i = 0}$

- $\frac{\partial SCR}{\partial \hat{\beta}_2} = \sum_{i=1}^{n} 2 \cdot \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot \left(- X_{2i} \right) =$

$$= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{2i} = 0 ;$$

esto es:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{2i} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right) \cdot X_{2i} = \boxed{\sum_{i=1}^{n} e_i \cdot X_{2i} = 0}$$

Si seguimos calculando la primera derivada parcial de *SCR* respecto a los sucesivos parámetros o coeficientes de regresión $\hat{\beta}_j$, tendremos que:

- $\frac{\partial SCR}{\partial \hat{\beta}_j} = \sum_{i=1}^{n} 2 \cdot \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot \left(- X_{ji} \right) =$

$$= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{ji} = 0 ;$$

por tanto:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{ji} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right) \cdot X_{ji} = \boxed{\sum_{i=1}^{n} e_i \cdot X_{ji} = 0}$$

Finalmente, para el caso de la derivada respecto al último parámetro $\hat{\beta}_k$, obtendremos:

- $\frac{\partial SCR}{\partial \hat{\beta}_k} = \sum_{i=1}^{n} 2 \cdot \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot \left(- X_{ki} \right) =$

$$= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{ki} = 0 ;$$

por consiguiente:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_j X_{ji} - ... - \hat{\beta}_k X_{ki} \right) \cdot X_{ki} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right) \cdot X_{ki} = \boxed{\sum_{i=1}^{n} e_i \cdot X_{ki} = 0}$$

 ISBN: 978-84-694-7251-4

Según hemos podido apreciar, la aplicación de la condición necesaria de optimalidad, es decir, el cálculo de las derivadas parciales de la función *SCR* respecto a cada uno de los parámetros $\hat{\beta}_j$ igualadas a cero, da lugar a un sistema de *k* ecuaciones, que reciben el nombre de *ecuaciones normales*.

La resolución del sistema de ecuaciones normales nos daría los valores de los estimadores de los parámetros β_j (es decir, los $\hat{\beta}_j$), que podrían hacer mínimo el valor de la función *SCR*.[6] Para poder asegurar que, en efecto, minimizan la *SCR*, habría luego que aplicar la condición suficiente, como ya se indicó anteriormente. Al llevar a cabo el estudio del signo de la matriz Hessiana, comprobaríamos que al ser ésta definida positiva[7], podemos asegurar que nuestros valores $\hat{\beta}_j$ obtenidos como solución del sistema de ecuaciones normales constituyen efectivamente un mínimo para la *SCR*.

De este modo, habríamos obtenido la estimación por MCO de nuestro modelo:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, 2, ..., n .$$

La resolución del sistema de ecuaciones normales y, consiguientemente, la obtención de los estimadores $\hat{\beta}$ MCO podemos realizarla mediante el análisis matricial.

Como ya sabemos, nuestro modelo de regresión lineal general puede expresarse:

$$Y = X\beta + u .$$

La estimación del vector β supone la estimación de nuestro modelo: $\hat{Y} = X\hat{\beta}$.

De este modo, podemos definir el vector de residuos muestrales como:

$$e = Y - \hat{Y} = Y - X\hat{\beta} ,$$

siendo *e* la matriz-vector columna:

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} .$$

[6] En el caso de que estuviésemos considerando un modelo de regresión lineal simple, obtendríamos los valores de $\hat{\beta}_1$ y $\hat{\beta}_2$ ya conocidos de las materias de Estadística:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \qquad \text{y} \qquad \hat{\beta}_2 = \frac{Cov(X,Y)}{Var(X)} = \frac{S_{XY}}{S_X^2} .$$

[7] Este paso se obviará por motivos de simplificación de nuestra exposición, pudiendo encontrarse en cualquier manual de Econometría.

 ISBN: 978-84-694-7251-4

Si desarrollamos la expresión de la *SCR* en el caso matricial[8], obtenemos:

$$SCR = \sum_{i=1}^{n} e_i^2 = e'e = \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = \left(Y' - \hat{\beta}' X'\right)\left(Y - X\hat{\beta}\right) =$$

$$= Y'Y - \hat{\beta}' X'Y - Y' X\hat{\beta} + \hat{\beta}' X' X\hat{\beta}\,.$$

En esta expresión se cumple que: $\hat{\beta}' X'Y = Y' X\hat{\beta}$, puesto que un lado de la ecuación es el traspuesto del otro y se trata de un escalar (un número); así, pues, tenemos que:

$$SCR = Y'Y - 2\hat{\beta}' X'Y + \hat{\beta}' X' X\hat{\beta}\,.$$

Por tanto, el desarrollo del método de estimación por MCO de manera matricial, podría escribirse como sigue:

$$\underset{\hat{\beta}}{Min}\ SCR \equiv \underset{\hat{\beta}}{Min}\ \sum_{i=1}^{n} e_i^2 \equiv \underset{\hat{\beta}}{Min}\ e'e = \underset{\hat{\beta}}{Min}\ \left(Y'Y - 2\hat{\beta}' X'Y + \hat{\beta}' X' X\hat{\beta}\right)$$

La aplicación de la condición necesaria de primer orden daría como resultado:

$$\frac{\partial SCR}{\partial \hat{\beta}} = -2X'Y + 2X' X\hat{\beta} = \theta\,.$$

A partir de ahí, se obtendría la expresión matricial del sistema de ecuaciones normales:

$$X' X\hat{\beta} = X'Y\,.$$

Y despejando la matriz $\hat{\beta}$, esto es, el vector de coeficientes de regresión estimados, tendríamos que:

$$\boxed{\hat{\beta} = \left(X' X\right)^{-1} X'Y}.$$

La aplicación de la condición suficiente de segundo orden nos demostraría posteriormente que esta solución (punto crítico) representa efectivamente un mínimo de la función objetivo *SCR*.

[8] A la hora de trabajar con el análisis matricial, deben tenerse en cuenta las propiedades de las operaciones con matrices; en particular, las más importantes son:

1. $(A + B)' = A' + B'$.
2. $(A \cdot B)' = B' \cdot A'$
3. $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$
4. $(A')^{-1} = (A^{-1})'$

Para poder realizar estas operaciones, las matrices deben cumplir los requisitos exigibles en lo que a sus órdenes respectivos se refiere.

 ISBN: 978-84-694-7251-4

La utilización del análisis matricial a la obtención de los estimadores del modelo por el método de MCO presenta como ventaja que se obtienen "de una vez" los valores de todos los parámetros del vector $\hat{\beta}$: $\left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_j, ..., \hat{\beta}_k\right)$.

En este punto, cabe hacer las siguientes observaciones:

1. La matriz $X'X$, que es cuadrada, simétrica y de orden *k*, debe ser no singular; esto es, su determinante ha de ser distinto de cero. O lo que es lo mismo, no puede haber multicolinealidad (relación lineal) perfecta entre las variables explicativas del modelo. De este modo, el rango de $X'X$ será *k* y se asegurará así la existencia de la matriz inversa de $X'X$, posibilitando la obtención del vector de los coeficientes de regresión $\hat{\beta}$.
2. El número de observaciones debe ser sensiblemente superior al número de variables explicativas: $n \geq k$. Si $n < k$, entonces el estimador MCO no está unívocamente definido (matemáticamente, el sistema de ecuaciones normales es compatible indeterminado). Y si $n = k$, el modelo no tiene grados de libertad (concepto que veremos más adelante).

Como se ha visto, la obtención de $\hat{\beta}$ supone la consideración de las matrices $X'X$ y $X'Y$. Seguidamente se muestran las expresiones operativas de éstas para cuando se precisen:

$$X'X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \cdot \begin{pmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{3i} & \cdots & \sum_{i=1}^{n} X_{ki} \\ & \sum_{i=1}^{n} X_{2i}^2 & \sum_{i=1}^{n} X_{2i}X_{3i} & \cdots & \sum_{i=1}^{n} X_{2i}X_{ki} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ & & & & \sum_{i=1}^{n} X_{ki}^2 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{2i}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ki}Y_i \end{pmatrix}.$$

Propiedades de los estimadores MCO

Los estimadores MCO poseen una serie de propiedades, que podemos diferenciar en dos tipos. Por un lado, las propiedades numéricas, y por otro, las propiedades estadísticas o probabilísticas.

 ISBN: 978-84-694-7251-4

- ***Propiedades numéricas de los estimadores MCO***

Las propiedades numéricas de los estimadores MCO son aquéllas que se mantienen como consecuencia de la aplicación del método de MCO sobre el modelo de regresión, sin considerar la forma en la que se generaron los datos. Son:

1. Los estimadores $\hat{\beta}$ por MCO están expresados en términos de cantidades observables muestrales. Por consiguiente, pueden ser fácilmente calculados.
2. Los estimadores $\hat{\beta}$ por MCO son estimadores puntuales; es decir, dada la muestra, cada estimador proporciona un solo valor (puntual) del parámetro poblacional correspondiente.
3. Una vez determinada por MCO la FRM (recta, o hiperplano de regresión en el caso múltiple), se cumplirá que:
   a) La FRM pasa necesariamente por las medias muestrales de la variable explicada Y y de todas las variables explicativas X.
   b) La suma de los residuos mínimo-cuadráticos es igual a cero: $\sum_{i=1}^{n} e_i = 0$, siempre y cuando el modelo tenga ordenada en el origen (como señala la primera ecuación normal, referida al parámetro β_1).
   c) Los residuos e_i no están correlacionados con las variables explicativas X_j, $\forall j = 1,2,...,k$, es decir: $\sum_{i=1}^{n} e_i \cdot X_{ji} = 0$.

      Esta propiedad es consecuencia directa de lo expresado por las ecuaciones normales derivadas de todos y cada uno de los parámetros β_j que acompañan a las correspondientes variables explicativas X_j.

      Matricialmente, esta propiedad, de ortogonalidad entre las variables explicativas y los residuos mínimo-cuadráticos, se puede expresar: $X'e = \theta$.

      Su demostración sería muy sencilla:

      $$X'e = X'\left(Y - X \cdot \hat{\beta}\right) = X'Y - X'X\hat{\beta} = X'Y - X'X\left(X'X\right)^{-1}X'Y = X'Y - X'Y = \theta.$$

   d) Los residuos e_i están incorrelados con $\hat{Y}_i$, es decir: $\sum_{i=1}^{n} e_i \cdot \hat{Y}_i = 0$, $\forall i = 1,2,...,n$.

      Matricialmente, esta propiedad, de ortogonalidad entre la variable explicada del modelo y los residuos mínimo-cuadráticos, se puede expresar: $\hat{Y}'e = \theta$. La demostración, muy sencilla, sería:

      $$\hat{Y}'e = \left(X\hat{\beta}\right)'e = \hat{\beta}'X'e = \hat{\beta}'\theta = \theta.$$

e) El valor promedio del valor estimado de *Y* coincide con el valor medio de la variable real *Y*: $\overline{\hat{Y}} = \overline{Y}$.

La demostración de esta propiedad comienza teniendo en cuenta que: $e_i = Y_i - \hat{Y}_i$, o lo que es lo mismo: $Y_i = \hat{Y}_i + e_i$. De este modo, se cumple que: $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \hat{Y}_i$; dividiendo ambos miembros de la igualdad por *n*, se cumple finalmente que: $\overline{Y} = \overline{\hat{Y}}$.

- ***Propiedades estadísticas de los estimadores MCO***

Frente a las anteriores, las propiedades estadísticas o probabilísticas de los estimadores MCO están basadas en los supuestos del modelo (ya analizados anteriormente) y están avaladas por el Teorema de Gauss-Markov. En este sentido, dados los supuestos del modelo clásico de regresión lineal, los valores estimados de los parámetros por MCO poseen algunas propiedades ideales u óptimas.

1. El vector de parámetros o coeficientes de regresión estimados del modelo, $\hat{\beta}$, es un vector aleatorio que sigue una distribución de probabilidad normal.

   Demostración:

   Sea el modelo: $Y = X\beta + u$. Como sabemos, la estimación por MCO de β se obtiene a través de la expresión: $\hat{\beta} = (X'X)^{-1}X'Y$.

   Desarrollando ésta se tiene que:

   $$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u =$$
   $$= \beta + (X'X)^{-1}X'u$$

   Así pues: $\hat{\beta} = \beta + (X'X)^{-1}X'u$. Dado que, de acuerdo con esta expresión, $\hat{\beta}$ depende de *u*, y *u* es un vector aleatorio normal, entonces se deduce que $\hat{\beta}$ también es un vector aleatorio normal.

2. Si $E(u) = \theta$, entonces $\hat{\beta}$ es un estimador insesgado de β; es decir, $E(\hat{\beta}) = \beta$.

   Demostración:

   $$E(\hat{\beta}) = E(\beta + (X'X)^{-1}X'u) = \beta + (X'X)^{-1}X'E(u) = \beta + \theta = \beta.$$

3. Si $Var-Cov(u) = \sigma_u^2 \cdot I$, entonces $Var-Cov(\hat{\beta}) = \sigma_u^2 \cdot (X'X)^{-1}$.

   Demostración:

   $$Var-Cov(\hat{\beta}) = E\left[(\hat{\beta} - E(\hat{\beta}))\cdot(\hat{\beta} - E(\hat{\beta}))'\right] = E\left[(\hat{\beta} - \beta)\cdot(\hat{\beta} - \beta)'\right];$$

teniendo en cuenta que: $\hat{\beta} = \beta + (X'X)^{-1}X'u$, entonces: $\hat{\beta} - \beta = (X'X)^{-1}X'u$.

Por tanto, retomando la expresión anterior tendremos que:

$$Var - Cov(\hat{\beta}) = E\left[\left((X'X)^{-1}X'u\right)\cdot\left((X'X)^{-1}X'u\right)'\right] = E\left[(X'X)^{-1}X'u\,u'X(X'X)^{-1}\right] =$$

$$= (X'X)^{-1}X'E[u\,u']X(X'X)^{-1} = (X'X)^{-1}X'Var - Cov(u)X(X'X)^{-1} =$$

$$= (X'X)^{-1}X'\sigma_u^2 \cdot I\,X(X'X)^{-1} = \sigma_u^2 \cdot (X'X)^{-1}X'I\,X(X'X)^{-1} = \sigma_u^2 \cdot (X'X)^{-1}.$$

Como conclusión de las tres propiedades anteriores, tenemos en definitiva que:

$$\boxed{\hat{\beta} \to N_k\left(\beta; \sigma_u^2 \cdot (X'X)^{-1}\right)}.$$

4. Teorema de Gauss-Markov[9]: <u>El estimador MCO es un estimador lineal, insesgado y óptimo (ELIO)</u>, entendiendo por óptimo que tiene mínima varianza.

   Por su propia definición, el método de estimación por MCO nos proporciona los estimadores óptimos a nivel muestral. El Teorema de Gauss-Markov es fundamental, ya que nos garantiza, además, que este método de estimación nos proporciona los mejores resultados posibles también a nivel inferencial, ya que los estimadores cumplen las propiedades deseables que se le exigen a un buen estimador.

5. La combinación lineal $C'\hat{\beta}$ es ELIO de $C'\beta$, donde C' es un vector de constantes numéricas de orden 1 x *k*.

   Para entender bien esta propiedad, podemos indicar que la estimación ELIO, por ejemplo, de $\beta_1 + 2\beta_2$ es $\hat{\beta}_1 + 2\hat{\beta}_2$.

6. El vector de residuos MCO puede expresarse como una transformación lineal de la variable dependiente y también de las perturbaciones aleatorias.

   $$e = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = \left(I - X(X'X)^{-1}X'\right)Y = \underline{M\,Y} =$$
   $$= M\left(X\hat{\beta} + u\right) = M\,X\hat{\beta} + M\,u = \theta + M\,u = \underline{M\,u}$$

   Una conclusión importante de esta propiedad es que, al poder expresarse el vector de los residuos linealmente en función del vector de perturbaciones aleatorias, los residuos también resultan ser un vector aleatorio que, además, sigue una distribución de probabilidad normal (como *u*).

   De manera adicional, <u>también se va a verificar que</u>: $\underline{e'e = u'M\,u}$.

[9] Obviamos aquí la demostración de este Teorema, pudiendo consultarse en cualquier manual de Econometría.

 ISBN: 978-84-694-7251-4

El cumplimiento de lo establecido en esta propiedad se basa, a su vez, en una serie de propiedades que posee la matriz *M*:

a) M es singular: $|M| = 0$.

b) M es simétrica: $M = M'$.

c) M es idempotente: $M = M^n$, $\forall n = 2,3,....$

d) $M.X = \theta$.

Demostración:

Basándonos en las propiedades de *M*, se puede demostrar la última parte de lo indicado en esta propiedad del modo que sigue:

$$e'e = (M\,u)'M\,u = u'M'M\,u = u'M\,M\,u = u'M^2\,u = u'M\,u\,.$$

7. El valor esperado del vector de residuos mínimo-cuadráticos y su matriz de varianzas-covarianzas son, respectivamente:

$$E(e) = \theta \quad \text{y} \quad Var-Cov(e) = \sigma_u^2 \cdot M\,.$$

Obsérvese, a partir de la expresión de la matriz de varianzas-covarianzas, cómo mientras las perturbaciones aleatorias son incorreladas[10], los residuos en cambio son linealmente dependientes.

Demostración:

- $E(e) = E(M\,u) = M\,E(u) = M \cdot \theta = \theta$
- $Var-Cov(e) = Var-Cov(M\,u) = M^2 \cdot Var-Cov(u) = M^2 \cdot \sigma_u^2 \cdot I =$
  $= \sigma_u^2 \cdot M^2 \cdot I = \sigma_u^2 \cdot M^2 = \sigma_u^2 \cdot M$

Como conclusión de las dos últimas propiedades, 6 y 7, tenemos que:

$$\boxed{e \to N_n(\theta; \sigma_u^2 \cdot M)}.$$

- ***Propiedades del estimador MCO de la varianza de la perturbación aleatoria***

Además de los coeficientes de regresión β, según se ha podido ir viendo, en el modelo hay otro importante parámetro a estimar: σ_u^2, esto es, la varianza de la perturbación aleatoria. En tanto que este parámetro no se estime, todas las expresiones en las que aparezca tendrán carácter poblacional. Sólo cuando se estime a partir de datos

[10] Recuérdese que la matriz de varianzas-covarianzas de *u* hemos asumido que es diagonal: $Var-Cov(u) = \sigma_u^2 \cdot I$, esto es, las covarianzas referidas a distintas observaciones valen 0. Sin embargo, la matriz $Var-Cov(e) = \sigma_u^2 \cdot M$ no es diagonal, por lo que las covarianzas entre distintas observaciones son distintas de 0.

 ISBN: 978-84-694-7251-4

muestrales, dichas expresiones se convertirán en estimaciones y, por tanto, en cantidades numéricas concretas de carácter muestral.

Si nos centramos en el método de MCO, podemos citar tres importantes propiedades[11] relativas a σ_u^2:

1. El estimador de σ_u^2 viene dado por la expresión: $\boxed{\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{e'e}{n-k} = \frac{SCR}{n-k}}$.
2. Este estimador resulta ser insesgado; esto es: $E\left[\hat{\sigma}_u^2\right] = \sigma_u^2$.
3. $\hat{\sigma}_u^2$ es una variable aleatoria de la que se no conoce exactamente su distribución de probabilidad. Sin embargo, sí se sabe que la siguiente expresión, donde se incluye, sigue una distribución de tipo *chi-cuadrado*. En particular:

$$\boxed{\frac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k) \rightarrow \chi_{n-k}^2}.$$

Interpretación de los coeficientes de regresión

Consideremos el modelo de regresión lineal general:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n.$$

Como se ha visto, una vez estimado, tenemos que:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n.$$

Dados los supuestos del modelo de regresión clásico, la estimación obtenida en el modelo de regresión permite predecir el valor esperado de la variable dependiente en función de los valores dados de las variables independientes.

Los parámetros o coeficientes de regresión estimados $\hat{\beta}$ del modelo tienen un significado muy concreto. En concreto:

- $\hat{\beta}_j$, $\forall j = 2,...,k$, expresa la variación media que experimenta el valor de la variable dependiente Y cuando se incrementa en 1 unidad[12] el valor de la variable explicativa X_j, manteniendo constantes los valores del resto de variables explicativas presentes en el modelo. La variación referida puede ser

[11] No vamos a demostrar ni deducir matemáticamente ninguna de las propiedades referidas a σ_u^2. Éstas pueden ser consultadas por el alumno en cualquier manual de Econometría.

[12] Habitualmente, en el ámbito de la Economía y la Empresa, se suelen considerar variaciones unitarias. No obstante, dichas variaciones pueden ser perfectamente infinitesimales, dependiendo del problema analizado.

 ISBN: 978-84-694-7251-4

positiva o negativa, dependiendo del signo que tenga $\hat{\beta}_j$, expresando así si la relación entre la variable Y y la variable X_j en cuestión es directa (signo positivo) o inversa (signo negativo).

- $\hat{\beta}_1$ es la ordenada en origen del modelo. Proporciona el valor esperado de la variable dependiente Y cuando se consideran nulas todas las variables independientes X. En ocasiones, dependiendo del problema que se esté analizando, este valor puede no tener sentido económico.

Efecto marginal

El efecto marginal de una variable explicativa X_j sobre la variable explicada Y es un concepto económico que expresa la relación entre variaciones absolutas de Y ante variaciones absolutas de X_j. En concreto, expresa la variación media que experimenta la variable dependiente Y cuando se incrementa en 1 unidad el valor de la variable explicativa X_j, manteniendo constantes las demás variables explicativas que pueda haber en el modelo[13]. Es decir:

$$\boxed{\text{Efecto marginal} \equiv \frac{\Delta Y}{\Delta X_j}}.$$

Si nos fijamos, el efecto marginal coincide con el significado de los coeficientes de regresión en el caso de un modelo de regresión lineal. No obstante, como se verá más adelante, en otros tipos de modelos los coeficientes de regresión no van a coincidir con este concepto.

2.4. Bondad del ajuste. El coeficiente de determinación. El coeficiente de determinación corregido.-

Bondad del ajuste

Después de estimar la función de regresión poblacional, en concreto sus coeficientes y de ver sus propiedades, vamos a analizar en este punto la bondad del ajuste obtenido, esto es, cómo es el ajuste de nuestro modelo al conjunto de datos de que disponemos.[14]

[13] Este concepto se corresponde matemáticamente con el de la derivada parcial: $\frac{\partial Y}{\partial X_j}$.

[14] En este punto, cabe resaltar la diferencia entre los conceptos de "buen ajuste" y "mejor ajuste". Ya hemos visto que el método de estimación por MCO nos proporciona el mejor ajuste posible a los datos de la muestra disponible (nube de puntos); sin embargo, ello no significa necesariamente que dicho ajuste sea "bueno". Puede que el mejor ajuste posible sea "malo". Así pues, lo que estudiaremos ahora es si nuestro ajuste obtenido, aun siendo el mejor posible, es bueno o no.

Para responder a esta cuestión, analizamos e intentamos determinar qué parte de la variabilidad de la variable dependiente *Y* puede atribuirse a la variabilidad de las variables explicativas *X* y qué parte no, atribuyéndose al efecto de la perturbación aleatoria *u*. En otras palabras, parte del comportamiento de *Y* es explicado por las variables *X*; en cambio hay otra parte que no consigue ser explicada y, consiguientemente, se le atribuye a la perturbación aleatoria *u*, a la variable aleatoria que no "controlamos". En este sentido, pues, cuanto mayor sea la parte de la variabilidad de *Y* que es explicada por la variabilidad de las variables independientes, mejor se ajusta nuestro modelo estimado de regresión a la muestra de datos considerada.[15]

La descomposición de la variabilidad de *Y* podemos analizarla *analítica* y *gráficamente* (esto último es posible si nos fijamos en un modelo de regresión simple).

La forma más sencilla de representar a un conjunto de datos muestrales Y_i de una determinada variable es tomar el valor de su media muestral ($\overline{Y}$). Por ello, cuando analizamos la variabilidad de la variable dependiente, hacemos referencia a la variabilidad que muestra en torno a su valor medio: $Y_i - \overline{Y}$.

Un indicador adecuado para considerar el conjunto de las desviaciones que muestran los datos individuales en relación a su media es la Suma de Cuadrados Totales (SCT): $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$. Si nos fijamos bien, la SCT no es más que el numerador de la expresión de la varianza de la variable explicada *Y* (el denominador sería *n*).

Teniendo en cuenta lo expuesto hasta ahora, nuestro objetivo va a ser descomponer en dos partes la SCT, de tal manera que, por un lado, se recojan las variaciones de *Y* atribuibles al conjunto de variables explicativas *X* presentes en el modelo de regresión y, por otro, aquéllas que provienen del término aleatorio *u* del mismo.

Podemos ver esto fácilmente de manera gráfica si tomamos como ejemplo un modelo de regresión lineal simple: $Y_i = \beta_1 + \beta_2 X_i + u_i, \quad i = 1, 2, ..., n$.

Como se puede ver en la *Figura 4*, para un valor concreto de la variable Y_0, tenemos:

$$Y_0 - \overline{Y} = \left(\hat{Y}_0 - \overline{Y}\right) + \left(Y_0 - \hat{Y}_0\right) = \left(\hat{Y}_0 - \overline{Y}\right) + e_0 .$$

Es decir, la variación de cada valor observado respecto a su media muestral puede descomponerse como la suma de la desviación del valor estimado proporcionado por la recta respecto al valor de la media muestral de la variable dependiente, más el valor del residuo de la recta de regresión.

[15] Cuando hablemos de bondad del ajuste, nos referiremos a la obtenida en relación a una muestra considerada. Teóricamente también podríamos referirnos a la población, pero no es lo que solemos tener en la realidad, dada la habitual imposibilidad de disponer de toda la información poblacional.

 ISBN: 978-84-694-7251-4

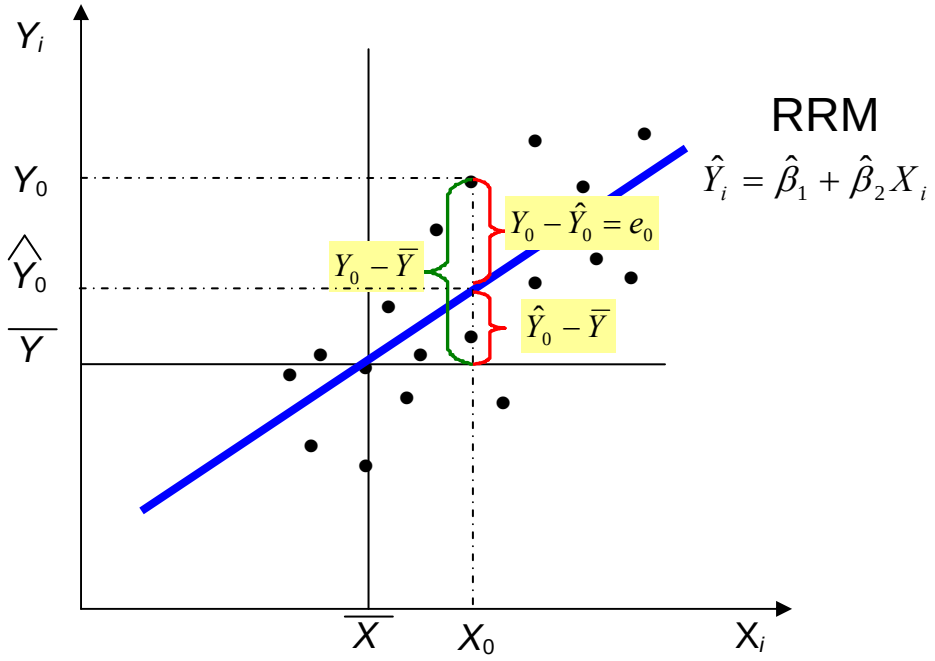


Figura 4

Si nos fijamos bien, esta descomposición puede perfectamente hacerse extensiva a un modelo de regresión lineal múltiple.

Por tanto, si tomamos en consideración la variabilidad total de *Y* a través de la SCT, vemos que ésta se podría escribir así:

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left[\left(\hat{Y}_i-\overline{Y}\right)+e_i\right]^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}e_i^2+2\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)\cdot e_i$$

Utilizando las propiedades numéricas 3-b) y 3-d) de los estimadores MCO (ya vistas), tenemos que[16]:

$$\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)\cdot e_i=\sum_{i=1}^{n}\hat{Y}_i\cdot e_i-\overline{Y}\cdot\sum_{i=1}^{n}e_i=0\,.$$

Así pues, hemos llegado a que:

$$\boxed{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}e_i^2}\,.$$

Esto significa que, al analizar la variabilidad de *Y*, parte de este comportamiento viene explicado por el modelo estimado (y con ello por las variables explicativas que conforman éste) y parte no, que se corresponde con los errores o residuos muestrales. El primero de los sumandos recibe el nombre de Suma de Cuadrados Explicados (SCE), en

[16] Si el modelo no tuviera ordenada en el origen, al no cumplirse la propiedad numérica 3-b) de los estimadores MCO, esta expresión no tendría entonces por qué ser 0 y, por tanto, el desarrollo matemático que se obtiene a partir de este punto no sería válido para este tipo de especificación del modelo.

 ISBN: 978-84-694-7251-4

tanto que el segundo es, como ya sabemos, la Suma de Cuadrados Residuales (SCR). Según esto, en definitiva tenemos que:

$$\boxed{SCT = SCE + SCR}.$$

De acuerdo con esta igualdad[17], resulta evidente que a la hora de llevar a cabo un ajuste de regresión, lo deseable es que el valor de la SCE sea lo mayor posible y, consecuentemente, el de la SCR lo menor posible. Es decir, el ajuste del modelo será tanto mejor, en tanto más alto sea la SCE en relación con la SCT. De este modo, se define un indicador para medir la bondad del ajuste: el coeficiente de determinación, que se denota por R^2.

El coeficiente de determinación

El coeficiente de determinación se define concretamente como: $\boxed{R^2 = \frac{SCE}{SCT}}$.

R^2 mide la proporción (en tanto por uno) o el porcentaje (si se multiplica por cien) de la variación total de la variable dependiente que es explicada por el conjunto de las variables explicativas que conforman el modelo de regresión.[18]

El coeficiente de determinación puede expresarse también como: $\boxed{R^2 = 1 - \frac{SCR}{SCT}}$.

Ello se deduce directamente al dividir la igualdad de la descomposición de las sumas de cuadrados entre STC:

$$\frac{STC}{STC} = \frac{SCE}{SCT} + \frac{SCR}{SCT} \Rightarrow 1 = R^2 + \frac{SCR}{SCT} \Rightarrow R^2 = 1 - \frac{SCR}{SCT}.$$

De estas expresiones puede deducirse, igualmente, que el rango de valores de R^2 es:

$$\boxed{0 \leq R^2 \leq 1}.$$

[17] Esta igualdad también se mantendría si dividiésemos todas las sumas de cuadrados por el tamaño de la muestra considerada (n); en este caso, nos encontraríamos con la conocida expresión estadística:

$$\frac{SCT}{n} = \frac{SCE}{n} + \frac{SCR}{n} \Rightarrow VT = VE + VNE;$$

esto es, la varianza total de Y (VT) es igual a la varianza explicada (VE) más la varianza no explicada o residual (VNE).

[18] No debe confundirse el concepto de coeficiente de determinación R^2 con el de coeficiente de correlación lineal R, que numéricamente se corresponde con su raíz cuadrada positiva. R expresa el grado de asociación lineal existente entre las variables en cuestión, moviéndose sus valores entre 0 y 1, si la relación entre las variables es directa, y entre -1 y 0, si la relación es inversa. Cuanto más cerca se halle el valor de 1 (ó -1), más fuerte es esta relación.

Si $R^2 = 1$, esto significaría un ajuste perfecto del modelo estimado, es decir: $SCE = SCT$, y lógicamente: $SCR = 0$.

Por el contrario, si $R^2 = 0$, ello supondría que no habría relación alguna entre la variable explicada y las variables explicativas: $SCE = 0$ y $SCR = SCT$. En este caso, $\hat{Y}_i = \hat{\beta}_1 = \overline{Y}$, es decir, la mejor predicción de cualquier valor de *Y* sería simplemente el valor de su media.

Por tanto, cuanto más cerca de 1 se halle el valor de R^2, tanto mejor será la bondad del ajuste analizado. En términos generales, la literatura estadístico-econométrica considera que un ajuste de regresión puede considerarse bueno si el coeficiente de determinación presenta un valor que se sitúa a partir, aproximadamente, de 0,75.

- ***Expresiones operativas de SCT, SCE y SCR***

En este punto vamos a mostrar distintas expresiones que pueden adoptar las sumas de cuadrados que hemos visto.[19]

- $SCT = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2 = Y'Y - n \cdot \overline{Y}^2$
- $SCE = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \hat{Y}_i^2 - n \cdot \overline{Y}^2 = \hat{Y}'\hat{Y} - n \cdot \overline{Y}^2 = \hat{\beta}'X'X\hat{\beta} - n \cdot \overline{Y}^2 = \hat{\beta}'X'Y - n \cdot \overline{Y}^2$
- $SCR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = Y'Y - \hat{Y}'\hat{Y} = Y'Y - \hat{\beta}'X'X\hat{\beta} = Y'Y - \hat{\beta}'X'Y$

A partir de éstas, pueden a su vez configurarse distintas expresiones del coeficiente de determinación R^2, dependiendo de cuáles se tomen en concreto. Ello dependerá normalmente del problema particular que se esté estudiando.[20]

El coeficiente de determinación corregido

A la hora de evaluar si un modelo es adecuado o no, es evidente que el coeficiente de determinación juega un papel importante; sin embargo, no es el único indicador a tener en cuenta. De hecho, su papel no debe sobrevalorarse. Obtener un valor elevado de R^2 no implica que nuestros resultados puedan ser fiables. Puede darse el caso extremo de tener un $R^2 = 1$, es decir, que el modelo presente un ajuste perfecto entre las variables

[19] Nos vamos a limitar simplemente a exponer estas expresiones. Sus deducciones y demostraciones pueden consultarse en cualquier manual de Econometría.

[20] En el caso particular de que consideremos un modelo de regresión lineal simple, las expresiones más habituales (y de sobra conocidas por las materias de Estadística) del coeficiente de determinación son:

$$R^2 = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = \hat{\beta}_2 \cdot \frac{S_{XY}}{S_Y^2} = \hat{\beta}_2^2 \cdot \frac{S_X^2}{S_Y^2}.$$

explicativas X y la variable explicada Y, pero sin embargo, que no tenga sentido económico o estadístico.

Podemos poner un ejemplo para ilustrar esto. Consideremos, por simplicidad, un modelo de regresión lineal simple, con ordenada en el origen (es decir, $k = 2$), y que para su estimación contamos únicamente con un tamaño muestral n de 2 observaciones. Así pues, en este caso: $n = k = 2$. Pensemos que nuestro modelo pretende estudiar el precio del transporte público de nuestra ciudad (Y) en función del precio del pan en un país de Extremo Oriente (X). Gráficamente, nuestro resultado podría ser el que se muestra en la *Figura 5*; esto es, puesto que nuestro modelo es una recta y la nube de puntos se compone únicamente de 2 puntos, obviamente la mejor recta posible pasará necesariamente por los 2 puntos. De acuerdo con ello, en este caso nuestro coeficiente de determinación arrojará un valor igual a 1.

Si atendemos a este resultado, nos encontramos con que el ajuste de nuestro modelo es perfecto y, sin embargo, a todas luces, desde el punto de vista de la interpretación económica, el resultado es absurdo.

¿Cuál es entonces el fallo? Pues éste radica en que estamos considerando el mismo número de observaciones muestrales (n) que de variables explicativas en el modelo (k). En este caso, matemáticamente el ajuste resulta necesariamente perfecto, pese a no tener ningún sentido.

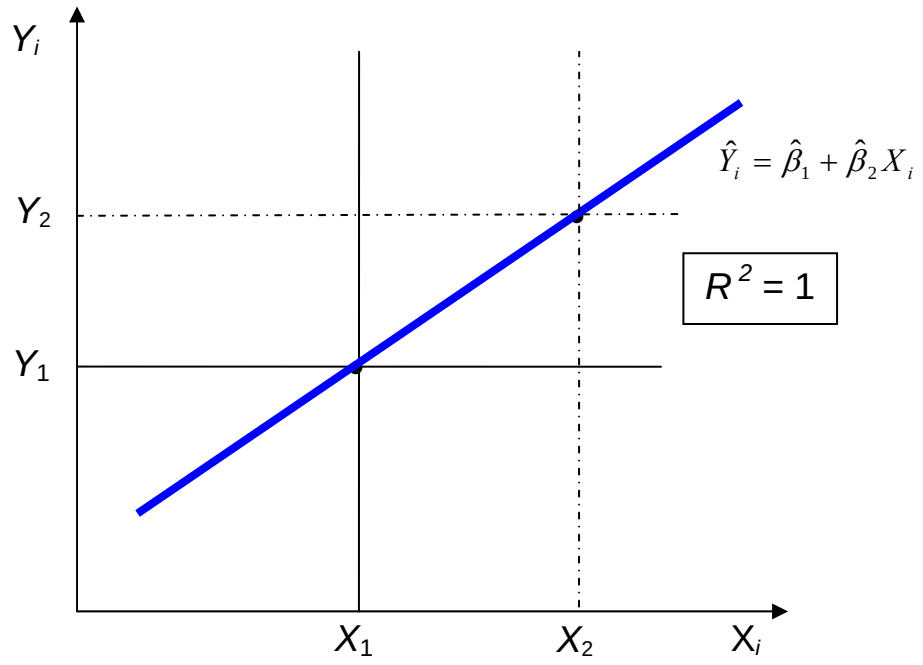


Figura 5

Si considerásemos en lugar de 2 observaciones muestrales, 3, el ajuste ya no sería perfecto y, consiguientemente, el valor de R^2 disminuiría. Si fuesen 4 las observaciones, R^2 sería todavía menor y así podríamos ir actuando sucesivamente. Es

 ISBN: 978-84-694-7251-4

decir, con un mayor tamaño muestral en relación al número de variables explicativas, el valor de R^2 iría decreciendo y la conclusión a la que se llegaría sería más realista. Volviendo a nuestro ejemplo, si el tamaño de la muestra fuese "grande", con toda seguridad acabaríamos teniendo un coeficiente de determinación con un valor bajo, siendo así coherente el resultado obtenido.

Por tanto, hay que tomar con precaución las conclusiones que se extraen sobre la bondad de ajuste de un modelo al interpretar el valor de R^2. Debemos tener presente el tamaño muestral con el que se trabaja en comparación con el número de variables explicativas del modelo. La diferencia entre ambos, $n-k$, es lo que se conoce como grados de libertad. Cuanto más elevado sea el valor de los grados de libertad de un ajuste, más realista, y por consiguiente fiable, será la explicación proporcionada por el coeficiente de determinación de un modelo. El modelo captará mejor la relación entre las variables cuanta más información muestral (considerada a través de sus observaciones) incorporemos en el estudio, es decir, cuantos más grados de libertad existan.

En definitiva, el fallo en nuestro ejemplo era que nuestro ajuste no tenía grados de libertad: $n-k=0$.

Otra forma de abordar la cuestión de los grados de libertad de un modelo es ver qué sucede si se produce un incremento del número de variables explicativas *k* del modelo. Esto conlleva, en la generalidad de los casos, un incremento del valor de R^2.

Cuando se introduce una nueva variable explicativa en el modelo, ésta en general siempre va a "aportar", es decir, va a mejorar la explicación de la variable dependiente del modelo, lo que supone un incremento de R^2. Obsérvese que en el peor de los casos simplemente no va a "aportar" nada, permaneciendo entonces invariable el valor del coeficiente de determinación.

Así pues, de acuerdo con esto, debemos tener presente que el incremento del número de variables explicativas en un modelo podría estar arrojándonos un valor elevado del R^2 un tanto "artificial", es decir, podría "enmascarar" un ajuste aparentemente bueno, aun no siéndolo en la realidad.

Si nos fijamos bien, si el número de observaciones muestrales *n* permanece constante y aumentamos el número de variables explicativas *k* del modelo, estamos reduciendo el número de grados de libertad $n-k$ de éste, lo que resta "credibilidad" al valor de R^2.

Tras lo visto en los puntos anteriores, cuando se habla del coeficiente de determinación que presenta un modelo, también es preciso hacerlo de los grados de libertad que tiene. Cuanto mayor sea el número de grados de libertad que presente un modelo, más fiable resultará el valor de su R^2, y viceversa.

 ISBN: 978-84-694-7251-4

Dicho esto, lo ideal sería poder tener un indicador que aunase las dos informaciones del modelo, esto es, el valor del coeficiente de determinación y el número de grados de libertad. De alguna manera, este indicador debería "penalizar" o corregir el valor de R^2 en función del número de grados de libertad, de forma que cuanto menor fuese éste, también fuese menor el valor de R^2.

En este sentido, surge entonces el denominado *coeficiente de determinación corregido*, que se denota por R_c^2 o $\overline{R}^2$, el cual matiza o penaliza, de alguna manera, la inclusión de nuevas variables explicativas en el modelo, o bien el escaso número de observaciones muestrales en relación al número de variables explicativas consideradas. Este coeficiente se define:

$$\boxed{R_c^2 = 1 - \left[\left(1 - R^2\right) \cdot \left(\frac{n-1}{n-k} \right) \right]}.$$

Nótese, según esta expresión, que si se produce un incremento del número de variables explicativas en el modelo, tendremos que:

$$\uparrow k \Rightarrow \downarrow (n-k) \Rightarrow \uparrow \frac{n-1}{n-k} \Rightarrow \downarrow R_c^2 .$$

Obsérvese también que este coeficiente nos sirve para ver si el número de observaciones muestrales que estamos considerando es "suficiente" o no, pues manteniendo fijo el número de variables explicativas *k*, cuando *n* tiende a infinito se tiene que:

$$R_c^2 = \lim_{n\to\infty} \left(1 - \left[\left(1 - R^2\right) \cdot \left(\frac{n-1}{n-k} \right) \right] \right) = 1 - \left[\left(1 - R^2\right) \cdot \lim_{n\to\infty} \left(\frac{n-1}{n-k} \right) \right] = 1 - \left[\left(1 - R^2\right) \cdot 1 \right] =$$

$$= 1 - \left(1 - R^2\right) = 1 - 1 + R^2 = R^2 .$$

Es decir, cuando el número de observaciones es suficientemente elevado, ambos coeficientes tienden a ser iguales. Esto nos puede servir para hacer el razonamiento inverso, esto es, si nos encontramos que los valores de R^2 y de R_c^2 son muy similares, ello significará que estamos ante un tamaño muestral "suficiente".

Pero debemos tener muy presente que, ante muestras pequeñas, el valor de R_c^2 nos arrojará información sobre la bondad del ajuste del modelo más fiable que la del R^2.

- ***Consideraciones sobre el uso del coeficiente de determinación y del coeficiente de determinación corregido***

1. El valor del coeficiente de determinación hay que tomarlo con precaución, puesto que en función del tamaño muestral, puede estar mostrando resultados engañosos y no del todo fiables.

 ISBN: 978-84-694-7251-4

2. En el caso de un modelo sin ordenada en el origen, según las razones matemáticas que se expusieron en su momento, no se tiene por qué cumplir la igualdad $SCT = SCE + SCR$. Este hecho tiene como consecuencia que el rango de valores de R^2 no se ciñe al intervalo (0, 1). De hecho, bajo determinadas circunstancias, en estos casos el valor del coeficiente de determinación puede llegar a ser incluso negativo. Tan sólo puede asegurarse que su valor es, como máximo, 1: $R^2 \leq 1$. Así pues, el coeficiente de determinación hay que tomarlo con mucha precaución.
3. Si no hay ordenada en el origen, y aun cuando $R^2 < 0$, el coeficiente de determinación se puede utilizar para comparar modelos que no presenten ordenada en el origen, siempre que tengan la misma variable dependiente, igual número de variables explicativas y utilicen una muestra del mismo tamaño, pero nunca se podrá tomar para comparar un modelo con ordenada en el origen con otro sin ordenada en el origen.
4. El coeficiente de determinación se puede utilizar para comparar modelos que presenten la misma variable dependiente, el mismo número de variables explicativas y una muestra de igual tamaño. En el caso de que la variable dependiente sea diferente, es necesario utilizar otros indicadores para juzgar la bondad de un modelo de regresión, como por ejemplo, el criterio de información de Akaike, que se verá más adelante.
5. Para modelos anidados con igual variable dependiente[21], la comparación debe realizarse utilizando el coeficiente de determinación corregido.
6. Para establecer si el modelo que se analiza resulta adecuado o no, no es suficiente estudiar únicamente la bondad del ajuste. El investigador debe preocuparse antes por estudiar la relevancia lógica o teórica que tienen las variables explicativas para la variable dependiente, así como la significatividad estadística de sus coeficientes, aspecto éste que se estudiará más adelante.

2.5. Introducción en el modelo de variables ficticias. Interpretación de los coeficientes de regresión.-

Introducción en el modelo de variables ficticias

Hasta este momento, hemos estado asumiendo que todas las variables de nuestro modelo econométrico eran de tipo cuantitativo, esto es, variables que toman de forma continua valores reales.

[21] Los modelos anidados son aquéllos que tienen en común una serie de variables explicativas, a las cuales se les suman además otras variables explicativas diferentes. Por ejemplo:

$$\text{Consumo}_i = \beta_1 + \beta_2 \cdot \text{Precio}_i + u_i$$

$$\text{Consumo}_i = \beta_1 + \beta_2 \cdot \text{Precio}_i + \beta_3 \cdot \text{Renta}_i + u_i$$

 ISBN: 978-84-694-7251-4

Sin embargo, la realidad está también plagada de factores de tipo cualitativo cuya inclusión en los modelos se puede hacer igualmente necesaria: sexo, estado civil, nivel de estudios, localización geográfica…; es decir, los modelos pueden tener entre sus variables explicativas, tanto variables cuantitativas como variables cualitativas. Aparecen entonces las denominadas variables ficticias, también conocidas como binarias o dicotómicas, o *dummy* (en terminología anglosajona), que reflejan la presencia o no de un determinado atributo.

Las variables ficticias se caracterizan porque:

- Toman únicamente los valores 1 y 0, de manera que éstos indican:
  $$\begin{cases} 1 \text{ – Presencia de determinado atributo.} \\ 0 \text{ – Ausencia del atributo considerado.} \end{cases}$$
- Suelen referirse a variables cualitativas: sexo, localización geográfica, etc.
- También son susceptibles de utilización para variables cuantitativas (por ejemplo, la edad), refiriéndose a "tramos" en los que se puede dividir el rango de valores de éstas.

A continuación, mostraremos un ejemplo ilustrativo para entender la interpretación y uso de las variables ficticias.

Interpretación de los coeficientes de regresión

- ***Ejemplo de una variable cualitativa con dos modalidades***

Supóngase que se pretende estudiar el "salario" que gana una población de "titulados universitarios", donde éstos pueden ser exclusivamente "licenciados" o "doctores", y se quiere conocer si existen diferencias salariales entre los dos tipos de titulados debidas a esa diferente condición. Para ello se establece el siguiente modelo de regresión lineal:

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

donde:

$Y_i =$ Salario del titulado i-ésimo

$$X_i = \begin{cases} 1, & \text{si el titulado } i \text{ es doctor} \\ 0, & \text{si el titulado } i \text{ es licenciado} \end{cases}$$

$u_i =$ Perturbación aleatoria que cumple los supuestos del modelo clásico de regresión lineal.

De acuerdo con ello, si se toman los valores esperados del salario para los distintos valores que puede adoptar la variable ficticia considerada, se tiene que:

 ISBN: 978-84-694-7251-4

$E[Y_i \mid X_i = 1] = \beta_1 + \beta_2 = \mu_1 \longrightarrow$ Salario medio poblacional que tienen los doctores.

$E[Y_i \mid X_i = 0] = \beta_1 = \mu_0 \longrightarrow$ Salario medio poblacional que tienen los licenciados.

Según las expresiones anteriores, se deduce que:

$\beta_1 = \mu_0 \longrightarrow$ Salario medio poblacional de los licenciados.

$\beta_2 = \mu_1 - \mu_0 \longrightarrow$ Diferencia en el salario medio poblacional entre los doctores y los licenciados.

Obsérvese que la categoría referente a los "licenciados" (aquélla en la que la variable ficticia toma el valor 0) es sobre la que se compara la otra categoría (en este caso, "doctores"). Los "licenciados" forman lo que se denomina la *categoría base* de la variable relativa al grado de los "titulados universitarios".

Teniendo en cuenta la interpretación dada de los coeficientes de regresión del modelo planteado, si se quisiera comprobar si las retribuciones entre ambos grados de titulados son iguales, se podría plantear un contraste de hipótesis[22], donde la hipótesis nula sería:

$$H_0 : \beta_2 = 0 .$$

Los coeficientes de regresión β_1 y β_2 se estiman por el método de MCO habitual: $\hat{\beta} = (X'X)^{-1}X'Y$, siendo las matrices de datos de las variables del modelo:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} \text{ y } X = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix},$$

donde la 2ª columna de la matriz *X*, relativa a los valores de la variable explicativa, está formada sólo por "ceros" y "unos", dependiendo de si la observación en cuestión se refiere a un licenciado o a un doctor, respectivamente.

Si tenemos en cuenta que el número de datos (n) es tal que $n = n_0 + n_1$, siendo n_0 el número de datos de *X* cuyo valor es 0 y n_1 el número de datos que son 1, y denotamos por Y_0 a la variable dependiente del modelo asociada a los valores de X que son 0 y por Y_1 a la que corresponde a los valores de X iguales a 1, entonces tendremos que:

[22] En el siguiente tema, trataremos los contrastes de hipótesis. No obstante, el alumno conoce ya estos conceptos, por anteriores materias de Estadística que ha estudiado previamente.

$$X'X = \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix} \qquad X'Y = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{2i} Y_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n_1} Y_i \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ n_1\overline{Y}_1 \end{pmatrix}.$$

Si se desarrolla la expresión de la estimación de β por MCO, $\hat{\beta} = (X'X)^{-1}X'Y$, finalmente se llega a:

$$\boxed{\hat{\beta} = \begin{pmatrix} \overline{Y}_0 \\ \overline{Y}_1 - \overline{Y}_0 \end{pmatrix}},$$

cuyo significado coincide, en términos muestrales, con el de los β poblacionales.

- ***Ejemplo de una variable cualitativa con tres modalidades***

Supóngase ahora que se quiere estudiar, a través de un modelo de regresión lineal, el "salario" de una población (Y_i) en función de su "nivel de estudios", distinguiéndose aquí tres posibles: "primarios", "secundarios" y "superiores".

Estas tres modalidades vienen expresadas por las siguientes variables ficticias:

$$D_{2i} = \begin{cases} 1, & \text{si el individuo } i \text{ tiene estudios secundarios} \\ 0, & \text{en caso contrario} \end{cases}$$

$$D_{3i} = \begin{cases} 1, & \text{si el individuo } i \text{ tiene estudios superiores} \\ 0, & \text{en caso contrario} \end{cases}$$

Obsérvese que no es necesario definir una tercera variable para el caso de los individuos con estudios primarios, ya que esta circunstancia es la que se da cuando D_{2i} y D_{3i} toman el valor 0 simultáneamente. La modalidad "estudios primarios" se considerará la categoría base de la variable "nivel de estudios".

De acuerdo con todo esto, el modelo que se especifica es:

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

$E[Y_i \mid D_{2i} = 1, D_{3i} = 0] = \beta_1 + \beta_2 = \mu_{SEC}$ ⟶ Salario medio de los individuos con estudios secundarios.

$E[Y_i \mid D_{2i} = 0, D_{3i} = 1] = \beta_1 + \beta_3 = \mu_{SUP}$ ⟶ Salario medio de los individuos con estudios superiores.

$E[Y_i \mid D_{2i} = 0, D_{3i} = 0] = \beta_1 = \mu_{PRI}$ ⟶ Salario medio de los individuos con estudios primarios.

 ISBN: 978-84-694-7251-4

A partir de aquí, puede deducirse fácilmente el significado de los coeficientes de regresión del modelo:

$\beta_1 = \mu_{PRI}$ ⟶ Salario medio de los individuos con estudios primarios.

$\beta_2 = \mu_{SEC} - \mu_{PRI}$ ⟶ Diferencia en el salario medio de los individuos con estudios secundarios respecto a los que tienen estudios primarios.

$\beta_3 = \mu_{SUP} - \mu_{PRI}$ ⟶ Diferencia en el salario medio de los individuos con estudios superiores respecto a los que tienen estudios primarios.

De nuevo se observa que las comparaciones de los salarios para las distintas modalidades de la variable "nivel de estudios" se hacen con respecto a la categoría base.

- ***Tipos de especificaciones en la construcción de modelos con variables ficticias***

A la hora de construir un modelo con variables ficticias, puede optarse por una de las dos especificaciones siguientes:

- *Aditiva*: $Y_i = \beta_1 + \beta_2 D_i + \beta_3 Z_i + u_i$
- *Multiplicativa*[23]: $\beta_1 + \beta_2 D_i + \beta_3 Z_i + \beta_4 D_i Z_i + u_i$

donde D_i y Z_i son, en este caso, variables ficticias en ambas especificaciones y se supone que la perturbación aleatoria u_i cumple los supuestos del modelo clásico de regresión lineal.

Para comprobar la diferencia entre ambas alternativas, nos serviremos del siguiente ejemplo.

Pensemos en un modelo de regresión lineal mediante el cual pretendemos explicar el "salario" de una población en función del "sexo" del individuo (cuyas dos posibles modalidades resultan ser: "varón" o "mujer") y su "zona geográfica de residencia" (donde se consideran también sólo dos posibles modalidades o categorías: "zona urbana" o "zona rural"). Se definen para ello las siguientes variables:

$Y_i =$ Salario del individuo i

$$D_i = \begin{cases} 1, & \text{si el individuo } i \text{ es mujer} \\ 0, & \text{si el individuo } i \text{ es varón} \end{cases}$$

$$Z_i = \begin{cases} 1, & \text{si el individuo } i \text{ reside en una zona rural} \\ 0, & \text{si el individuo } i \text{ reside en una zona urbana} \end{cases}$$

[23] Existen otras variantes de la especificación multiplicativa, como es el caso de:

$$\beta_1 + \beta_2 D_i + \beta_3 D_i Z_i + u_i .$$

Nótese que las *categorías base* para el caso del "sexo" (variable D_i) y la "zona geográfica de residencia" (variable Z_i) son "varón" y "zona urbana", respectivamente.

Especificación aditiva:

En el caso de considerar la especificación aditiva, el modelo es:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 Z_i + u_i .$$

Si se toman los valores esperados del salario para las distintas combinaciones de valores que pueden adoptar las dos variables ficticias introducidas en el modelo, se deduce de manera inmediata que:

$E[Y_i \mid D_i = 1, Z_i = 1] = \beta_1 + \beta_2 + \beta_3 = \mu_{M,R}$ ⟶ Salario medio de las mujeres que residen en una zona rural.

$E[Y_i \mid D_i = 1, Z_i = 0] = \beta_1 + \beta_2 = \mu_{M,U}$ ⟶ Salario medio de las mujeres que residen en una zona urbana.

$E[Y_i \mid D_i = 0, Z_i = 1] = \beta_1 + \beta_3 = \mu_{V,R}$ ⟶ Salario medio de los varones que residen en una zona rural.

$E[Y_i \mid D_i = 0, Z_i = 0] = \beta_1 = \mu_{V,U}$ ⟶ Salario medio de los varones que residen en una zona urbana.

Así, el significado de los distintos coeficientes de regresión resulta ser entonces:

$\beta_1 = \mu_{V,U}$ ⟶ Salario medio de los varones que residen en una zona urbana.

$\beta_2 = \mu_{M,U} - \mu_{V,U} = \mu_{M,R} - \mu_{V,R}$ ⟶ Diferencia en el salario medio de las mujeres respecto a los varones, dentro de cada zona de residencia.

$\beta_3 = \mu_{V,R} - \mu_{V,U} = \mu_{M,R} - \mu_{M,U}$ ⟶ Diferencia en el salario medio de los que residen en una zona rural respecto a los que lo hacen en una zona urbana, dentro de cada sexo.

Una vez más puede observarse cómo las comparaciones se hacen sobre las categorías base: "varón" y "zona urbana".

Especificación multiplicativa:

En este caso, el modelo que se especifica resulta ser:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 Z_i + \beta_4 D_i Z_i + u_i ,$$

 ISBN: 978-84-694-7251-4

Los valores esperados del salario, teniendo en cuenta los posibles valores de las variables ficticias consideradas, son en este modelo:

$E[Y_i \mid D_i = 1, Z_i = 1] = \beta_1 + \beta_2 + \beta_3 + \beta_4 = \mu_{M,R}$ ⟶ Salario medio de las mujeres que residen en una zona rural.

$E[Y_i \mid D_i = 1, Z_i = 0] = \beta_1 + \beta_2 = \mu_{M,U}$ ⟶ Salario medio de las mujeres que residen en una zona urbana.

$E[Y_i \mid D_i = 0, Z_i = 1] = \beta_1 + \beta_3 = \mu_{V,R}$ ⟶ Salario medio de los varones que residen en una zona rural.

$E[Y_i \mid D_i = 0, Z_i = 0] = \beta_1 = \mu_{V,U}$ ⟶ Salario medio de los varones que residen en una zona urbana.

A partir de las expresiones anteriores, se deduce el significado de los coeficientes de regresión del modelo:

$\beta_1 = \mu_{V,U}$ ⟶ Salario medio de los varones que residen en una zona urbana.

$\beta_2 = \mu_{M,U} - \mu_{V,U}$ ⟶ Diferencia en el salario medio, en las zonas urbanas, de las mujeres respecto a los varones.

$\beta_3 = \mu_{V,R} - \mu_{V,U}$ ⟶ Diferencia en el salario medio, en los varones, de los que residen en una zona rural respecto a los de una zona urbana.

Obsérvese, por su parte, que las otras dos diferencias posibles, distinguiendo sexo y zona geográfica de residencia, vienen dadas por combinaciones de coeficientes de regresión y no por coeficientes aislados:

$\beta_2 + \beta_4 = \mu_{M,R} - \mu_{V,R}$ ⟶ Diferencia en el salario medio, en las zonas rurales, de las mujeres respecto a los varones.

$\beta_3 + \beta_4 = \mu_{M,R} - \mu_{M,U}$ ⟶ Diferencia en el salario medio, en las mujeres, de las que residen en una zona rural respecto a las de una zona urbana.

El significado de β_4 se puede ver a partir de las relaciones anteriores:

β_4 ⟶ Compara la diferencia media de salarios entre mujeres y varones de zona rural a zona urbana ($(\beta_2 + \beta_4) - \beta_2$); asimismo ofrece la comparación de la diferencia media de salarios entre zona rural y zona urbana de mujeres a varones ($(\beta_3 + \beta_4) - \beta_3$).

 ISBN: 978-84-694-7251-4

La principal aportación del esquema multiplicativo frente al aditivo es que permite tener en cuenta la interacción entre variables. En concreto, en el ejemplo que nos ocupa, se pueden apreciar distintas diferencias de salario entre mujeres y varones según cual sea su zona de residencia, mientras que con el esquema aditivo la diferencia salarial entre varones y mujeres resulta ser la misma tanto en la zona rural como en la zona urbana. Del mismo modo, la especificación multiplicativa hace posible la percepción de distintas diferencias de salario entre zona rural y zona urbana según se trate de mujeres o varones, hecho que no ocurre con el esquema aditivo.

- ***La trampa de las variables ficticias***

A lo largo de todos los ejemplos que se han ido exponiendo en este tema, obsérvese que cuando las variables cualitativas incluidas en los modelos tenían *m* modalidades, se han utilizado *m*-1 variables ficticias para su consideración. La razón de ello, en lugar de utilizar *m* variables ficticias, no sólo reside en el hecho de que no resulta necesario, sino en evitar la presencia de multicolinealidad perfecta entre los términos independientes y las variables ficticias en los correspondientes modelos.

Si, por ejemplo, tuviésemos una variable con tres modalidades y empleásemos tres variables ficticias (D_1, D_2, D_3) en el modelo:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + u_i ,$$

tendríamos que: $D_1 + D_2 + D_3 = 1$; esto es, la suma de estas tres variables sería igual a la ordenada en el origen: X_1 . Por tanto, estaríamos ante una situación de multicolinealidad perfecta, que nos impediría calcular de forma unívoca la estimación de los parámetros.

Para evitarla, caben dos soluciones alternativas:

a) incluir sólo $m-1$ variables ficticias en el modelo; o bien,

b) incluir m variables ficticias, pero en un modelo sin ordenada en el origen.

- ***¿Por qué valores 0 y 1?***

Otro aspecto importante a tener en cuenta en la construcción de variables ficticias es el hecho de que no sería correcto emplear valores distintos a 0 y 1 en ellas, tal como las hemos estudiado. El porqué de ello lo ilustraremos mediante el siguiente ejemplo.

Considérese el modelo: $Y_i = \beta_1 + \beta_2 Z_i + u_i$, donde:

$Y_t =$ Salario anual del individuo *i*

$$Z_i = \begin{cases} 0, & \text{si el individuo } i \text{ tiene un nivel de cualificación "A"} \\ 1, & \text{si el individuo } i \text{ tiene un nivel de cualificación "B"} \\ 2, & \text{si el individuo } i \text{ tiene un nivel de cualificación "C"} \end{cases}$$

 ISBN: 978-84-694-7251-4

$E[Y_i \mid Z_i = 0] = \beta_1 \equiv \mu_A$ ⟶ Salario anual medio de los individuos con un nivel de cualificación "A".

$E[Y_i \mid Z_i = 1] = \beta_1 + \beta_2 \equiv \mu_B$ ⟶ Salario anual medio de los individuos con un nivel de cualificación "B".

$E[Y_i \mid Z_i = 2] = \beta_1 + 2\beta_2 \equiv \mu_C$ ⟶ Salario anual medio de los individuos con un nivel de cualificación "C".

Nótese que con esta especificación se está asumiendo que:

$$\mu_C - \mu_B = \mu_B - \mu_A = \beta_2 \quad \text{y} \quad \mu_C - \mu_A = 2\beta_2 .$$

Es decir, que la diferencia en los salarios entre los individuos con un nivel de cualificación "A" respecto de los de "B" es igual que la diferencia de los de un nivel "B" respecto de los de "C" y que, por tanto, entre "A" y "C" esta diferencia es el doble. Todo ello, evidentemente, no tiene por qué ser así. Ésta es la razón, pues, por la que no deben darse a las variables ficticias valores distintos a 0 y 1.

- ***Ejemplos ilustrativos***

Consideremos un modelo de regresión lineal que explica los ingresos netos familiares en función de diversas características de la persona principal del hogar: edad, estado civil y nivel máximo de estudios alcanzado. En concreto, las variables del modelo son:

ING = Ingresos netos familiares (en €)

EDAD = Edad de la persona principal del hogar (en años)

$$\text{CASADO} = \begin{cases} 1, & \text{si la persona principal del hogar está casada} \\ 0, & \text{en caso contrario} \end{cases}$$

$$\text{SECUNDAR} = \begin{cases} 1, & \text{si los máximos estudios de la persona ppal. del hogar son secundarios} \\ 0, & \text{en caso contrario} \end{cases}$$

$$\text{SUPERIOR} = \begin{cases} 1, & \text{si los máximos estudios de la persona principal del hogar son superiores} \\ 0, & \text{en caso contrario} \end{cases}$$

En el nivel de estudios, se considera un tercer nivel: el PRIMARIO, que constituye la categoría base (esto es, cuando SECUNDAR y SUPERIOR toman a la vez valores 0).

Un aspecto importante que cabe señalar es el hecho de que la asignación de los valores 0 y 1, así como la elección de la categoría o modalidad básica, resultan arbitrarias.

Únicamente habrá que estar atentos a la interpretación de los signos y valores de los correspondientes coeficientes de regresión.

A partir de una muestra de 3.000 hogares españoles con datos relativos a 1998, se han planteado dos modelos distintos: según un esquema aditivo y según un esquema multiplicativo. Se muestran a continuación los resultados obtenidos en ambos casos.

Especificación aditiva:

De acuerdo con la definición de las variables realizada, el modelo a estimar es:

$$\boxed{\text{ING} = \beta_1 + \beta_2 \text{EDAD} + \beta_3 \text{CASADO} + \beta_4 \text{SECUNDAR} + \beta_5 \text{SUPERIOR} + u}$$

El resultado de la estimación por MCO de este modelo resulta ser:

$$\widehat{\text{ING}} = 6.030{,}49 + 199{,}90*\text{EDAD} + 1.205{,}63*\text{CASADO} + 2.897{,}10*\text{SECUNDAR} + 13.181{,}26*\text{SUPERIOR}$$

El significado de los coeficientes de regresión estimados sería:

- $\hat{\beta}_1$ - Su significado es el habitual de todo modelo de regresión lineal. Es la ordenada en el origen y, a veces, no tiene sentido económico. En este caso, sin embargo, sí lo tiene; 6.030,49 € serían los ingresos netos familiares que, en promedio, tendrían como mínimo aquellos hogares cuya persona principal no está casada y tiene estudios primarios, sin tener en cuenta el efecto de la edad.
- $\hat{\beta}_2$ - Es el efecto marginal medio de la edad sobre los ingresos netos familiares. Cada año de edad de la persona principal de hogar supondría, por término medio, un incremento de 199,90 € en los ingresos netos familiares.
- $\hat{\beta}_3$ - Este parámetro recoge el efecto del estado civil en nuestro modelo. En este caso viene a significar que estar casado supone, por término medio, unos ingresos netos adicionales en el hogar de 1.205,63 €, frente a otros estados civiles.
- $\hat{\beta}_4$ - El valor de este coeficiente muestra el efecto diferencial de las personas principales del hogar que tienen un nivel máximo de estudios secundarios frente a los de estudios primarios. Como parecía lógico esperar, el signo de este parámetro es positivo y, en concreto, su valor indica que los ingresos netos de estos hogares son, por término medio, 2.897,10 € más altos que los de la categoría base.
- $\hat{\beta}_5$ - El coeficiente que acompaña a los estudios superiores muestra, por su parte, cuál es la influencia que éstos tienen sobre los ingresos netos, en comparación con los correspondientes a los estudios primarios. Al igual que en el caso anterior, de nuevo el signo es positivo, reflejando su valor que los ingresos netos de los hogares

cuya persona principal alcanza los estudios superiores son 13.181,26 € más elevados que los de la categoría base, por término medio.

Si se quisiera comparar la diferencia, en sus efectos sobre los ingresos netos familiares, entre los hogares cuya personal principal tiene estudios superiores y los de estudios secundarios, bastaría con ver la diferencia entre los valores de los respectivos coeficientes de regresión; es decir: 13.181,26 – 2.897,10 = 10.284,16 €.

Obsérvese que, aplicando el significado de las variables ficticias, llegamos a distintas ecuaciones según sea el perfil de la persona principal del hogar:

- Persona principal no casada y con estudios primarios:

$$\widehat{\text{ING}} = 6.030,49 + 199,90*\text{EDAD}$$

- Persona principal no casada y con estudios secundarios:

$$\widehat{\text{ING}} = 6.030,49 + 2.897,10 + 199,90*\text{EDAD} = 8.927,59 + 199,90*\text{EDAD}$$

- Persona principal no casada y con estudios superiores:

$$\widehat{\text{ING}} = 6.030,49 + 13.181,26 + 199,90*\text{EDAD} = 19.211,75 + 199,90*\text{EDAD}$$

- Persona principal casada y con estudios primarios:

$$\widehat{\text{ING}} = 6.030,49 + 1.205,63 + 199,90*\text{EDAD} = 7.236,12 + 199,90*\text{EDAD}$$

- Persona principal casada y con estudios secundarios:

$$\widehat{\text{ING}} = 6.030,49 + 1.205,63 + 2.897,10 + 199,90*\text{EDAD} = 10.133,22 + 199,90*\text{EDAD}$$

- Persona principal casada y con estudios superiores:

$$\widehat{\text{ING}} = 6.030,49 + 1.205,63 + 13.181,26 + 199,90*\text{EDAD} = 20.417,38 + 199,90*\text{EDAD}$$

Especificación multiplicativa:

Planteamos ahora un modelo alternativo que permitirá distinguir posibles diferencias en el efecto marginal medio de la edad sobre los ingresos netos familiares, dependiendo del nivel máximo de estudios alcanzado por la persona principal del hogar:

$$\text{ING} = \beta_1 + \beta_2\text{EDAD} + \beta_3\text{CASADO} + \beta_4\text{SECUNDAR} + \beta_5\text{SUPERIOR} + \\ + \beta_6\text{EDAD}*\text{SECUNDAR} + \beta_7\text{EDAD}*\text{SUPERIOR} + u$$

Los ingresos estimados vienen dados ahora por:

$$\widehat{\text{ING}} = 11.778,99 + 73,06*\text{EDAD} + 1.000,14*\text{CASADO} + 4.398,71*\text{SECUNDAR} + \\ 7.333,88*\text{SUPERIOR} + 170,00*\text{EDAD}*\text{SECUNDAR} + 478,99*\text{EDAD}*\text{SUPERIOR}$$

En cuanto al significado de los coeficientes de regresión estimados, resulta que:

- $\hat{\beta}_1$ - 11.778,99 € serían los ingresos familiares netos que, en promedio, tendrían como mínimo aquellos hogares cuya persona principal no está casada y tiene estudios primarios, sin tener en cuenta el efecto de la edad.
- $\hat{\beta}_2$ - Es el efecto marginal medio de la edad sobre los ingresos netos familiares en el caso en que la persona principal del hogar tiene estudios primarios. Cada año de edad de dicha persona supondría, por término medio, un incremento de 73,06 € en los ingresos netos familiares.
- $\hat{\beta}_3$ - Los ingresos netos familiares de los hogares cuya persona principal está casada superan en 1.000,14 €, por término medio, a los ingresos netos familiares de los hogares en los que el estado civil de la persona principal es otro (estando en igualdad de condiciones para el resto de variables).
- $\hat{\beta}_4$ - Los ingresos netos familiares mínimos de los hogares cuya persona principal tiene estudios secundarios serán, en promedio y sin tener en cuenta el efecto de la edad, de 4.398,71 € más que los correspondientes cuando la persona principal tiene estudios primarios.
- $\hat{\beta}_5$ - Los ingresos netos familiares mínimos de los hogares cuya persona principal tiene estudios superiores serán, en promedio y sin tener en cuenta el efecto de la edad, de 7.333,88 € más que los correspondientes cuando la persona principal tiene estudios primarios.
- $\hat{\beta}_6$ - Este coeficiente muestra la diferencia en el efecto marginal medio de la edad sobre los ingresos familiares netos de las personas principales del hogar que tienen un nivel máximo de estudios secundarios frente a los que tienen estudios primarios. En el caso de estudios secundarios, cada año de edad supone 170,00 € más de ingresos, por término medio, que con estudios primarios.
- $\hat{\beta}_7$ - El significado de este coeficiente es similar al de $\hat{\beta}_6$, referido a los estudios superiores. Es decir, si la persona principal del hogar tiene estudios superiores, cada año de edad supondrá un incremento de 478,99 € más en los ingresos que si tuviera únicamente estudios primarios, por término medio.

Para comparar los efectos marginales medios de la edad sobre los ingresos netos de los hogares cuyas personas principales tienen estudios superiores respecto a las que tienen estudios secundarios, de nuevo sería suficiente con calcular la diferencia entre los valores de los respectivos coeficientes de regresión; es decir: 478,99 – 170,00 = 308,99 €.

Igualmente se podrían comparar los ingresos mínimos de los hogares, en término medio y sin tener en cuenta el efecto de la edad, cuando la personal principal tiene estudios

 ISBN: 978-84-694-7251-4

superiores frente a estudios secundarios. En este caso, dichos ingresos se diferenciarían en 7.333,88 – 4.398,71 = 2.935,17 €, esto es, serían 2.935,17 € más cuando se tienen estudios superiores.

Los distintos perfiles de las personas principales de los hogares dan lugar a las siguientes ecuaciones para la estimación de los ingresos netos familiares:

- Persona principal no casada y con estudios primarios:

$$\widehat{ING} = 11.778,99 + 73,06*EDAD$$

- Persona principal no casada y con estudios secundarios:

$$\widehat{ING} = 11.778,99 + 4.398,71 + 73,06*EDAD + 170,00*EDAD = 16.177,70 + 243,06*EDAD$$

- Persona principal no casada y con estudios superiores:

$$\widehat{ING} = 11.778,99 + 7.333,88 + 73,06*EDAD + 478,99*EDAD= 19.112,87 + 552,05*EDAD$$

- Persona principal casada y con estudios primarios:

$$\widehat{ING} = 11.778,99 + 1.000,14 + 73,06*EDAD = 12.779,13 + 73,06*EDAD$$

- Persona principal casada y con estudios secundarios:

$$\widehat{ING} = 11.778,99 + 1.000,14 + 4.398,71 + 73,06*EDAD + 170,00*EDAD = 17.177,84 + 243,06*EDAD$$

- Persona principal casada y con estudios superiores:

$$\widehat{ING} = 11.778,99 + 1.000,14 + 7.333,88 + 73,06*EDAD + 478,99*EDAD= 20.113,01 + 552,05*EDAD$$

En las ecuaciones anteriores, puede observarse que el efecto marginal de la edad sobre los ingresos netos familiares varía dependiendo del nivel de estudios de la persona principal del hogar. Esto no ocurría con la especificación aditiva, sino que dicho efecto marginal era entonces el mismo en todos los casos.

2.6. Formas funcionales linealizables. Elasticidad *vs.* efecto marginal. Comparación entre modelos.-

En este apartado, se van a considerar diversas formas funcionales alternativas a la lineal. Todas ellas se caracterizan por el hecho de que, pese a no ser lineales, a partir de sencillas transformaciones matemáticas se convierten en tales; por esta razón, reciben el nombre de formas funcionales linealizables.

Vamos a ver, para cada uno de los modelos que consideramos, cuál es el significado de sus coeficientes de regresión, así como la forma en que se obtienen el efecto marginal y la elasticidad. Por último, nos fijaremos en el modo en el que debemos actuar para poder establecer comparaciones entre los distintos modelos, a fin de elegir el mejor.

Por simplicidad en la explicación, se van a considerar en todos los casos modelos de regresión simple (con ordenada en el origen), donde Y es la variable cuyo comportamiento se trata de explicar según los valores de la variable explicativa X.

Elasticidad *vs.* efecto marginal

Comenzamos exponiendo los conceptos de "efecto marginal" y de "elasticidad":

- $\boxed{Efecto\,marginal \equiv \frac{dY}{dX}}$. Expresa la relación entre variaciones absolutas de Y y de X.
- $\boxed{Elasticidad \equiv \frac{dY/Y}{dX/X} = \frac{dY}{dX}\cdot\frac{X}{Y}}$. Expresa la relación entre variaciones relativas de Y y de X.

Análisis de los modelos

Antes de abordar el estudio de las diferentes formas funcionales linealizables, vamos a tratar el modelo lineal brevemente para poder contrastar sus resultados frente a los de los demás modelos.

- ***Modelo lineal***

Como bien sabemos ya, la expresión de este modelo es:

$$\boxed{Y_i = \beta_1 + \beta_2 X_i + u_i}, \text{ donde } i = 1,\ldots,n.$$

- El coeficiente de regresión β_2 expresa: $\boxed{\beta_2 = \frac{dY}{dX}}$.

  Su interpretación puede verse con el ejemplo: $\widehat{CONSUMO} = 3{,}28 + 0{,}37\,RENTA$.

  Si ambas variables viniesen expresadas en euros, un incremento de renta de 1 € supondría un incremento medio en el consumo de 0,37 €.

- $\boxed{Efecto\,marginal \equiv \beta_2}$.

  Obsérvese que este valor es constante.

- $\boxed{\text{Elasticidad} \equiv \frac{dY/Y}{dX/X} = \frac{dY}{dX}\cdot\frac{X}{Y} = \beta_2\cdot\frac{X}{Y}}$.

  El valor de la elasticidad en el modelo lineal es variable, dependiendo de los valores del par de observaciones concreto de X e Y que se considere. Por este motivo, se suele hablar de la elasticidad media, que no es sino el valor de la elasticidad correspondiente a los valores medios de X e Y.

 ISBN: 978-84-694-7251-4

- ***Modelo log-log o log-lineal***

Este modelo puede proceder originariamente de un modelo potencial. Así, si se parte de éste: $Y_i = \alpha_1 \cdot X_i^{\alpha_2} \cdot e^{u_i}$, tomando logaritmos neperianos podemos llegar a: $\ln Y_i = \ln \alpha_1 + \alpha_2 \ln X_i + u_i$. Si hacemos que: $\ln \alpha_1 = \beta_1$ y $\alpha_2 = \beta_2$, entonces tendremos finalmente la expresión del modelo log-log, o también llamado log-lineal, o incluso doble log:

$$\boxed{\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i}, \text{ donde } i = 1, \ldots, n.$$

- El coeficiente de regresión β_2 expresa: $\boxed{\beta_2 = \dfrac{d \ln Y}{d \ln X} = \dfrac{dY \cdot \dfrac{1}{Y}}{dX \cdot \dfrac{1}{X}} = \dfrac{dY/Y}{dX/X}}$.

  Su interpretación podría verse con: $\widehat{\ln CONSUMO} = 4{,}37 + 0{,}1582 \ln RENTA$.

  Un incremento del 1% en la renta supone un incremento medio en el consumo del 0,1582%.

- $\boxed{Elasticidad \equiv \beta_2}$.

  Obsérvese que este valor es constante.

- $\boxed{Efecto\ marginal \equiv \dfrac{dY}{dX} = \dfrac{dY/Y}{dX/X} \cdot \dfrac{Y}{X} = \beta_2 \cdot \dfrac{Y}{X}}$.

  El valor del efecto marginal en este modelo es variable, dependiendo de los valores del par de observaciones concreto de X e Y que se considere. Por este motivo, se suele hablar del efecto marginal medio, que es el que corresponde a los valores medios de X e Y.

- ***Modelos semi-logarítmicos***

Dentro de éstos, pueden considerarse dos tipos: log-lin y lin-log.

- **Modelo log-lin.**

  Este modelo puede proceder en su origen de un modelo exponencial. De este modo, partiendo de éste: $Y_i = \alpha_1 \cdot \alpha_2^{X_i} \cdot e^{u_i}$, tomando logaritmos se puede llegar a: $\ln Y_i = \ln \alpha_1 + \ln \alpha_2 X_i + u_i$. Si hacemos que: $\ln \alpha_1 = \beta_1$ y $\ln \alpha_2 = \beta_2$, entonces se llegará finalmente a la expresión del modelo log-lin:

$$\boxed{\ln Y_i = \beta_1 + \beta_2 X_i + u_i}, \text{ donde } i = 1, \ldots, n.$$

 ISBN: 978-84-694-7251-4

- Su <u>coeficiente de regresión</u> β_2 expresa: $\boxed{\beta_2 = \frac{d \ln Y}{dX} = \frac{dY \cdot \frac{1}{Y}}{dX} = \frac{dY/Y}{dX}}$.

  Su interpretación puede verse mediante: $\widehat{\ln CONSUMO} = 8{,}23 + 0{,}0023 RENTA$.

  Un incremento de renta de 1€ supone un incremento medio en el consumo del 0,23%.

- $\boxed{Efecto\, marginal \equiv \frac{dY}{dX} = \frac{dY/Y}{dX} \cdot Y = \beta_2 \cdot Y}$.

  El valor del efecto marginal en este modelo depende del valor concreto de Y que se considere. Por ello, suele hablarse del <u>efecto marginal medio</u>, que es el que corresponde con el valor medio de dicha variable.

- $\boxed{Elasticidad \equiv \frac{dY/Y}{dX/X} = \frac{dY/Y}{dX} \cdot X = \beta_2 \cdot X}$.

  El valor de la elasticidad también es variable, dependiendo del valor concreto que se tome de X. Debido a ello, suele hablarse de la <u>elasticidad media</u>, correspondiente al valor medio de X.

- **Modelo lin-log.**

  La expresión de este modelo es: $\boxed{Y_i = \beta_1 + \beta_2 \ln X_i + u_i}$, donde $i = 1, \ldots, n$.

  - El <u>coeficiente de regresión</u> β_2 expresa aquí: $\boxed{\beta_2 = \frac{dY}{d \ln X} = \frac{dY}{dX \cdot \frac{1}{X}} = \frac{dY}{dX/X}}$.

    Sobre su interpretación, considérese: $\widehat{CONSUMO} = 1{,}23 + 348{,}27 \ln RENTA$.

    Un incremento de renta del 1% supone un incremento medio de 3,4827 € en el consumo.

  - $\boxed{Efecto\, marginal \equiv \frac{dY}{dX} = \frac{dY}{dX/X} \cdot \frac{1}{X} = \beta_2 \cdot \frac{1}{X}}$.

    El valor del efecto marginal en este modelo depende del valor concreto de X que se tome. Debido a esto, suele hablarse del <u>efecto marginal medio</u>, que es el que corresponde con el valor medio de X.

  - $\boxed{\text{Elasticidad} \equiv \frac{dY/Y}{dX/X} = \frac{dY}{dX/X} \cdot \frac{1}{Y} = \beta_2 \cdot \frac{1}{Y}}$.

 ISBN: 978-84-694-7251-4

El valor de la elasticidad depende del valor concreto de Y que se considere. Por esta razón, se suele hablar de la elasticidad media, que es la correspondiente al valor medio de dicha variable.

- ***Modelo recíproco***

Este modelo viene expresado a través de la forma funcional:

$$\boxed{Y_i = \beta_1 + \beta_2\left(\frac{1}{X_i}\right) + u_i}, \text{ donde } i = 1,\ldots,n\,.$$

- Su coeficiente de regresión nos da: $\boxed{\beta_2 = \dfrac{dY}{d\left(\dfrac{1}{X}\right)} = \dfrac{dY}{-dX/X^2}}$.

La interpretación de β_2 no resulta inmediata en este modelo, por lo que la obviaremos.

- $\boxed{Efecto\ marginal \equiv \dfrac{dY}{dX} = \dfrac{dY}{-dX/X^2}\cdot\left(\dfrac{-1}{X^2}\right) = \beta_2\cdot\left(\dfrac{-1}{X^2}\right)}$.

El valor del efecto marginal en este caso depende del valor concreto que adopte X. Por ello, suele considerarse el efecto marginal medio, que es el relativo al valor medio de dicha variable.

- $\boxed{Elasticidad \equiv \dfrac{dY/Y}{dX/X} = \dfrac{dY}{-dX/X^2}\cdot\left(\dfrac{-1}{XY}\right) = \beta_2\cdot\left(\dfrac{-1}{XY}\right)}$.

El valor de la elasticidad en el modelo recíproco también es variable, dependiendo de los valores del par de observaciones concreto de X e Y que se considere. Por este motivo, se suele hablar de la elasticidad media, que es la que corresponde a los valores medios de X e Y.

Comparación entre modelos

Para poder comparar modelos alternativos y elegir qué opción resulta mejor, debemos fijarnos en dos aspectos esenciales:

- El número de variables explicativas.
- La unidad de medida de la variable dependiente.

Los indicadores más habituales para la comparación de modelos son:

1. El coeficiente de determinación: R^2, que evalúa la bondad del ajuste de un modelo, siendo el ajuste tanto mejor cuanto más cercano a 1 sea su valor.

2. El coeficiente de determinación corregido o ajustado, R_c^2, que evalúa de igual modo la bondad del ajuste de un modelo, pero teniendo en cuenta los grados de libertad de éste, y que señala también que el ajuste es tanto mejor cuanto más cercano a 1 sea su valor.
3. El criterio de información de Akaike: *AIC*, que expresa que el ajuste de un modelo es tanto mejor cuanto menor sea el valor de este estadístico.

Ya sabemos que cuando se consideran modelos donde la variable dependiente es la misma y tienen igual número de variables explicativas, puede emplearse el coeficiente de determinación R^2 para su comparación y dilucidar de este modo qué modelo es el que presenta un mejor ajuste.

Por su parte, en el caso de modelos anidados, considerando la misma variable dependiente, debe utilizarse el coeficiente de determinación corregido R_c^2. Como es ya sabido, este coeficiente penaliza el aumento de la capacidad explicativa del modelo (resultante de añadir nuevas variables) por el hecho de la pérdida de grados de libertad en el mismo.

La comparación de modelos lineales y logarítmicos no puede realizarse, en cambio, de forma directa, puesto que la unidad de medida de la variable dependiente es distinta: en un caso se trata de Y, mientras que en el otro se refiere a $\ln Y$. Una forma de comparar los resultados del ajuste de modelos lineales y no lineales es a través del *AIC*.

En concreto habría que comparar el valor del *AIC* del modelo en el que Y fuese lineal, con el de un AIC^* transformado para el caso del modelo en el que se considerase el $\ln Y$. Este AIC^* se obtendría del siguiente modo:

$$\boxed{AIC^* = AIC + 2 \cdot \overline{\ln Y}}$$

- ***Ejemplo ilustrativo***

La siguiente tabla muestra los valores de los indicadores comentados para la relación entre las variables *CONSUMO* y *RENTA*, según las distintas formas funcionales consideradas:

Modelo	R^2	R_c^2	*AIC*
Lineal	0,458367	0,444113	5,2374
Log-log	0,518989	0,506331	0,2364
Lin-log	0,456735	0,442439	5,2404
Log-lin	0,497778	0,484561	0,2794
Recíproco	0,428070	0,413019	5,2918

 ISBN: 978-84-694-7251-4

En este ejemplo el número de variables explicativas es siempre el mismo en los diferentes modelos planteados; sin embargo, la variable dependiente en ocasiones es *CONSUMO* y en otras es ln *CONSUMO*.

Teniendo en cuenta esto, se pueden comparar directamente entre sí, mediante el coeficiente de determinación R^2, los modelos lineal, lin-log y recíproco (cuya variable explicada es en todos ellos *CONSUMO*). De ellos, el mejor ajuste correspondería al modelo lineal: 0,458367.

De igual forma, también se podrían comparar entre sí los modelos log-log y log-lin (ya que en éstos la variable explicada resulta ser ln *CONSUMO*). En este caso, el valor más alto de R^2 se registra en el modelo log-log: 0,518989.

Para poder, seguidamente, dilucidar qué modelo presenta en nuestro ejemplo un mejor ajuste: el lineal o el log-log, deberíamos hacer lo siguiente (sabiendo que el valor medio[24] de la variable ln *CONSUMO* es 2,3719):

$$AIC^* = 0{,}2364 + (2 \,.\, 2{,}3719) = 4{,}9802.$$

Una vez calculado AIC^*, se puede comparar directamente con el AIC del modelo lineal. En este caso, como 5,2374 (mod. lineal) > 4,9802 (mod. log-log), entonces el mejor ajuste resulta ser el del modelo log-log, ya que presenta un valor menor.

2.7. Introducción al uso de EViews (I).-

En este apartado comenzamos nuestro aprendizaje del programa informático *Econometric Views* (más conocido abreviadamente como *EViews*). En concreto, nos vamos a centrar en su versión 3.1.

Nuestra primera aproximación a este *software* de extendido uso en el ámbito econométrico, se va a estructurar en los siguientes puntos:

- Acerca de *EViews*
- Ficheros de trabajo (*Workfiles*)
- Series de datos: introducción e importación
- Análisis de una serie de datos
- Estimación de un modelo lineal por MCO
- Ejemplo de estimación de un modelo no lineal por MCO: el modelo log-log
- Cómo guardar un archivo de trabajo

[24] Al igual que los coeficientes mostrados en la tabla de este ejemplo, este valor ha sido calculado de forma externa a partir de los datos originales que se han empleado para el mismo.

Acerca de *EViews*

Al iniciar una sesión de *EViews*, la primera imagen que aparece es la pantalla que se muestra en la *Figura 6*, donde se recogen diversas informaciones.

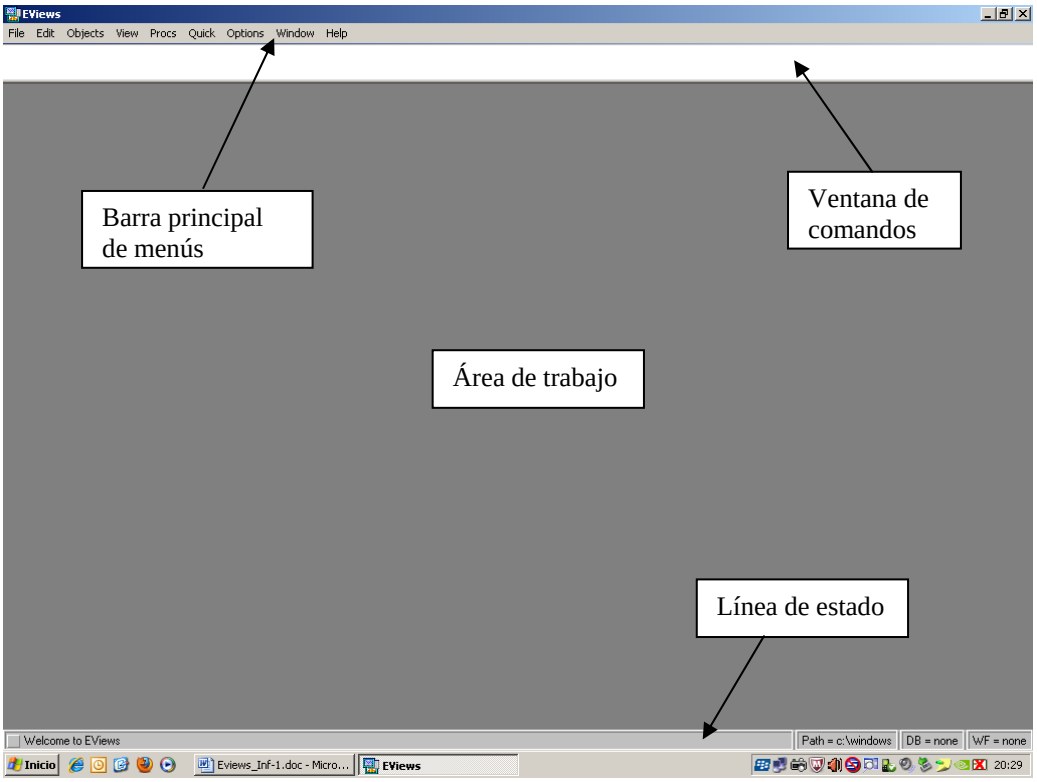


Figura 6

En la parte superior se nos muestra una barra de color azul con el título del programa y a continuación la *barra principal de menús*. La franja blanca que aparece bajo la barra principal de menús se denomina *ventana de comandos* y permite trabajar en *EViews* introduciendo directamente los comandos necesarios, ejecutándolos posteriormente con la tecla *Enter*.

En la parte inferior de la pantalla, se encuentra la llamada *línea de estado* donde se especifica, por ejemplo, el directorio donde por defecto se guardará el archivo en curso o en el caso de haber iniciado la sesión, si tenemos o no un fichero de trabajo en memoria (*WF = none*).

La zona amplia de color gris es el *área de trabajo* donde *EViews* irá desplegando las ventanas que vayamos utilizando a medida que sea necesario en el transcurso de la sesión.

- ***Elementos de la barra principal de menús (Menú principal de EViews)***

✓ **File;** Incluye las operaciones usuales relacionadas con ficheros, datos y programas dentro de un entorno Windows: abrir y cerrar ficheros de trabajo nuevos y ya guardados anteriormente, guardar ficheros de trabajo, imprimirlos, importar ficheros

desde una hoja de cálculo o base de datos o exportarlos hacia una hoja de cálculo o base de datos, salir del programa *EViews*, etc.

- ✓ **Edit;** Contiene las operaciones básicas (cortar, copiar, pegar, deshacer, buscar, reemplazar, etc.) de cualquier programa en entorno Windows.
- ✓ **Objects;** Contiene las funciones para manejar los distintos objetos que se almacenan en un fichero de trabajo: borrar, nombrar, imprimir, importar, etc.
- ✓ **View;** Para habilitar este menú desplegable es necesario antes abrir un fichero de trabajo (o *Workfile*). En función del tipo de ventana activa en cada caso, se obtendrán diferentes aspectos relacionados con la visualización en pantalla.
- ✓ **Procs;** Contiene las operaciones relacionadas con series de datos principalmente. De nuevo en este caso y al igual que en la opción VIEW es necesario tener un fichero de trabajo abierto para habilitar el menú desplegable, el cual será distinto según la ventana activa que utilicemos. En este menú podremos seleccionar una muestra de los datos, ordenarlos, generar nuevas series a partir de otras ya existentes, importar y exportar series de datos, así como cambiar el rango poblacional de las series.
- ✓ **Quick;** Proporciona acceso directo a los comandos que se utilizan con mayor frecuencia: generar series a partir de otras ya existentes, seleccionar una muestra, representar gráficamente las series de datos, editar las series, estimar modelos de regresión por MCO, representar las series a través de histogramas y sus estadísticos más representativos (media, mediana, curtosis, etc.), hallar las matrices de covarianzas y de correlaciones en un modelo de regresión, aplicar diversos métodos de tratamiento de series temporales (alisado exponencial, test de raíces unitarias, correlogramas, test de causalidad de *Granger*, test de cointegración, etc.) y estimar modelos VAR, entre otros. Al igual que en las opciones VIEW y PROCS, es necesario tener un fichero de trabajo abierto para habilitar el menú desplegable, el cual será distinto según la ventana activa que utilicemos.
- ✓ **Options;** Contiene los parámetros de funcionamiento general de *EViews*. Por ejemplo, el tamaño y las fuentes de las ventanas de resultados que obtendremos al estimar modelos (*Window and Font Options*), el comando que permite a *EViews* guardar la última versión actualizada de nuestro fichero de trabajo (*Backup files*), el número de iteraciones y grado de convergencia en procesos de estimación iterativos tales como el método de *Cochrane-Orcutt* en la autocorrelación (*Estimation Defaults*), tamaño, colores y fuentes de los gráficos (*Graphic Defaults*), etc.
- ✓ **Window;** Proporciona acceso directo a las distintas ventanas que tengamos abiertas durante la sesión de trabajo.
- ✓ **Help;** Es el menú de ayuda usual de un entorno Windows. Se organiza de acuerdo con varias opciones: referencias a objetos, comandos, funciones, matrices y programación. Además, en cada una de ellas se puede hacer uso de la ayuda en

función de un índice, una página de contenido y una opción de búsqueda, constituyendo una completa base de consulta de los más variados métodos econométricos que *EViews* es capaz de aplicar.

Ficheros de trabajo (*Workfiles*)

El elemento básico de *EViews* es el fichero de trabajo o *WORKFILE*; por ello, el primer paso antes de empezar a trabajar con este programa es la creación de un fichero de trabajo. Para crear un fichero de trabajo seleccionamos dentro del menú principal:

FILE / NEW / WORKFILE

De este modo nos aparecerá en pantalla (*Figura 7*) un menú llamado *Rango del Fichero de Trabajo* (*Workfile Range*).

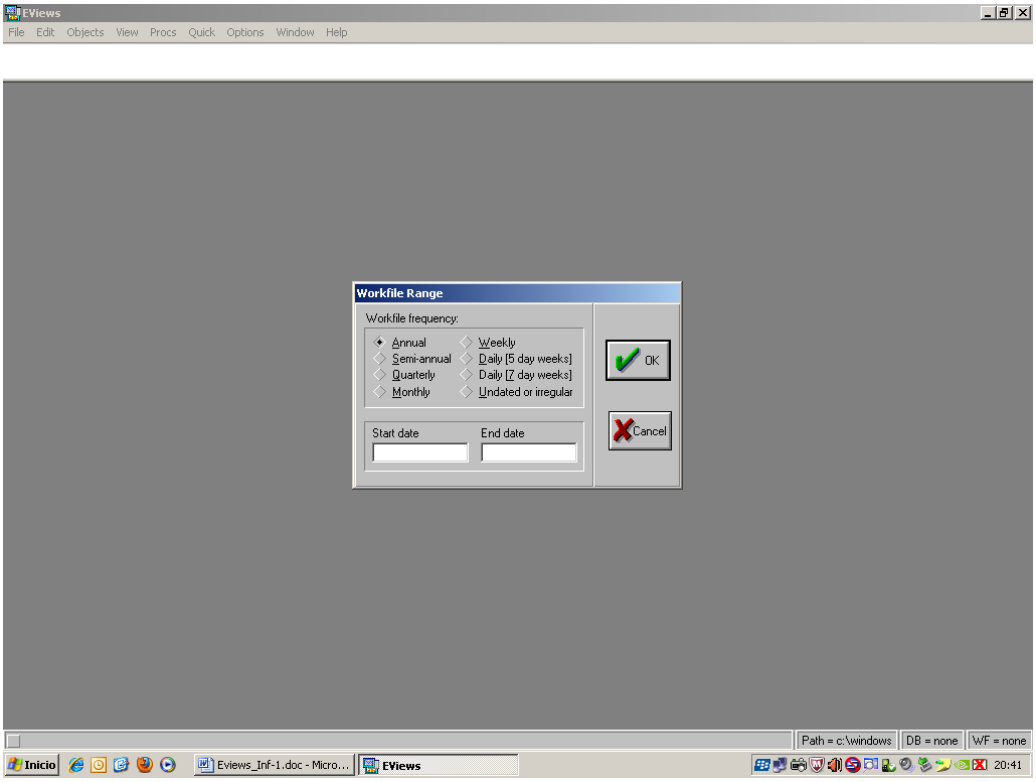


Figura 7

Aquí debemos especificar la frecuencia de los datos (anuales, semestrales, trimestrales, mensuales, semanales, diarios para semanas de 5 ó 7 días; o bien, datos irregulares o sin periodicidad específica). En series temporales, la fecha inicial (*Start date*) y la final (*End date*) permiten definir el rango en el que deberá moverse la serie, teniendo en cuenta que deberá incluirse el periodo de predicción, puesto que *EViews* no admitirá datos para series que superen el rango establecido pero sí que sean inferiores.

Al seleccionar, por ejemplo, la opción de frecuencia trimestral (*Quarterly*) e indicar la fecha de inicio como 1980:1 y 2005:4 como fecha final, se crearía una sesión de trabajo

con datos trimestrales cuyos valores extremos deberían situarse entre el primer trimestre del año 1980 y el cuarto trimestre de 2005.

El problema que vamos a plantear en esta sesión de trabajo es el Ejercicio nº 10 del Boletín del Tema 2, que nos ofrece datos de 20 valores contables y de mercado de las acciones correspondientes a otros tantos bancos españoles en un día determinado de agosto de 1995 (es decir, se trata de datos de corte transversal o no temporales). Nuestro objetivo será tratar de establecer una relación econométrica entre el valor de mercado de las acciones de los bancos y sus respectivos valores contables.

Por tanto, en la opción del menú que se nos ha abierto elegiremos *Undated or irregular* y, seguidamente, escribiremos: *1* en *Start observation*; y *20* en *End observation*, tal y como se indica en la *Figura 8*.

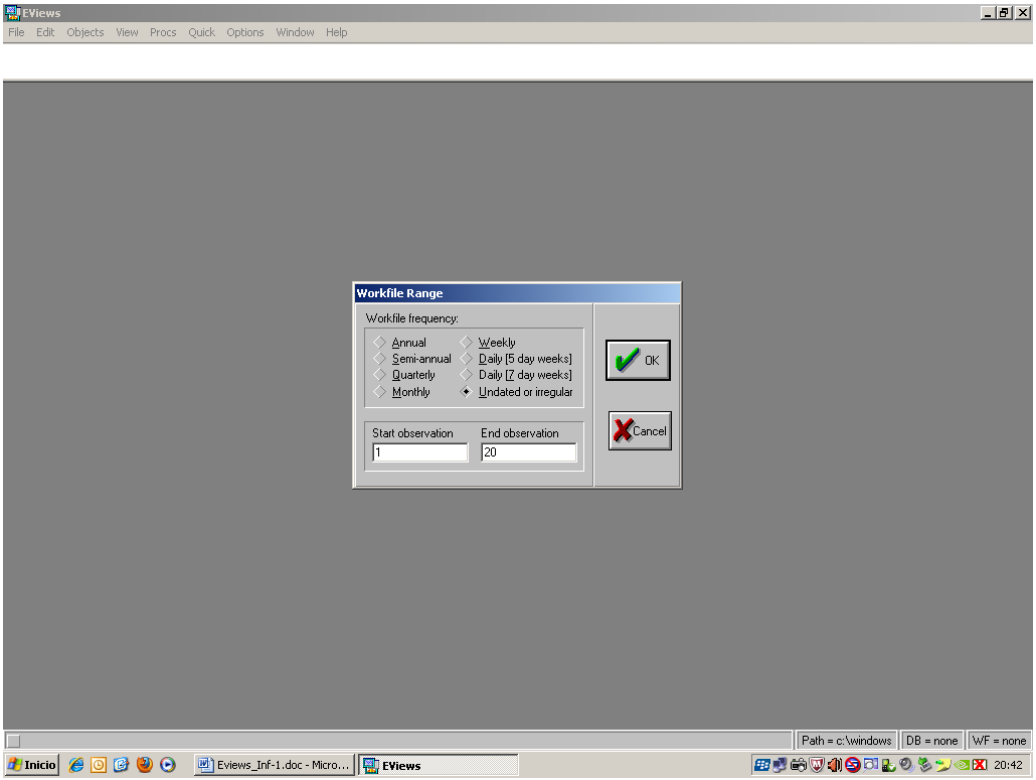


Figura 8

Tras esto, nos aparecerá la VENTANA DEL FICHERO DE TRABAJO (*Workfile: UNTITLED*). Esto se puede ver en la *Figura 9*.

En esta ventana, *RANGE* mostrará el rango en el que toda la serie de datos debe moverse; en cambio *SAMPLE* señalará el periodo o muestra concreta que se utiliza en el estudio. Aunque inicialmente ambos se igualan, es muy habitual que la muestra se cambie durante la sesión de trabajo para adecuarla a cada cálculo que se desee realizar.

El espacio en blanco en la ventana es el DIRECTORIO DE OBJETOS en el que irán apareciendo todos los objetos del fichero de trabajo con su nombre correspondiente y un icono que indica de qué tipo de objeto se trata: vector de coeficientes de regresión,

ecuación, gráfico, grupo de series, matrices, modelos, series, datos de panel, tablas, etc. Por defecto, aquí siempre aparecerán dos elementos: un vector de coeficientes de regresión denominado *c* y representado por α, que incluirá los coeficientes estimados de la última ecuación activa; y una serie llamada *resid* que, como su propio nombre sugiere, está destinada a guardar los residuos de la última ecuación estimada.

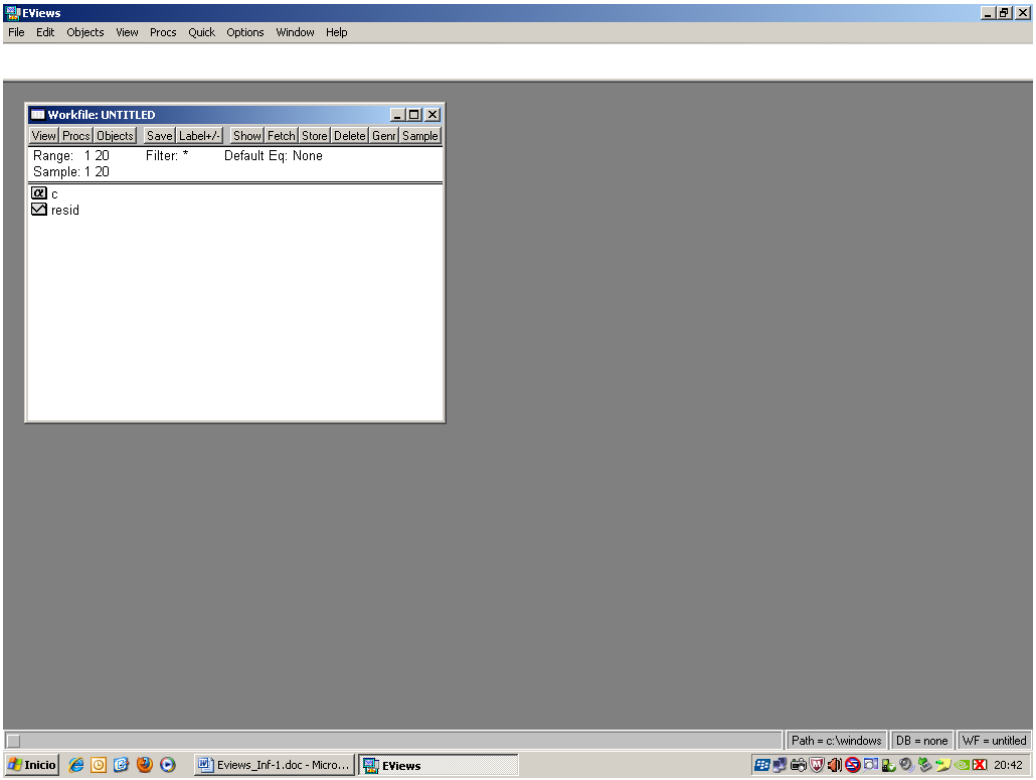


Figura 9

La ventana del Fichero de Trabajo contiene, además, una barra de menús propia cuyos elementos describiremos brevemente:

- ✓ **View, Procs y Objects;** Nos proporcionan los mismos menús desplegables que vimos en la barra de menús principal de *EViews*.
- ✓ **Save;** Permite guardar el fichero de trabajo en uso en el disco duro o en una unidad externa de almacenamiento (disquete, CD, *pen-drive*…).
- ✓ **Label +/-;** Permite visualizar el detalle de los objetos que se presentan en la ventana de trabajo, así como fecha y hora de creación.
- ✓ **Show;** Permite visualizar una serie de datos, una lista de series y gráficos en una misma ventana o una serie generada a través de una fórmula apropiada.
- ✓ **Fetch;** Permite importar distintos objetos (ecuaciones estimadas, series de datos, gráficos…) de otro fichero de trabajo distinto de *EViews*, que hayamos almacenado previamente.

 ISBN: 978-84-694-7251-4

- ✓ **Store;** Permite exportar distintos objetos (ecuaciones estimadas, series de datos, gráficos…) de otro fichero de trabajo distinto de *EViews*, que hayamos almacenado previamente.
- ✓ **Delete;** Permite borrar cualquier objeto que esté en la ventana del fichero de trabajo.
- ✓ **Genr;** Permite generar cualquier serie de datos a partir de una operación con otras ya establecidas.
- ✓ **Sample;** Permite seleccionar la amplitud de la muestra dentro del rango que se ha especificado para el fichero de trabajo.

Llegados a este punto, ya estamos en condiciones de empezar a trabajar con series de datos y de iniciar nuestra sesión de trabajo.

La opción puede ser doble: crear una nueva serie, o bien importarla desde otro fichero de trabajo de *EViews* o de cualquier base de datos u hoja de cálculo, como por ejemplo *Excel*.

A continuación vamos a explicar el procedimiento para introducir nuevas series, aunque posteriormente optaremos, para trabajar con los datos del Ejercicio que vamos a resolver, por la importación del fichero ***agosto95.xls***, que previamente deberemos descargar en el *Escritorio* de nuestro PC desde el espacio reservado a la Asignatura en la plataforma de docencia virtual *WebCT*.

Series de datos: introducción e importación

Para la introducción directa de los datos, a modo de ejemplo, haremos lo siguiente:

- En la opción *OBJECTS* del menú de *Workfil*e se selecciona *NEW OBJECT*, lo que dará lugar a una ventana (*Figura 10*) donde podemos elegir entre diversas opciones: series, ecuaciones, gráficos, etc., y nombrarlos como creamos oportuno. En este punto, se selecciona *SERIES* y le damos el nombre CONSUMO (sobrescribiendo en *Untitled*). Se pueden añadir etiquetas que ayuden a describir el contenido de las series; ello se hace abriendo la serie CONSUMO (haciendo "clic" en ella) y pulsando a continuación la opción *NAME* en la ventana correspondiente. Finalmente, se pulsa *OK*. El resultado de lo realizado puede verse posteriormente pulsando *LABEL +/-* en la ventana de trabajo.
- Una vez creada la serie CONSUMO, introduciremos los datos; por ejemplo:

CONSUMO	125	205	333	214	512

Para ello, podemos hacer doble "clic" sobre la nueva serie CONSUMO que aparece junto a *c* y *resid* en el *Directorio de Objetos* de *Workfile*. Tras abrirse la ventana

 ISBN: 978-84-694-7251-4

correspondiente, pulsaremos en *EDIT +/-* para poder comenzar a introducir los datos. La pantalla deberá quedar finalmente como aparece en la *Figura 11*.

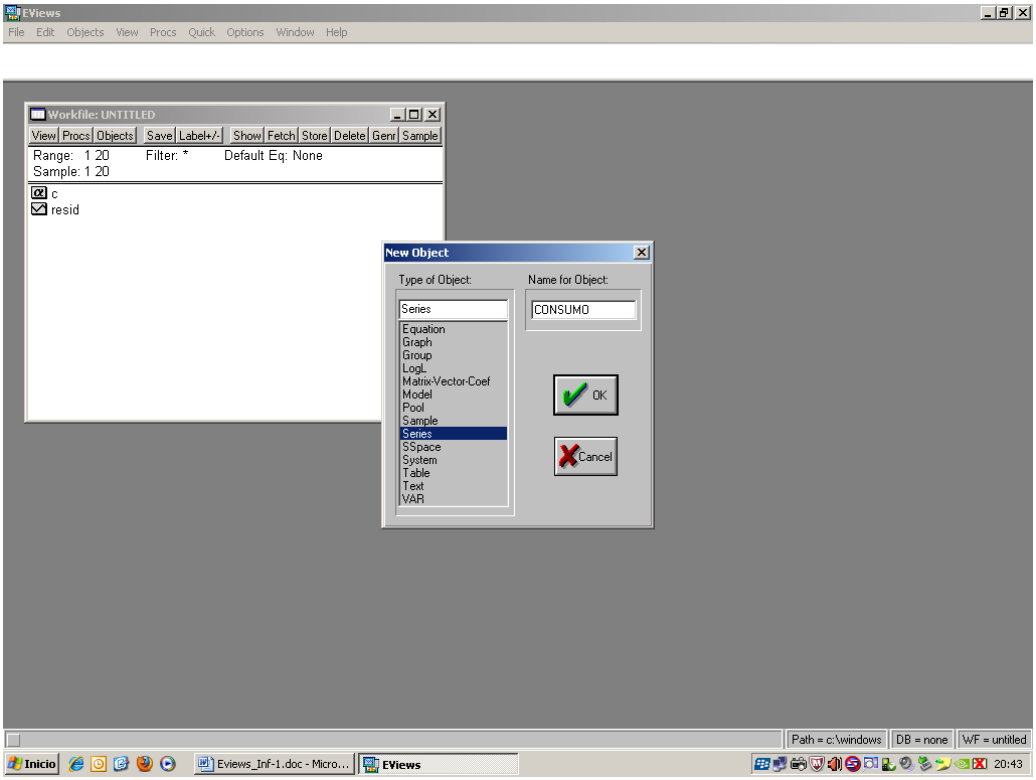


Figura 10

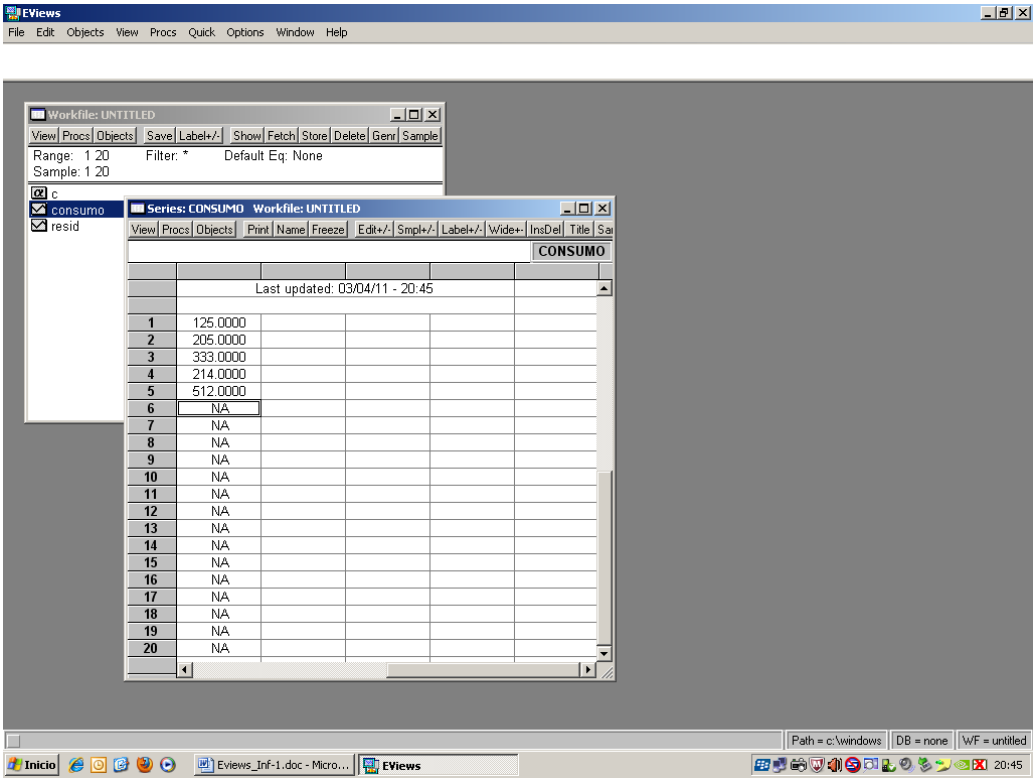


Figura 11

 ISBN: 978-84-694-7251-4

Obsérvese cómo antes de introducir los datos, la serie no tiene todavía valores, como lo indica su referencia NA (*Not Available*). Asimismo, otro aspecto que debemos reseñar aquí es que en *EViews* los decimales están en notación anglosajona, es decir, precedidos de un punto.

Después de haber visto este breve ejemplo de introducción directa de datos en *EViews*, pasamos ya a resolver nuestro ejercicio. Cerramos entonces la ventana de trabajo de la serie CONSUMO (e incluso, si queremos, podemos eliminarla situándonos sobre ella y pulsando la opción *DELETE* en el menú de *Workfile*).

Los datos relativos al problema sobre los valores bancarios que queremos analizar se encuentran en un fichero de *Excel*, del que deberemos importarlos. Para ello, desde el menú principal del fichero de trabajo deberemos seleccionar la opción: *FILE / IMPORT / READ TEXT-LOTUS-EXCEL….*

De este modo, se obtendrá la pantalla que vemos en la *Figura 12*, donde se debe especificar la ruta por la que se accede al fichero *agosto95.xls*, que es la siguiente:

Escritorio \ agosto95.xls

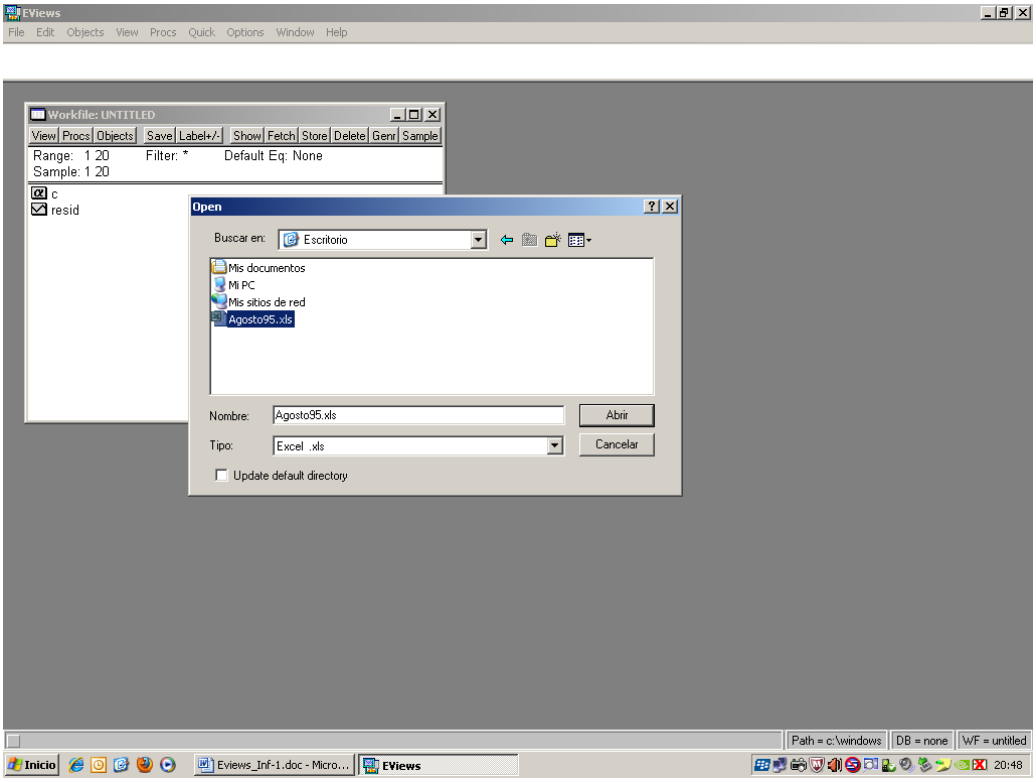


Figura 12

Tras haber seleccionado este fichero y pulsado la opción *ABRIR*, obtendremos el menú desplegable correspondiente a la importación de ficheros *Excel* (*Excel Spreadsheet Import*).

En este menú debemos atender a varias cuestiones importantes:

- En la opción *Order of Data*, el programa *EViews* nos permite especificar si los datos que vamos a insertar de *Excel* están ordenados por columnas (*columns*) o por filas (*rows*). En este caso, seleccionamos la primera opción.
- En segundo lugar, debemos indicar a *EViews* en qué celda de *Excel* se sitúa el primer dato de nuestras series (*Upper-left data cell*). Por defecto, nos propone *B2* y así lo vamos a aceptar puesto que el fichero está preparado en este sentido. Igualmente, debemos indicar el nombre de la hoja donde están nuestros datos (para versiones de *Excel* superiores a la 5, y siempre y cuando haya datos en más de una hoja).
- En tercer lugar, hemos de especificar el nombre de las series que vamos a utilizar o, en su defecto, si el fichero *Excel* ya trae en su primera fila los nombres, sólo el número de series que vamos a importar. Como este último es nuestro caso, sólo escribiremos en *Names for series or Number of series if names in file* un *2*.
- Por último, debemos cerciorarnos de que tenemos correctamente especificada la muestra de datos que queremos importar; esto es, en *Sample to import* debe figurar: *1 20*.

La *Figura 13* muestra el resultado final de la pantalla.

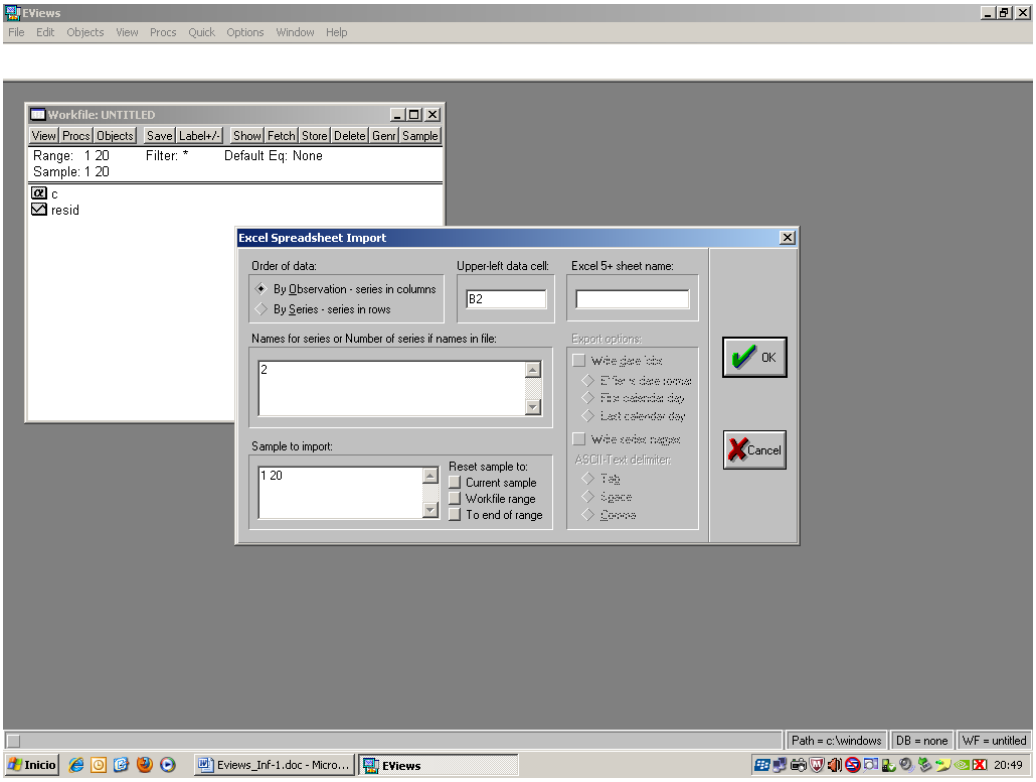


Figura 13

 ISBN: 978-84-694-7251-4

Una vez que hayamos seleccionado todos estos argumentos en este submenú, estaremos en condiciones de pulsar *OK* y proceder a la importación de las dos series de datos de nuestro ejercicio: VACC, como el valor de las acciones de los 20 bancos más importantes de España; y VCON, como el valor contable de las acciones de dichos bancos (ambas variables expresadas en la unidad monetaria de entonces; concretamente, en millones de pesetas).

La *Figura 14* representa el estado final de la ventana del fichero de trabajo después de importar los datos.

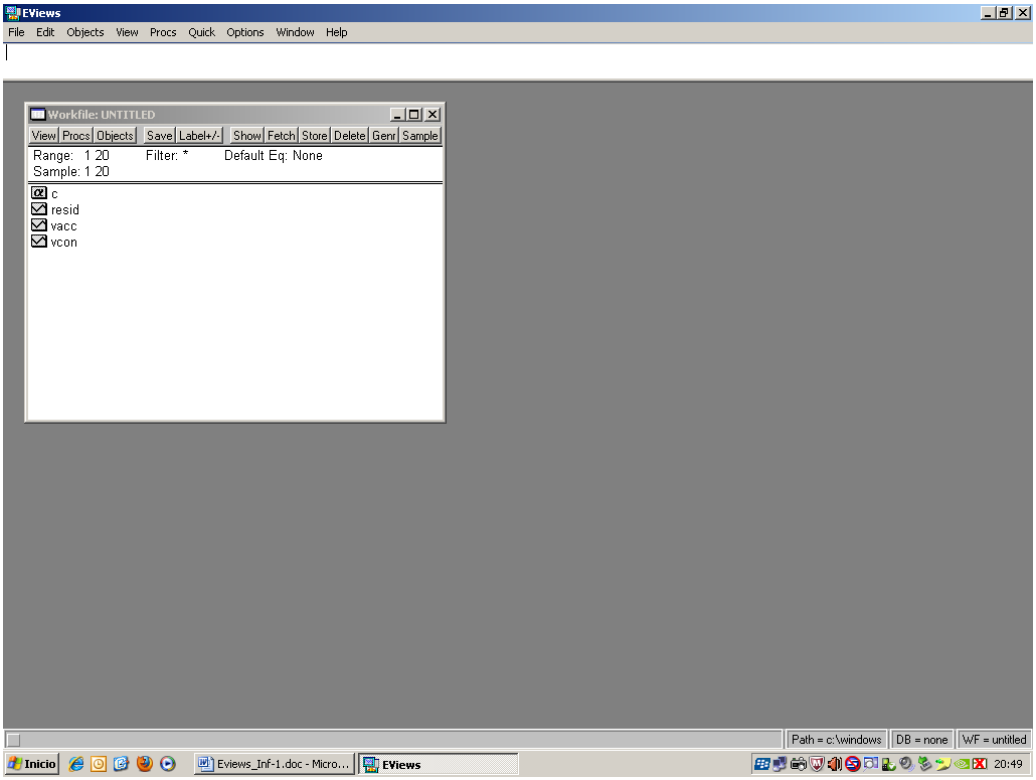


Figura 14

Análisis de una serie de datos

El programa *EViews* nos ofrece la posibilidad de realizar un rápido análisis estadístico de nuestras series de datos.

Si hacemos, por ejemplo, doble "clic" sobre la serie VACC, aparecerá, como ya vimos antes con la serie CONSUMO, la ventana correspondiente de esta serie (*Figura 15*).

En esta ventana surge una nueva barra de menús, cuyos componentes son: *View*, *Procs*, *Objects*, *Print*, *Name*, *Freeze*, *Edit+/-*, *Smpl+/-*, *Label+/-*, *Wide+-*, *InsDel*, *Title*, *Sample* y *Genr*. Cada uno de ellos, a su vez, contiene numerosas posibilidades.

 ISBN: 978-84-694-7251-4

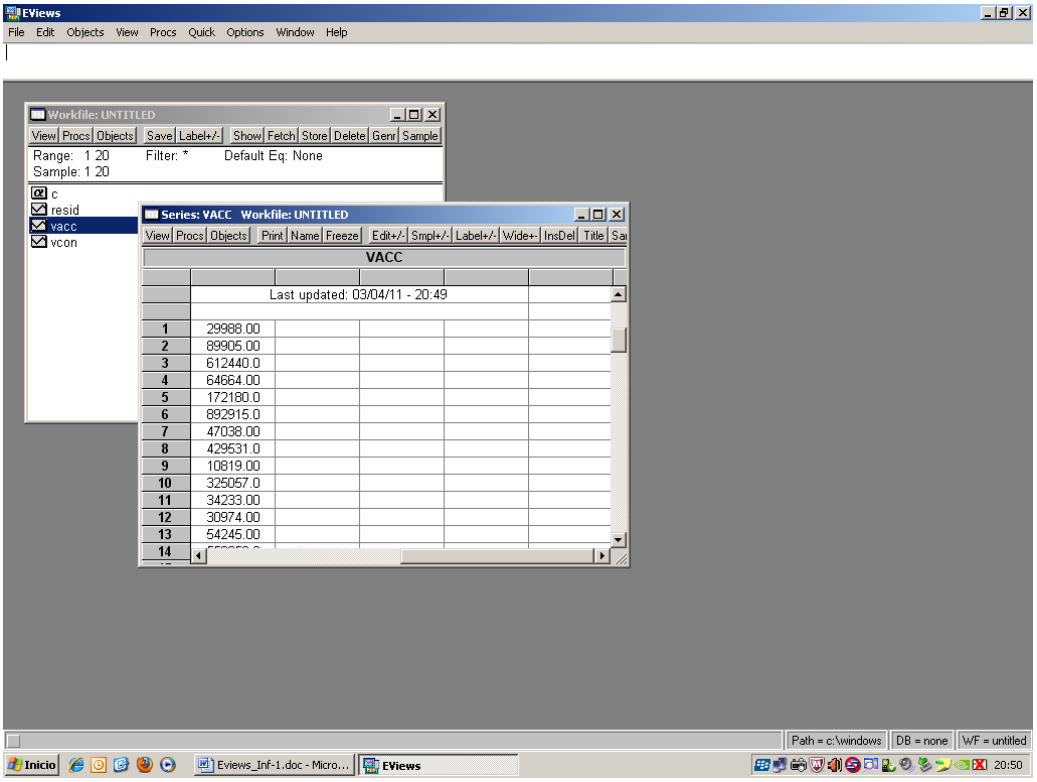


Figura 15

La opción *VIEW* resulta muy interesante. Permite, entre otras acciones (*Figura 16*):

- La representación gráfica de la serie seleccionada en función de un gráfico de líneas o uno de barras: *LINE GRAPH* o *BAR GRAPH*, respectivamente.
- El análisis del histograma y los estadísticos descriptivos de la serie: media, mediana, valor máximo, valor mínimo, desviación típica, coeficiente de asimetría de *Fisher* y coeficiente de curtosis de *Fisher*, así como el estadístico de *Jarque-Bera*, que permite contrastar la normalidad de la serie en cuestión. Todo esto se hace a través de: *DESCRIPTIVE STATISTICS / HISTOGRAM AND STATS*. El resultado puede observarse en la *Figura 17*.
- Otras posibilidades de esta opción del menú, tales como *CORRELOGRAM*, se verán más adelante.

Tras realizar cualquiera de las acciones indicadas en *VIEW*, puede que queramos volver al listado de datos de la variable. En este caso, bastará con elegir *SPREADSHEET*.

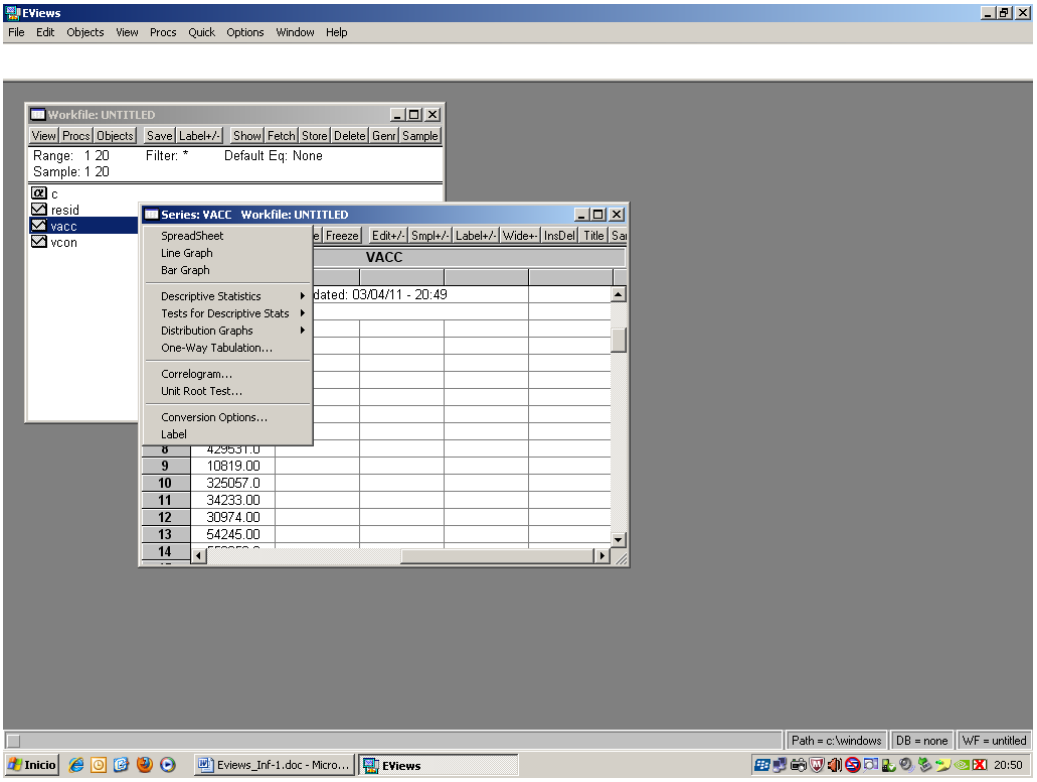


Figura 16

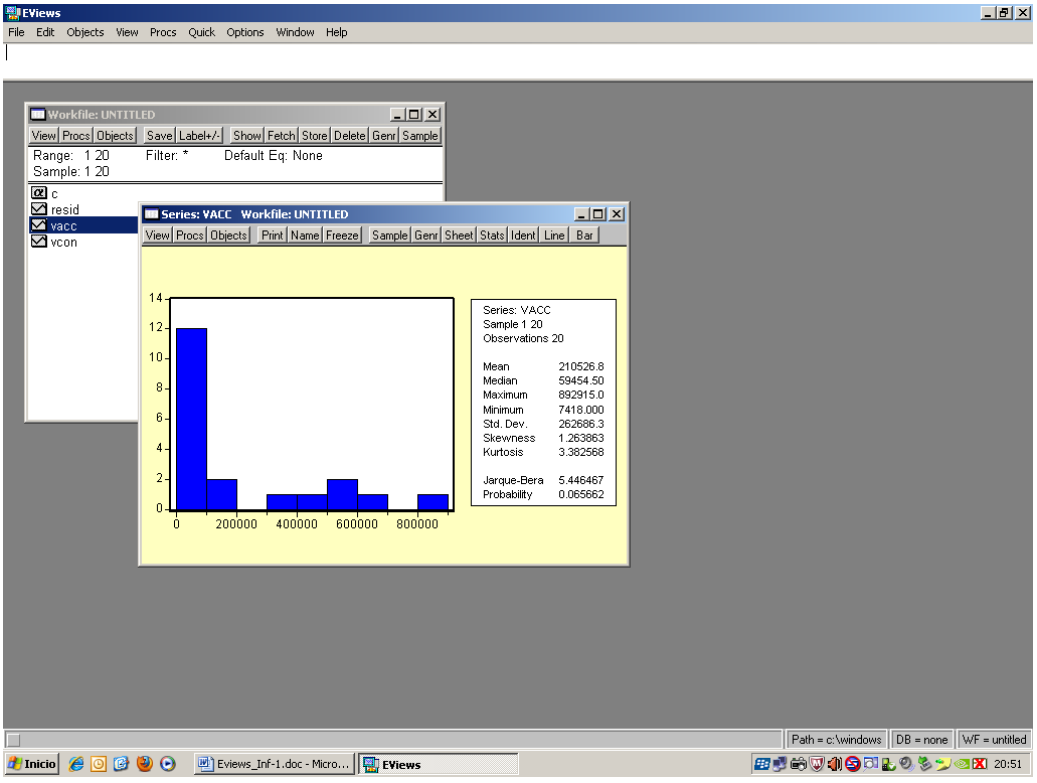


Figura 17

Estimación de un modelo lineal por MCO

El presente ejercicio trata de plantear la existencia de una relación econométrica entre el valor de las acciones de mercado de 20 bancos en un día determinado y sus correspondientes valores contables. En concreto, estableceremos una relación lineal.

En principio, sería interesante conocer la representación gráfica de la nube de puntos con objeto de tener una idea de la idoneidad o no de la forma lineal que vamos a plantear. Para ello, en la *barra principal de menús* seleccionaremos *QUICK / GRAPH*. Se creará así una nueva ventana (*Figura 18*) en la que escribiremos en primer lugar la variable independiente y luego la dependiente (en nuestro caso, VCON y VACC, respectivamente). Tras pulsar *OK*, se nos abrirá un nuevo cuadro de diálogo en el que podremos indicar el tipo de gráfico que queremos: aquí elegiremos *Scatter Diagram* (es decir, gráfico de dispersión o nube de puntos). Antes de aceptar, podemos pulsar *SHOW OPTIONS* y seleccionar *Regression line*; de este modo, el programa nos trazará también la recta de regresión que mejor se ajustaría a la nube de puntos (*Figura 19*). El resto de opciones no las tocaremos.

Para concluir, pulsaremos *OK*. El resultado final es el que aparece en la *Figura 20*.

Si queremos guardarlo, basta con pulsar *NAME* y llamarlo, por ejemplo, NUBE. Se creará así un objeto con formato "gráfico".

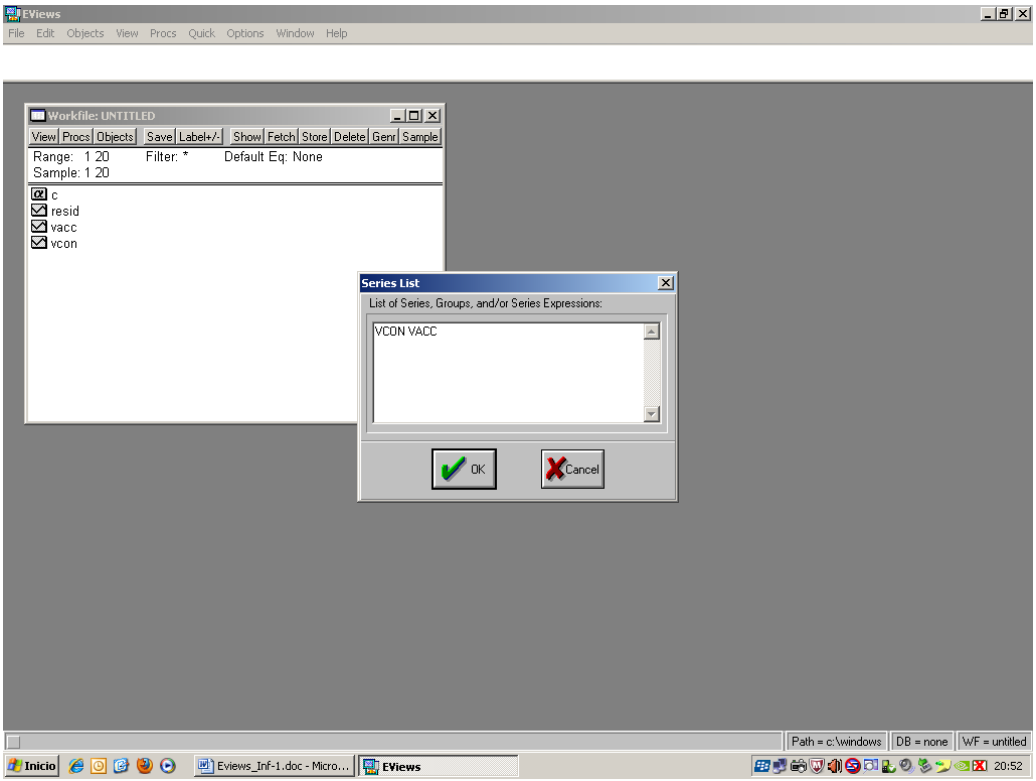


Figura 18

 ISBN: 978-84-694-7251-4

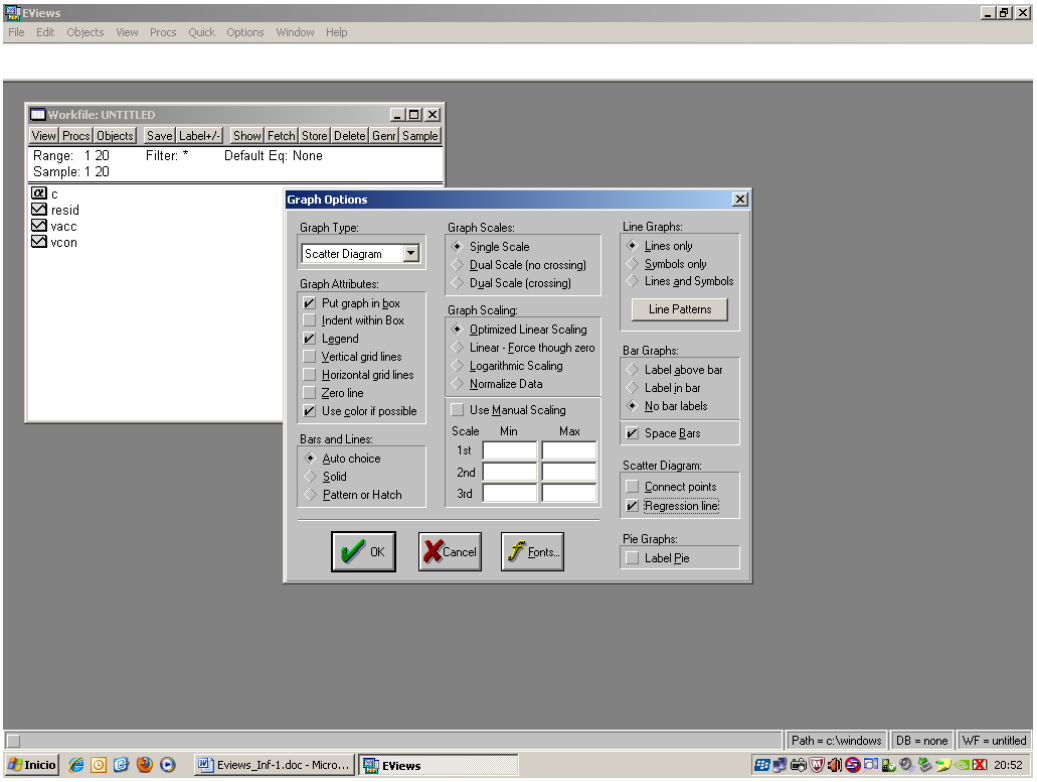


Figura 19

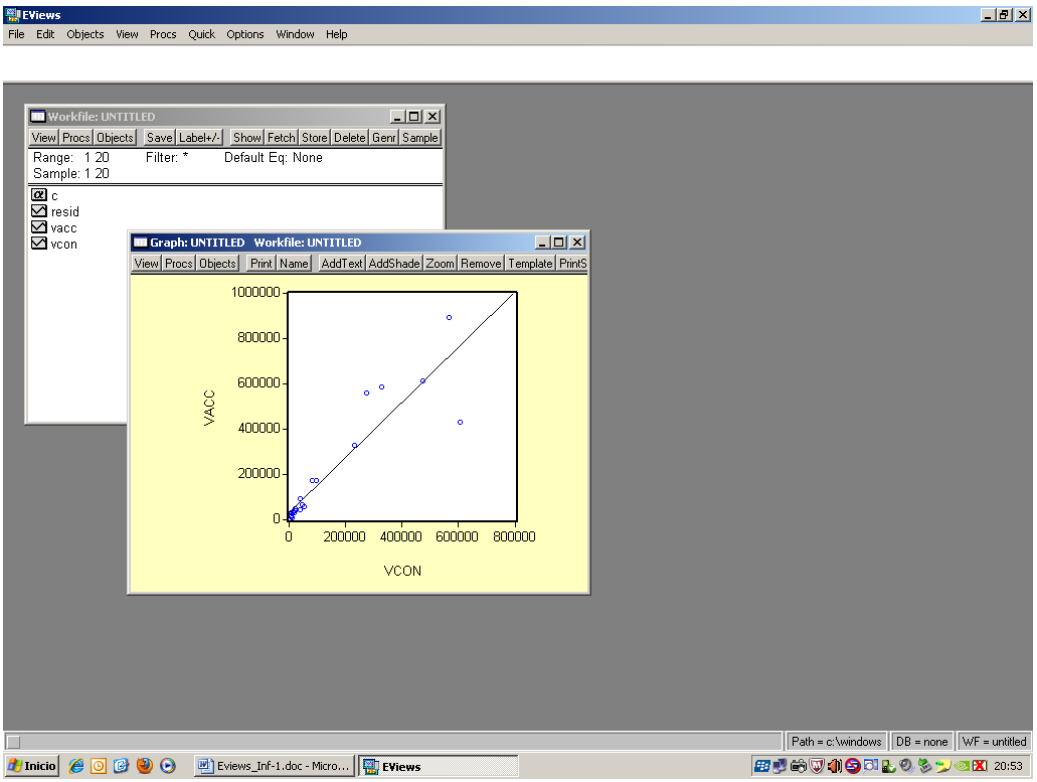


Figura 20

A continuación, para obtener la recta estimada de regresión, seleccionamos la opción *QUICK / ESTIMATE EQUATION* en la *barra principal de menús* (*Figura 21*).

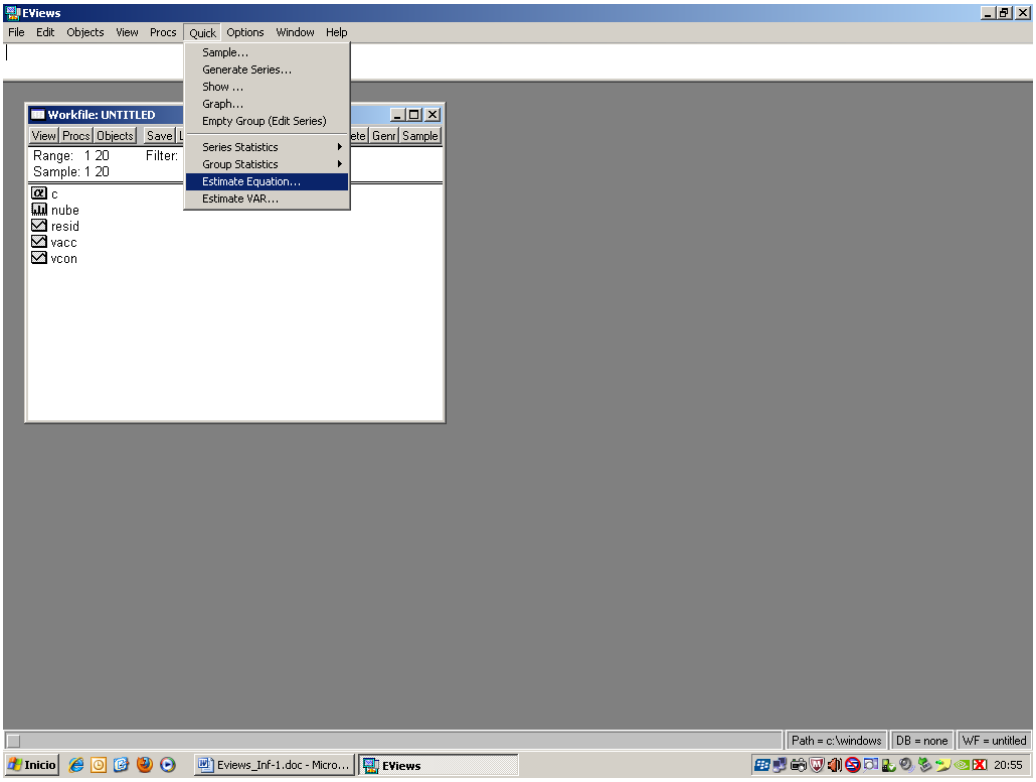


Figura 21

Como resultado, se obtiene la ventana de Especificación de la Ecuación (*Equation Specification*), donde deberemos atender a dos aspectos importantes:

- *Equation Specification*; en este apartado debemos especificar la regresión lineal que vamos a realizar. En primer lugar hay que escribir el nombre de la serie que actuará como variable dependiente. Tras ello deberemos escribir, separados por espacios, la relación de variables independientes o explicativas del modelo, comenzando por la constante u ordenada en el origen (que se denota por la letra *C*) si deseamos que esté presente en nuestro modelo.

  Así pues, dado que en este caso el modelo que planteamos es:

  $$VACC = \beta_1 + \beta_2 VCON + u,$$

  escribiremos entonces: VACC C VCON.

- *Estimation Settings*; aquí se configuran las características de la estimación, seleccionándose tanto el método de estimación (MCO, métodos para modelos binarios, etc.), como la muestra a utilizar.

  En nuestro caso elegiremos MCO, que en inglés es: *LS* (*Least Squares*).

Tras especificar la ecuación y el método de estimación (*Figura 22*), pulsaremos *OK*, obteniendo la *Ventana de Ecuación* (*Equation: Untitled*), donde se muestran tres tipos de resultados: los resultados básicos del proceso de estimación, las estimaciones de los coeficientes de regresión y los estimadores destinados a estudiar su significatividad y,

finalmente, los estadísticos más relevantes empleados en cualquier método econométrico (*Figura 23*).

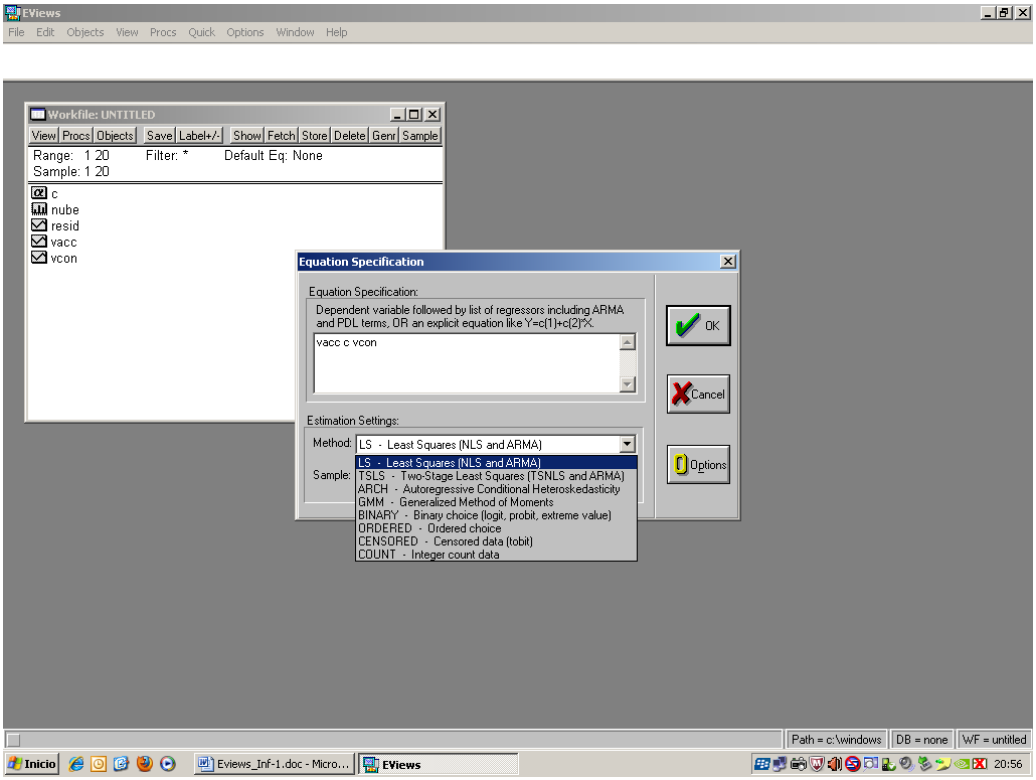


Figura 22

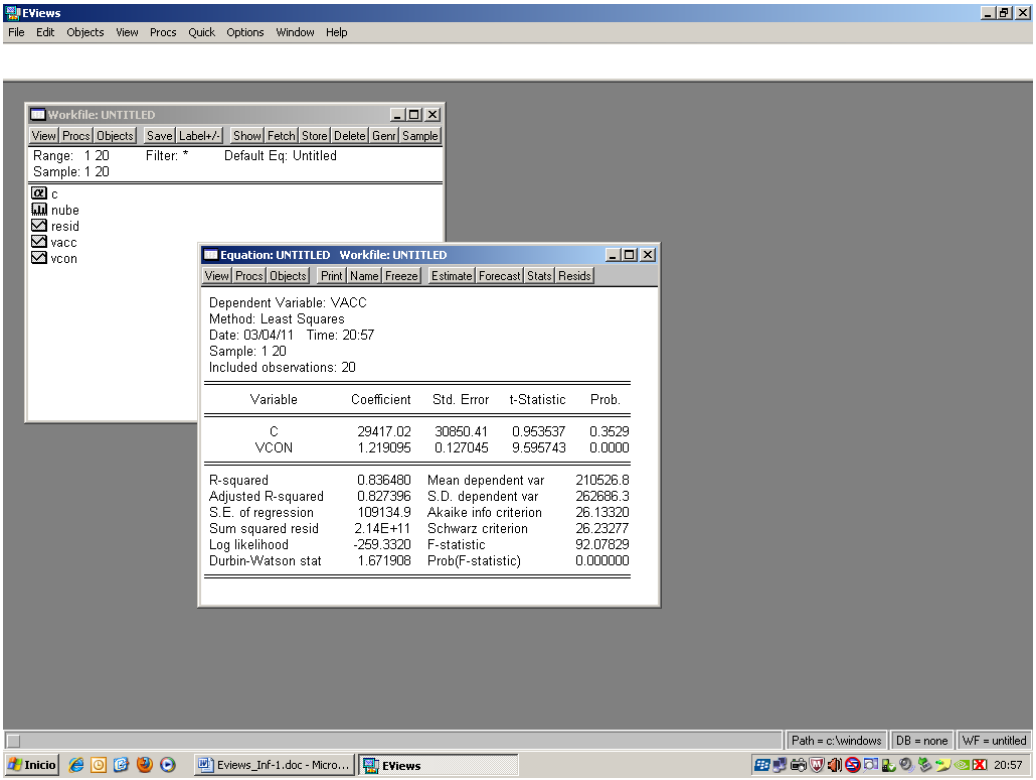


Figura 23

- ***Resultados básicos del proceso de estimación*:** (parte superior de la ventana)

- El nombre de la variable dependiente o explicada de la ecuación.
- El método de estimación aplicado. En este caso, MCO: *LS* (*Least Squares Method*).
- La fecha y hora de realización de la estimación.
- La muestra empleada en la estimación.
- El número de observaciones utilizadas en la estimación.

- ***Estimaciones de los coeficientes de regresión y estimadores relevantes para el análisis de su significatividad*:** (parte central de la ventana)

- El nombre de las variables independientes o explicativas de la ecuación.
- Estimación de los coeficientes de regresión asociados a las variables independientes o explicativas (*Coefficient*: $\hat{\beta}_1$ y $\hat{\beta}_2$, en este caso).
- Estimación de los errores estándar de los coeficientes de regresión (*Std. Error*).
- El valor del estadístico *t-Student* bajo la hipótesis nula de no significatividad de la variable correspondiente (*t-Statistic*).
- Nivel de significación mínimo al que se puede rechazar la hipótesis nula anterior bajo el supuesto de que fuera cierta (*Prob.*). Es lo que se conoce como ***p-valor*** y nos permite realizar un contraste sin necesidad de tablas estadísticas de distribución de probabilidades.

- ***Estadísticos relevantes de la estimación econométrica*:** (parte inferior de la ventana)

- El valor del coeficiente de determinación del modelo (*R-Squared*): R^2.
- El valor del coeficiente de determinación corregido del modelo (*Adjusted R-Squared*): R_c^2.
- Estimación de la raíz cuadrada positiva de la varianza (desviación típica o error estándar) de la perturbación aleatoria (*S.E. of regresión*: $\hat{\sigma}_u$).
- El valor de la suma de cuadrados residuales de la estimación (*Sum Squared Resid*).
- El valor del logaritmo de la función de verosimilitud sujeto al valor estimado de los coeficientes de regresión (*Log likelihood*).
- El valor del estadístico de *Durbin-Watson*, que permite la detección de problemas de autocorrelación en el modelo según esquemas AR de orden 1 (*Durbin-Watson stat*).
- El valor de la media de la variable dependiente (*Mean dependent var*).
- El valor de la cuasi-desviación típica de la variable dependiente (*S.D. dependent var*).
- El valor del estadístico del criterio de información de *Akaike*, el cual puede utilizarse para elegir entre diversas especificaciones de modelos cuando la variable

dependiente presenta diferentes formas funcionales, o bien también para elegir entre distintas especificaciones de modelos de elección discreta (*Akaike info criterion*).

- El valor del estadístico del criterio de *Schwarz*, como alternativa al criterio de información de *Akaike* (*Schwarz criterion*).
- El valor del estadístico *F de Fisher-Snedecor* bajo la hipótesis nula de no significatividad del modelo.
- Nivel de significación mínimo al que se puede rechazar la hipótesis nula de no significativad del modelo bajo el supuesto de que fuera cierta (*Prob.(F-statistic)*). Es el *p-valor* asociado a la significatividad global del modelo. Permite estudiar este aspecto sin usar las tablas estadísticas de distribución de probabilidades.

Esta ventana nos da una completa visión inicial del modelo especificado. En nuestro ejemplo, si observamos el signo del coeficiente de regresión estimado de VCON, deducimos que la relación entre las variables del modelo es directa (como cabría esperar según la Teoría Económica), concretándose en que cuando el valor contable se incrementa en 1 millón de Ptas., el de las acciones lo hace por término medio en casi 1,22 millones de Ptas. (nótese que en el modelo lineal, el significado de los coeficientes de regresión coincide con el concepto económico de *efecto marginal*). En cuanto a la bondad del ajuste muestral, tanto el valor de R^2 (0,836480), como el de R_c^2 (0,827396), son muy aceptables. Y en el terreno inferencial, que veremos en breve, la variable explicativa resulta significativa, como evidencia el *p-valor* asociado a su estadístico *t*-Student, o bien al estadístico *F* de significatividad global del modelo (que en este caso, por ser un modelo de regresión lineal simple, coinciden en significado).

Resulta conveniente guardar esta ventana como "objeto" para que a lo largo de la sesión, cuando se desee, podamos recuperarla en la ventana del fichero de trabajo y no haya que volver a realizar la estimación. Así, en la *Ventana de Ecuación* seleccionamos la opción *NAME* y le damos un nombre; por ejemplo: VACC_VCON.

EViews, además, nos permite conocer la serie de los residuos, la de los valores estimados de la variable dependiente y la de los valores reales, comparándolas a su vez en un gráfico. Para ello, debemos seleccionar desde el menú de la *Ventana de Ecuación*, la opción *VIEW*, en la que aparecerán a su vez varias opciones tales como *REPRESENTATIONS,* que nos indica el modelo estimado o, entre otras: *ACTUAL, FITTED, RESIDUAL* (*Figura 24*), donde podemos elegir entre cuatro más detalladas:

- **Actual, Fitted, Residual Table**; representa los valores reales (*actual*), estimados (*fitted*) y los residuales (*residual*) en una tabla, junto con un gráfico a su derecha.
- **Actual, Fitted, Residual Graph;** representa gráficamente los valores anteriores.
- **Residual Graph;** representa gráficamente sólo la serie residual.
- **Standardized Residual Graph;** representa gráficamente los residuos tipificados.

 ISBN: 978-84-694-7251-4

Las *Figuras 25* y *26* muestran, respectivamente, los resultados de elegir la primera y la segunda de las opciones indicadas.

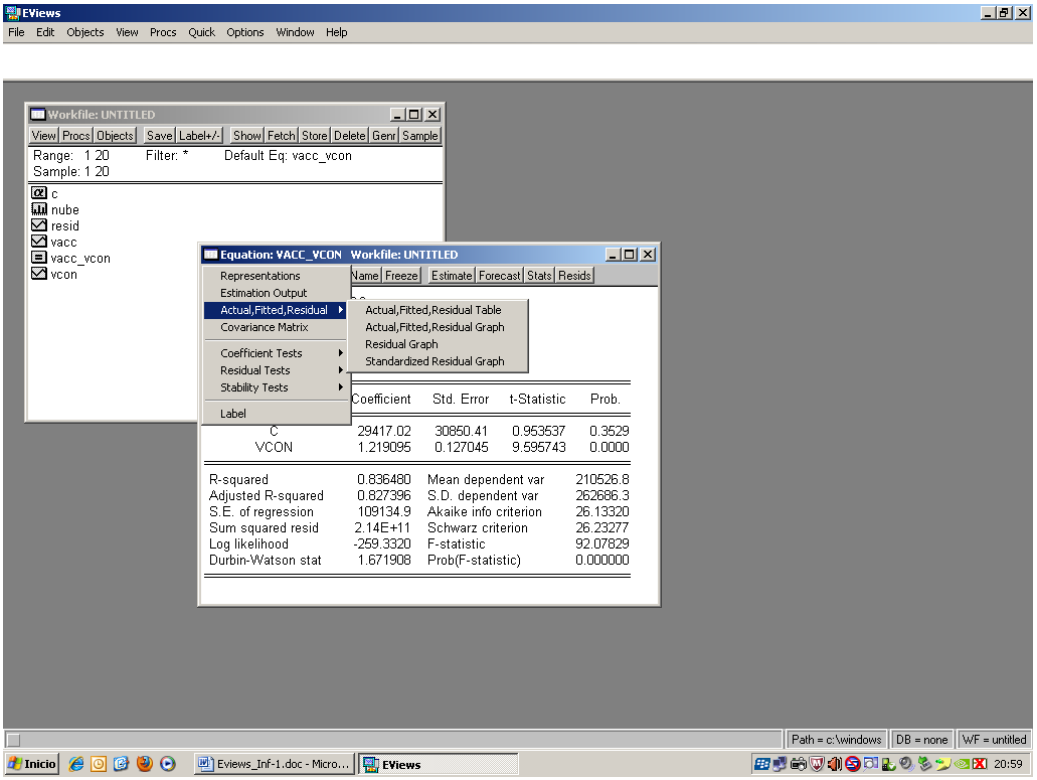


Figura 24

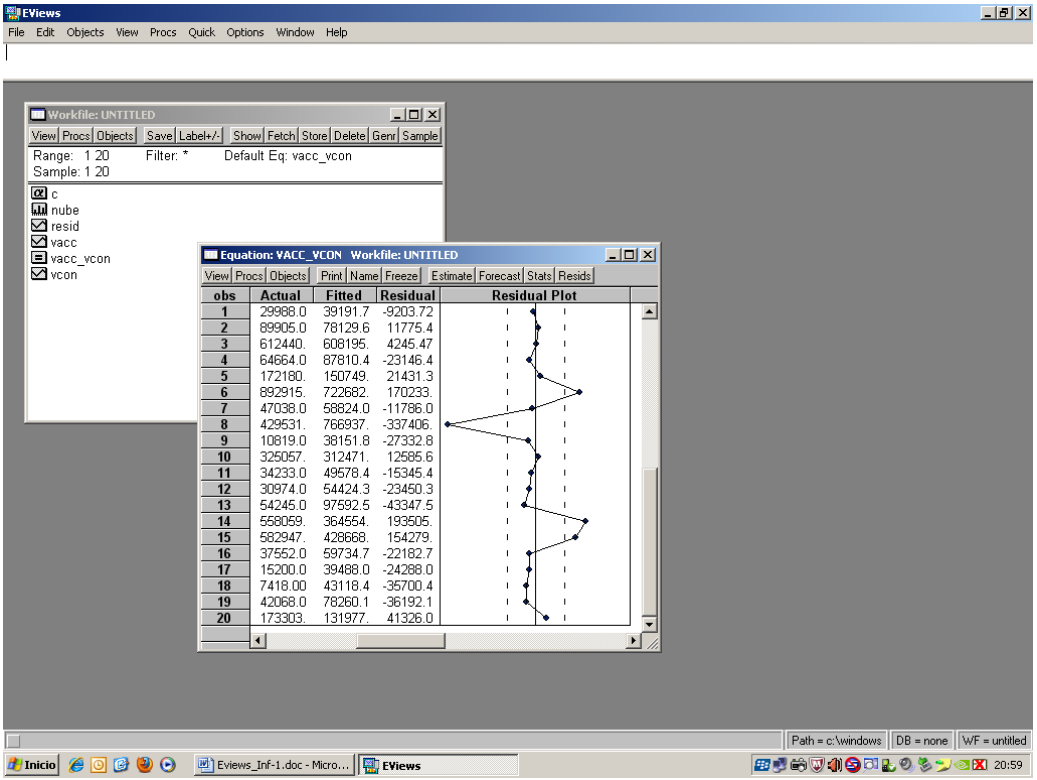


Figura 25

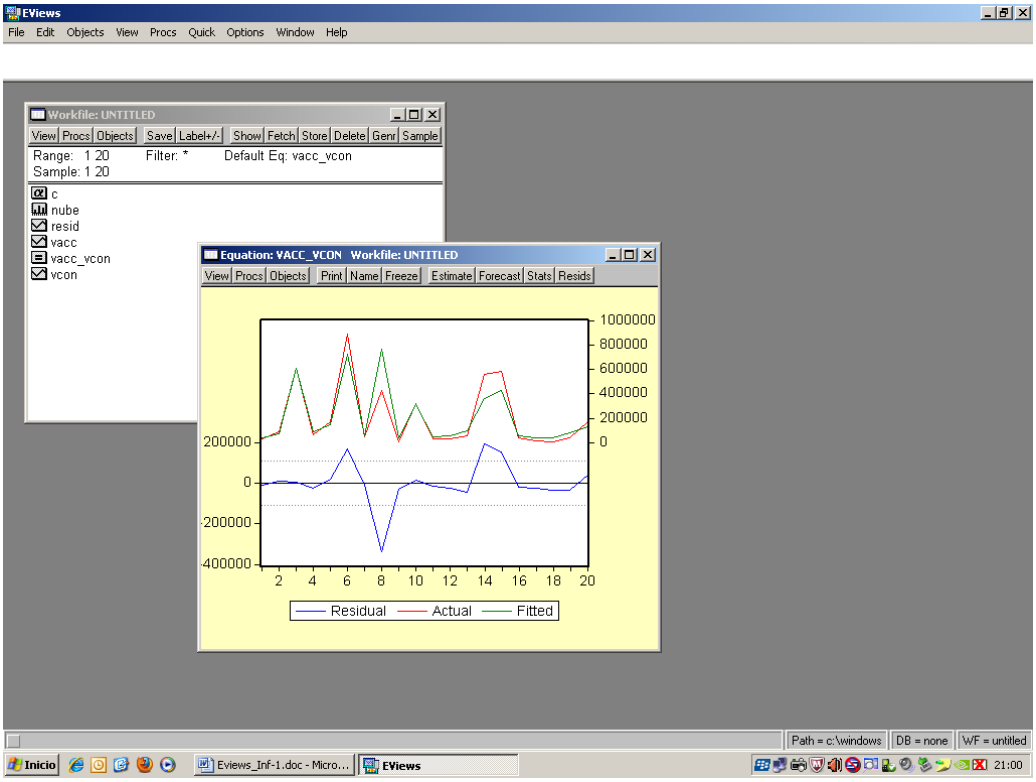


Figura 26

Otro resultado interesante es el cálculo de la matriz de varianzas-covarianzas estimada de los estimadores de los coeficientes de regresión (*Figura 27*).

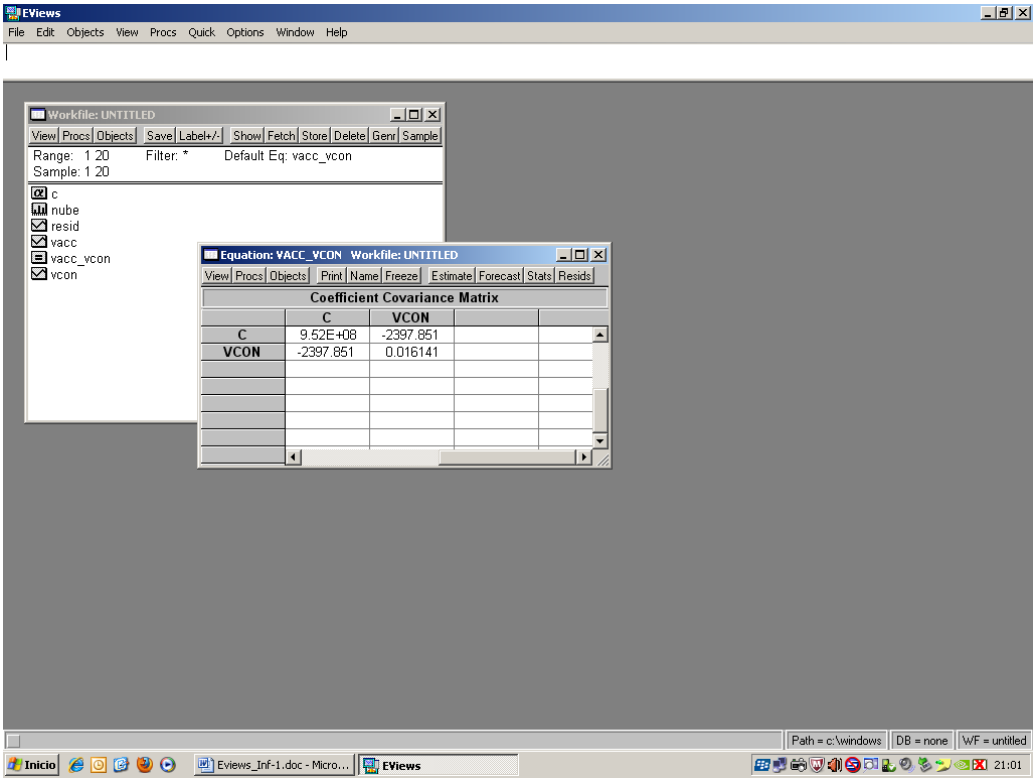


Figura 27

Este resultado, muy útil en los procesos de inferencia, se obtiene también a partir del menú *VIEW*; en concreto, pulsando dentro de él la opción *COVARIANCE MATRIX*.

Ejemplo de estimación de un modelo no lineal por MCO: el modelo log-log

En ocasiones, puede resultar que el modelo lineal no sea el más adecuado para explicar la relación entre distintas variables. Modelos como el potencial, el exponencial u otros no lineales, pero linealizables, pueden ser entonces opciones más apropiadas. El carácter linealizable de éstos permite que a través de determinadas transformaciones de las variables originales se pueda llegar a una sencilla especificación lineal. De este modo, se obtienen modelos como el log-log, el log-lin o el lin-log.

Para poder estimar por MCO un modelo de estas características, debemos en primer lugar definir las nuevas variables. *EViews* habilita la opción *GENR* para generar series a partir de otras ya existentes. Esta opción se encuentra tanto en la ventana principal del fichero de trabajo, como en la *barra principal de menús*: *QUICK / GENERATE SERIES*.

En este punto vamos a plantear como ejemplo un modelo log-log para las variables de nuestro ejercicio; es decir:

$$\boxed{\ln VACC = \beta_1 + \beta_2 \ln VCON + u}$$

Por tanto, tras seleccionar la referida opción *QUICK / GENERATE SERIES*, en el recuadro *Enter equation* escribiremos: LOGVACC = LOG(VACC), tal y como aparece en la *Figura 28*. Al aceptar pulsando *OK*, se obtiene la nueva serie.

Debemos reseñar que "LOGVACC" es el nombre que nosotros hemos querido asignar a la nueva serie; y puede cambiarse a gusto del usuario. Por su parte, en el miembro de la derecha de la expresión (a partir del signo "=") debemos escribir necesariamente el nombre de la función matemática correspondiente que entiende *EViews* (en este caso, LOG), a partir de la cual se genera la nueva serie.

De igual forma, se deberá proceder con la otra variable, VCON. Tendremos entonces LOGVCON.

Una vez creadas las nuevas variables logarítmicas, llevaremos a cabo la estimación del modelo de forma análoga a como se hizo anteriormente con el modelo lineal. Los resultados obtenidos se muestran en la *Figura 29*.

Podemos luego grabar el modelo nombrándolo en *NAME*, por ejemplo, LOGVACC_LOGVCON.

Recuérdese que en un modelo log-log, los coeficientes de regresión expresan *elasticidades*. Por tanto, en este caso, tendríamos que cuando el valor contable se

 ISBN: 978-84-694-7251-4

incrementa un 1%, el valor de las acciones se incrementa por término medio casi un 0,94%.

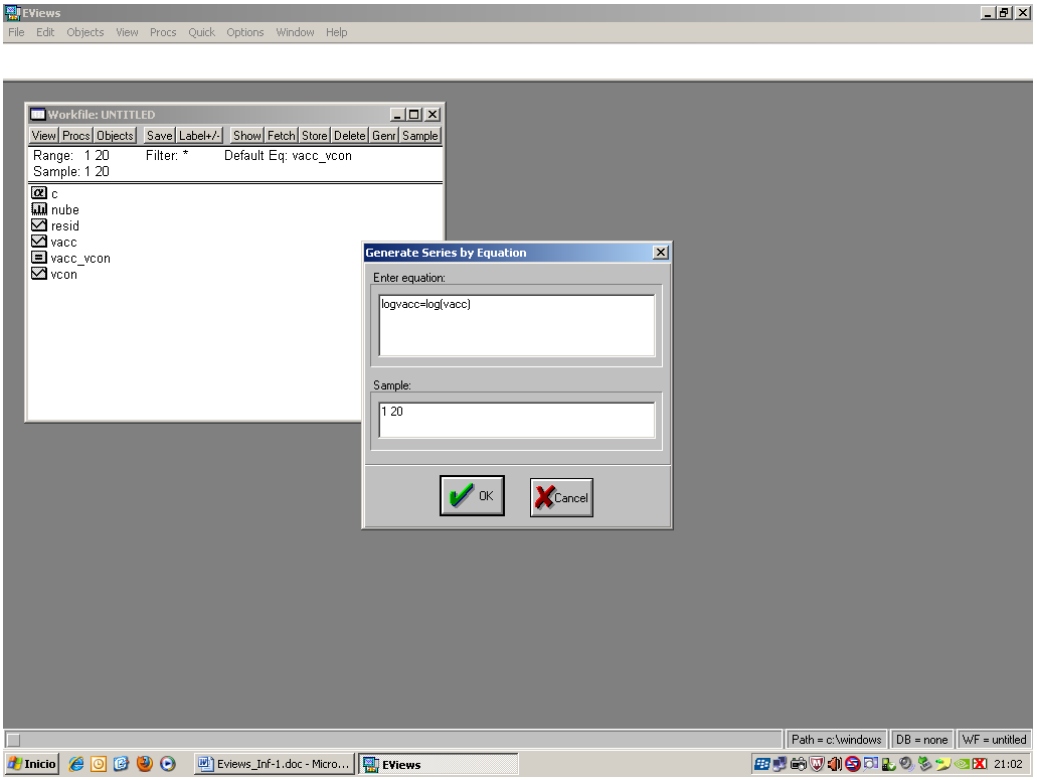


Figura 28

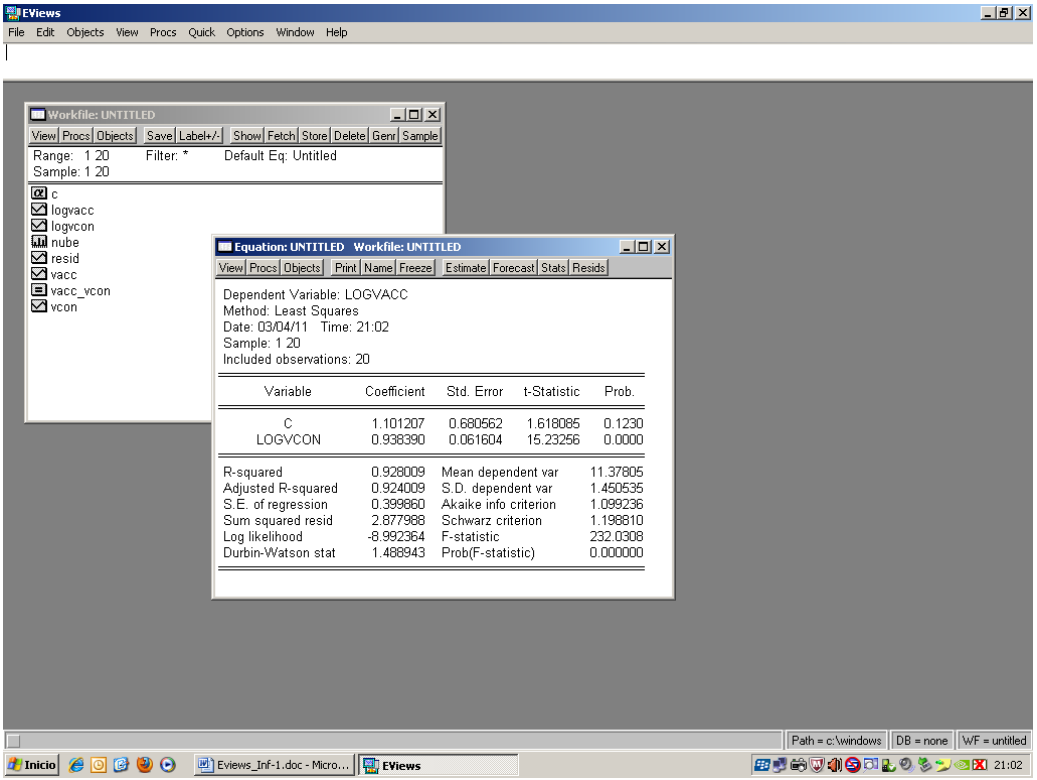


Figura 29

 ISBN: 978-84-694-7251-4

Cómo guardar un archivo de trabajo

Concluida la sesión de trabajo, una de las maneras más rápidas de salir de *EViews* y guardar el fichero de trabajo es haciendo "clic" en el aspa de la esquina superior derecha, como en cualquier otro programa que opera en el entorno *Windows*.

Otra forma es ir, en la barra de menús principal, a *FILE / EXIT*.

Antes de cerrarse, *EViews* procederá a confirmar si el usuario desea guardar el fichero, o bien no guardarlo o no registrar los últimos cambios en el caso de que ya existiera (*Figura 30*).

Para finalizar, cabe señalar que los ficheros realizados en *EViews* se distinguen porque, tras el nombre que les queramos dar, la terminación o sufijo que los caracteriza es "*.wf1*".

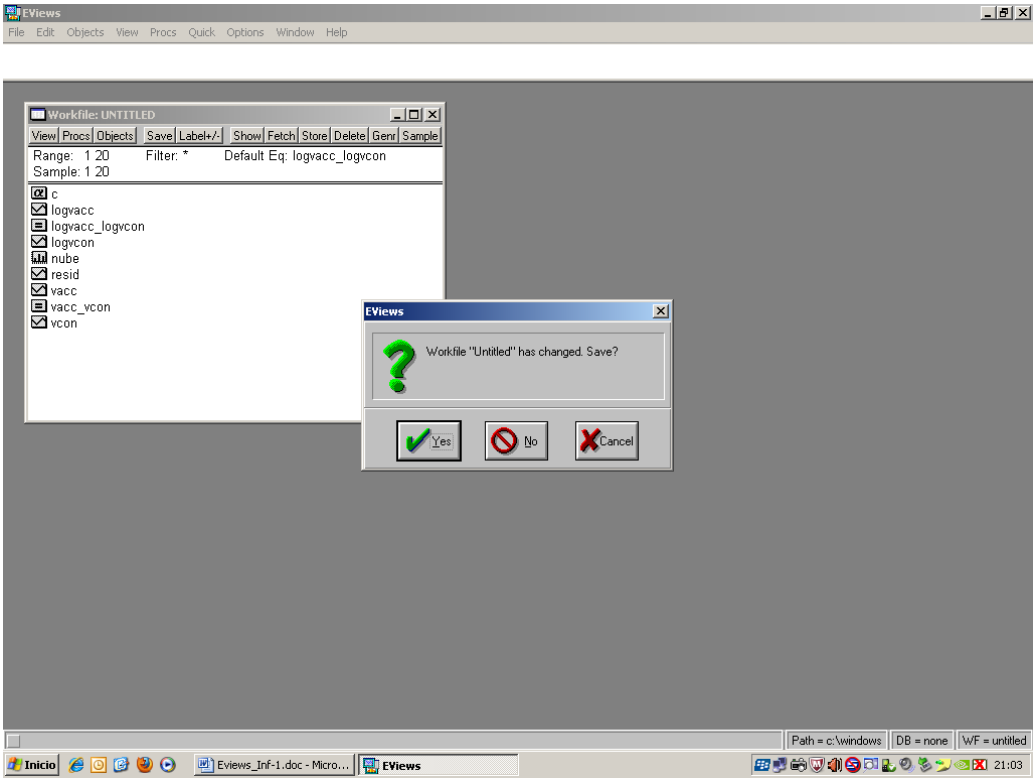


Figura 30

 ISBN: 978-84-694-7251-4

TEMA 3

El modelo clásico de regresión lineal: inferencia y predicción

3.1. Normalidad en las perturbaciones. Contraste de Jarque-Bera.-

Consideremos el modelo clásico de regresión lineal, en su expresión matricial: $Y = X\beta + u$. Como ya sabemos, la perturbación *u* es un vector aleatorio que sigue una distribución de probabilidad normal: $u \to N_n(\theta; \sigma_u{}^2 \cdot I)$.

Precisamente la normalidad de *u* es uno de los supuestos básicos del modelo clásico de regresión. A partir del mismo, hemos visto que tanto *Y* como el vector de coeficientes de regresión estimados $\hat{\beta}$ siguen también distribuciones de probabilidad normales. Y sobre esta premisa veremos que se basa todo nuestro análisis inferencial del modelo.

Así pues, resulta esencial comprobar que *u* se comporta efectivamente siguiendo una distribución normal de probabilidad, pues de ello dependerá la validez de todas las conclusiones que podamos extraer sobre los aspectos inferenciales del modelo.

Para llevar a cabo el estudio de la normalidad de *u*, se utiliza un contraste estadístico: el contraste de Jarque-Bera.

Como en todo contraste de hipótesis, debemos establecer una hipótesis nula y, frente a ella, una hipótesis alternativa. En particular, en este contraste, éstas son:

$$\boxed{\begin{array}{l} H_0 : u \to Normal \\ H_1 : u \to No\, normal \end{array}}$$

Una vez definidas las hipótesis nula y alternativa, en un contraste es preciso también establecer un estadístico de prueba, que tendrá carácter aleatorio (tomando diferentes valores según la muestra que se considere) y seguirá una determinada distribución de probabilidad. En este caso, el estadístico de Jarque-Bera sigue una distribución chi-cuadrado con 2 grados de libertad, siendo su expresión:

$$\boxed{\chi^2_{JB} = n \cdot \left(\frac{\gamma_1^2}{6} + \frac{(\gamma_2 - 3)^2}{24} \right) \to \chi^2_2},$$

donde *n* hace referencia al tamaño muestral, $\gamma_1 = \mu_3 / \sigma^3$ es el coeficiente de asimetría de la distribución y $\gamma_2 = \mu_4 / \sigma^4$ es su coeficiente de curtosis.

En una distribución normal, tendríamos que $\gamma_1 = 0$ (al ser simétrica) y $\gamma_2 = 3$, con lo que si la perturbación aleatoria cumpliese la hipótesis nula de normalidad, el estadístico

 ISBN: 978-84-694-7251-4

χ^2_{JB} valdría 0 (ó un valor muy próximo a 0); es decir, si la perturbación es normal, tendrá asociado un valor pequeño del estadístico χ^2_{JB}. Por tanto:

$$H_0 : u \to Normal \qquad (\chi^2_{JB} = 0)$$
$$H_1 : u \to No\,normal \quad (\chi^2_{JB} > 0)$$

En este punto debemos de hacer una observación importante. Nuestra variable objeto de estudio es la perturbación aleatoria; sin embargo, ésta resulta inobservable, por lo que no podremos analizarla directamente. Por ello, a la hora de estudiar *u*, tendremos que recurrir a una estimación de la misma: al residuo o error muestral. Recordemos que $e_i = \hat{u}_i$. Así pues, "a la hora de la verdad" nosotros estudiaremos la normalidad de los residuos, en tanto que éstos constituyen una estimación muestral de las perturbaciones.

Si denotamos por $\chi^2_{JB-\exp}$ el valor que toma el estadístico χ^2_{JB} para la serie de los residuos de la muestra que estamos considerando; y por $\chi^2_{2,1-\alpha}$ el valor teórico del mismo para un nivel de significación α, entonces tendremos que:

- Si $\chi^2_{JB-\exp} < \chi^2_{2,1-\alpha} \Rightarrow$ nos situaríamos en la <u>región de aceptación</u> (RA) y, por tanto, no habría evidencias para rechazar la hipótesis de normalidad de las perturbaciones; esto es, asumiríamos que éstas siguen una distribución normal.
- Si $\chi^2_{JB-\exp} > \chi^2_{2,1-\alpha} \Rightarrow$ nos encontraríamos en la <u>región crítica</u> (RC) y rechazaríamos la hipótesis nula de normalidad de las perturbaciones.

Gráficamente puede verse esto en la *Figura 1*.

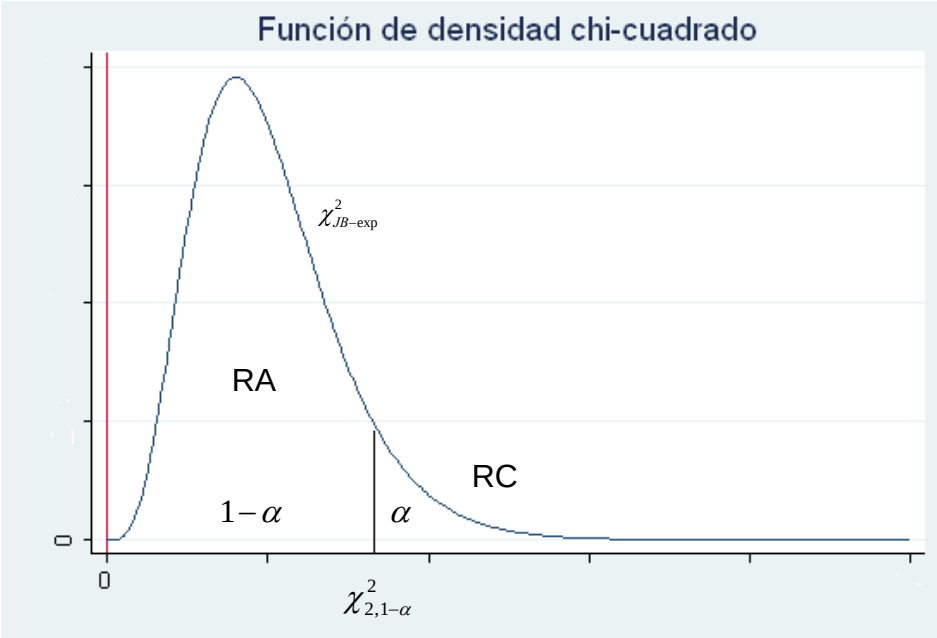


Figura 1

 ISBN: 978-84-694-7251-4

3.2. Intervalos de confianza.-

Intervalo de confianza para los coeficientes de regresión (β_j)

Una vez obtenida, mediante MCO, la estimación del vector de parámetros del modelo de regresión $\hat{\beta}$, y para valorar si ésta resulta ser una "aproximación" adecuada de los parámetros poblacionales β, podríamos en primera instancia atender a las propiedades que posee este estimador calculado por el citado método: es ELIO (esto es, resulta ser lineal, insesgado y de mínima varianza, como ya se ha estudiado).

Una forma adicional de valorar la precisión de la estimación consiste en establecer un intervalo de confianza: un intervalo de valores dentro del cual consideramos que se encuentran los parámetros poblacionales β con un determinado nivel de confianza[1].

Recordemos que el vector de estimadores $\hat{\beta}$ es un vector aleatorio que sigue una distribución normal multivariante. En particular: $\hat{\beta} \to N_k\left(\beta; \sigma_u^2 \cdot (X'X)^{-1}\right)$.

De este modo, la distribución de cada uno de los coeficientes de regresión estimados que conforma este vector ($\hat{\beta}_j$, $\forall j = 1,2,...,k$) es: $\hat{\beta}_j \to N\left(\beta_j; Var(\hat{\beta}_j)\right)$.

$Var(\hat{\beta}_j)$ es un elemento de la matriz de varianzas-covarianzas de $\hat{\beta}$; en concreto, de su diagonal principal:

$$Var - Cov(\hat{\beta}) = \sigma_u^{\,2} \cdot (X'X)^{-1} = \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1,\hat{\beta}_2) & Cov(\hat{\beta}_1,\hat{\beta}_3) & \dots & Cov(\hat{\beta}_1,\hat{\beta}_k) \\ & Var(\hat{\beta}_2) & Cov(\hat{\beta}_2,\hat{\beta}_3) & \dots & Cov(\hat{\beta}_2,\hat{\beta}_k) \\ & & \ddots & \ddots & \vdots \\ & & & & \vdots \\ & & Var(\hat{\beta}_j) & & \\ & & & \ddots & \vdots \\ & & & & Var(\hat{\beta}_k) \end{pmatrix} =$$

$$= \sigma_u^{\,2} \cdot \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & a_{1k} \\ & a_{22} & a_{23} & \dots & \dots & a_{2k} \\ & & \ddots & & & \vdots \\ & & & a_{jj} & & \vdots \\ & & & & \ddots & \\ & & & & & a_{kk} \end{pmatrix}.$$

[1] Podemos recordar brevemente el concepto de intervalo de confianza: supongamos que $\hat{\vartheta}$ es el estimador puntual de ϑ. Nuestro objetivo será determinar qué valores conforman el intervalo $\left(\hat{\vartheta} - \varepsilon, \hat{\vartheta} + \varepsilon\right)$, de tal forma que la probabilidad de que contenga a ϑ sea $1-\alpha$ (nivel de confianza).
Si nos fijamos, la estimación es el centro o *pivote* del intervalo y ε es un número positivo, es el radio de dicho intervalo, que sumado y restado al valor central configura finalmente la amplitud del intervalo. El valor de ε va a depender del nivel de confianza.

Así, podremos escribir que: $\hat{\beta}_j \to N\left(\beta_j ; \sigma_u^2 . a_{jj}\right)$, siendo a_{jj} el elemento j correspondiente de la diagonal principal de la matriz $(X'X)^{-1}$.

Nuestro objetivo en este apartado es establecer un intervalo de confianza de β_j a partir de su estimador $\hat{\beta}_j$. Si nos fijamos en la expresión anterior, vemos que en la caracterización de la distribución de probabilidad normal de $\hat{\beta}_j$ figura el parámetro poblacional, y por tanto desconocido, σ_u^2. Esto supone un problema. En las siguientes líneas vamos a proceder a realizar una serie de transformaciones que concluirán finalmente con una nueva expresión donde aparecerá la estimación del mismo, lo cual nos permitirá poder trabajar con ella. Veremos, además, cómo pasaremos de una distribución normal a una distribución *t*-Student.

Consideremos la última expresión que hemos expuesto:

$$\hat{\beta}_j \to N\left(\beta_j ; \sigma_u^2 . a_{jj}\right).$$

Tipificando esta variable aleatoria, tendríamos que:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma_u{}^2 . a_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{ES(\hat{\beta}_j)} \to N(0,1),$$

donde $ES(\hat{\beta}_j)$ es el error estándar de $\hat{\beta}_j$; esto es, su desviación típica: la raíz cuadrada positiva de la varianza de $\hat{\beta}_j$.

Por otro lado, tomemos en consideración también el siguiente estadístico, ya conocido, referido a σ_u^2:

$$\frac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k) \to \chi_{n-k}^2 .$$

A partir de los dos últimos estadísticos podemos generar un nuevo estadístico que seguiría una distribución de probabilidad *t*-Student, con $n-k$ grados de libertad[2]; en concreto, dividiendo el estadístico normal entre la raíz del cociente del estadístico chi-cuadrado entre sus grados de libertad:

[2] Si atendemos a la definición de una variable aleatoria que sigue una distribución *t*-Student, tenemos: que si se toman dos variables aleatorias independientes, Z y V, de modo que $Z \to N(0,1)$ y $V \to \chi_s^2$; entonces, la variable aleatoria $T = \dfrac{Z}{\sqrt{\dfrac{V}{s}}}$ se distribuye según una *t-Student con s grados de libertad.*

Se representa por: $T \to t_s$.

$$\frac{\dfrac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma_u{}^2.a_{jj}}}}{\sqrt{\dfrac{\dfrac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k)}{n-k}}}=\frac{\dfrac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma_u{}^2.a_{jj}}}}{\sqrt{\dfrac{\hat{\sigma}_u^2}{\sigma_u^2}}}=\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}_u{}^2.a_{jj}}}=\frac{\hat{\beta}_j-\beta_j}{\widehat{ES(\hat{\beta}_j)}}.$$

La expresión final de este estadístico la vamos a denotar por t_{β_j}, de modo pues que:

$$\boxed{t_{\beta_j}=\frac{\hat{\beta}_j-\beta_j}{\widehat{ES(\hat{\beta}_j)}}\to t_{n-k}.}$$

Obsérvese que la diferencia entre este estadístico y el de la normal tipificada expuesta anteriormente es que en su denominador aparece ahora no el error estándar de $\hat{\beta}_j$, sino su estimación, y ello es debido a que ya no se considera el parámetro poblacional σ_u^2, sino su estimación: $\hat{\sigma}_u^2$. De esta manera, cuando trabajemos con $\hat{\beta}_j$, no lo haremos con la distribución de probabilidad normal (que es la propia suya), sino con la *t*-Student, ya que así evitaremos tratar con un parámetro poblacional desconocido.

La *Figura 2* muestra cómo es una función de densidad de una distribución *t*-Student: una función de "dos colas" y simétrica, siendo el eje de simetría el 0.

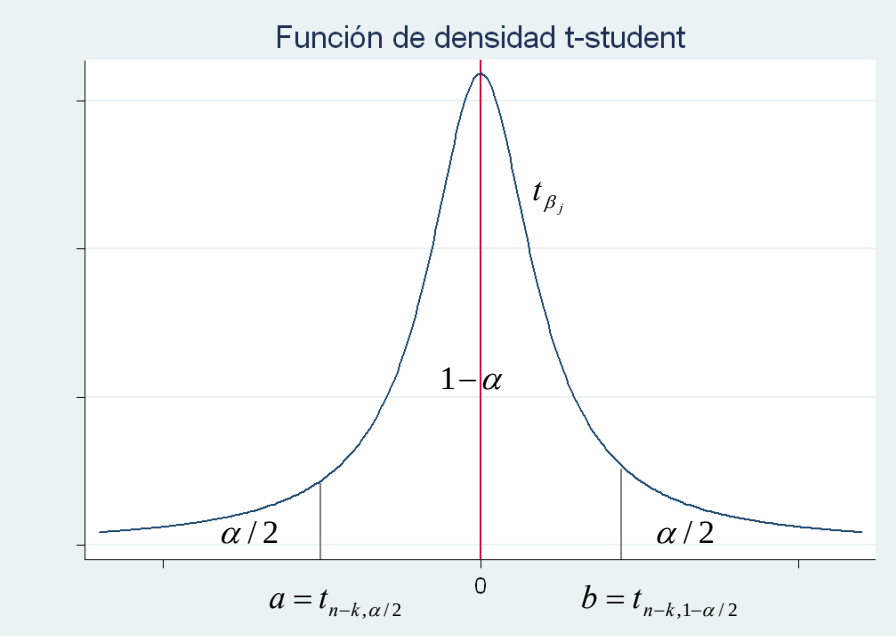


Figura 2

A partir de aquí, podemos deducir un intervalo de confianza para β_j, $\forall j=1,2,...,k$.

Como se puede ver en la *Figura 2*: $P\left[a \le t_{\beta_j} \le b\right] = 1-\alpha$, donde α es el nivel de significación y $a = -b$, dada la simetría de la función de densidad de la distribución *t*-Student. Desarrollando esta expresión, tenemos entonces que:

$$P\left[a \le t_{\beta_j} \le b\right] = 1-\alpha \quad \Leftrightarrow \quad P\left[-b \le t_{\beta_j} \le b\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[-t_{n-k,1-\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{\widehat{ES(\hat{\beta}_j)}} \le t_{n-k,1-\alpha/2}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[-t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)} \le \hat{\beta}_j - \beta_j \le t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[-\hat{\beta}_j - t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)} \le -\beta_j \le -\hat{\beta}_j + t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\hat{\beta}_j + t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)} \ge \beta_j \ge \hat{\beta}_j - t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\hat{\beta}_j - t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)} \le \beta_j \le \hat{\beta}_j + t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right] = 1-\alpha\,.$$

Si nos fijamos, llegados a este punto ya tenemos definido el intervalo de confianza de β_j, para un nivel de significación α:

$$\left(\hat{\beta}_j \pm t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right); \text{ es decir: } \boxed{\left(\hat{\beta}_j - t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)},\ \hat{\beta}_j + t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right)}$$

Esto viene a decir que el parámetro poblacional β_j se encuentra contenido en este intervalo[3] con un nivel de confianza cifrado en el $[(1-\alpha).100]\%$.

[3] Obsérvese cómo el centro o pivote del intervalo es la estimación $\hat{\beta}_j$ del parámetro poblacional. Por su parte, el radio del mismo viene dado por:

$$\varepsilon = t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\,,$$

que al sumarse y restarse a $\hat{\beta}_j$, determinan finalmente el intervalo de confianza.

 ISBN: 978-84-694-7251-4

Intervalo de confianza para la varianza de la perturbación aleatoria (σ_u^2)

El otro parámetro de interés en el modelo de regresión es la varianza de la perturbación aleatoria, σ_u^2, cuya estimación puntual ya se ha visto con anterioridad y para el que, de igual modo, podemos establecer un intervalo de confianza, esta vez basándonos directamente en la distribución de probabilidad chi-cuadrado que tiene asociada:

$$\boxed{\chi^2_{\sigma_u^2} = \frac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k) \rightarrow \chi^2_{n-k}}.$$

En la *Figura 3* se muestra gráficamente la distribución de este estadístico:

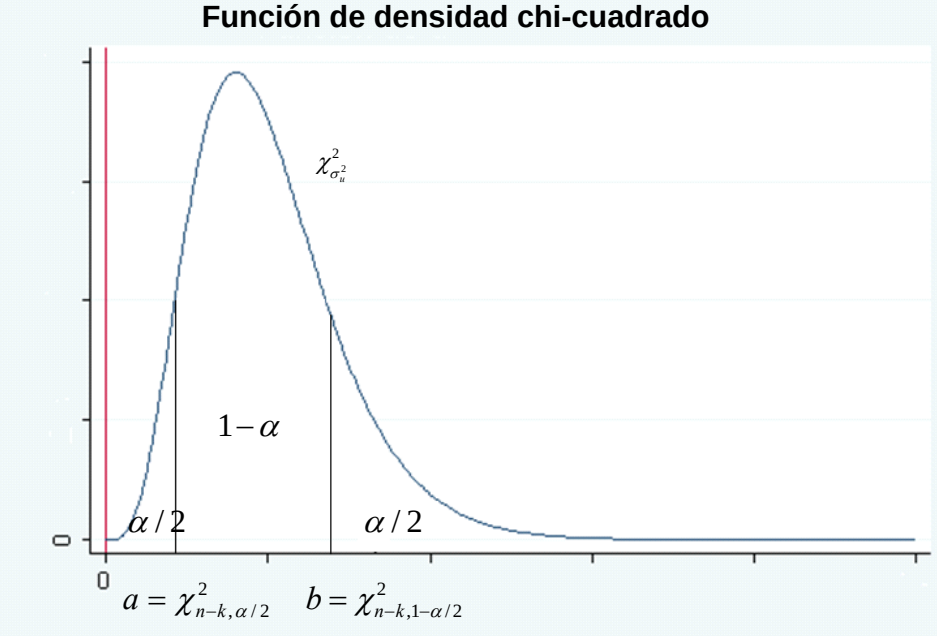


Figura 3

De aquí se deduce que:

$$P\left[a \le \chi^2_{\sigma_u^2} \le b\right] = 1-\alpha \quad \Leftrightarrow \quad P\left[\chi^2_{n-k,\alpha/2} \le \chi^2_{\sigma_u^2} \le \chi^2_{n-k,1-\alpha/2}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\chi^2_{n-k,\alpha/2} \le \frac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k) \le \chi^2_{n-k,1-\alpha/2}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\frac{\chi^2_{n-k,\alpha/2}}{\hat{\sigma}_u^2.(n-k)} \le \frac{1}{\sigma_u^2} \le \frac{\chi^2_{n-k,1-\alpha/2}}{\hat{\sigma}_u^2.(n-k)}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,\alpha/2}} \ge \sigma_u^2 \ge \frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,1-\alpha/2}}\right] = P\left[\frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,1-\alpha/2}} \le \sigma_u^2 \le \frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,\alpha/2}}\right] = 1-\alpha\,.$$

Así pues, el intervalo de confianza de σ_u^2, para un nivel de significación α, es:

$$\left(\frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,1-\alpha/2}}, \frac{\hat{\sigma}_u^2.(n-k)}{\chi^2_{n-k,\alpha/2}} \right).$$

El significado de este intervalo es análogo al del desarrollado anteriormente para el caso de los coeficientes de regresión del modelo.

Para concluir, puesto que: $\hat{\sigma}_u^2 = \frac{SCR}{n-k} \Rightarrow SCR = \hat{\sigma}_u^2.(n-k)$. De esta forma, el intervalo de σ_u^2 también puede expresarse de manera que:

$$\left(\frac{SCR}{\chi^2_{n-k,1-\alpha/2}}, \frac{SCR}{\chi^2_{n-k,\alpha/2}} \right).$$

3.3. Contraste de significatividad individual de las variables explicativas, de significatividad global del modelo y general de un conjunto de restricciones lineales. El modelo restringido.-

Volvemos a centrarnos en este apartado en los contrastes de hipótesis, todos ellos referidos a los coeficientes de regresión del modelo.

Contraste de significatividad individual de las variables explicativas

Tómese en consideración el modelo econométrico clásico de regresión lineal múltiple:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n.$$

En este punto vamos a estudiar si la variable X_j resulta significativa o no en el modelo; es decir, si realmente dicha variable explica, al menos parte, del comportamiento de la variable *Y* a nivel poblacional con un determinado nivel de error.

Para ello, planteamos la realización del siguiente contraste de hipótesis sobre el coeficiente de regresión β_j que lo acompaña:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

- Si se acepta $H_0 \Rightarrow X_j$ no sería relevante o significativa a la hora de explicar la variable *Y*, ya que desaparecería de la especificación del modelo.
- Si se rechaza $H_0 \Rightarrow X_j$ sería significativa, pues permanecería en el modelo.

 ISBN: 978-84-694-7251-4

Para llevar a cabo el contraste necesitamos un estadístico de prueba, que obtendremos partiendo del estadístico t_{β_j} que vimos en el apartado anterior:

$$t_{\beta_j} = \frac{\hat{\beta}_j - \beta_j}{\widehat{ES(\hat{\beta}_j)}} \rightarrow t_{n-k}\,.$$

Si asumimos como cierta la hipótesis nula ($\beta_j = 0$), y con los datos de la muestra que estemos considerando, obtendremos un valor concreto para dicho estadístico: el estadístico experimental, que adoptaría la forma[4]:

$$\boxed{t^{\exp} = \frac{\hat{\beta}_j}{\widehat{ES(\hat{\beta}_j)}} \rightarrow t_{n-k}}\,.$$

En la *Figura 4* se puede ver la distribución de este estadístico, mostrándose asimismo la región de aceptación (RA) y la región crítica (RC) (que en este caso estaría conformada por dos áreas simétricas separadas, dado que estamos ante una distribución *t*-Student), que vendrían delimitadas por los valores críticos $t^*_{n-k,1-\alpha/2}$ y $-t^*_{n-k,1-\alpha/2}$. Éstos últimos se establecerían a partir de los grados de libertad del modelo estudiado ($n-k$) y del nivel de significación α considerado.

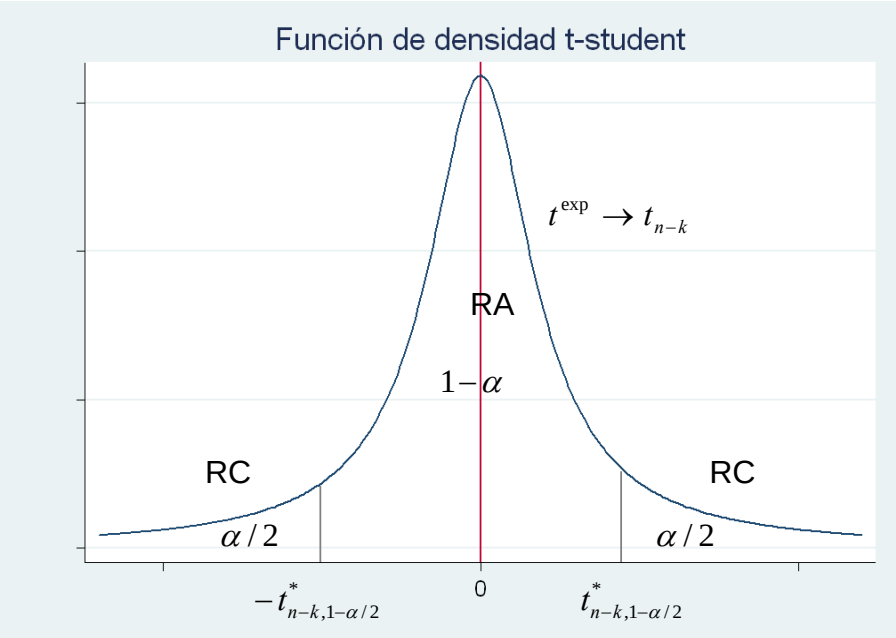


Figura 4

[4] Nótese que el estadístico de prueba podría igualmente utilizarse para cualquier otra hipótesis nula que quisiésemos contrastar en relación a un valor concreto β_0 del parámetro β_j; esto es: $H_0 : \beta_j = \beta_0$. Tenida en cuenta la información muestral del caso concreto analizado, el estadístico experimental sería:

$$\frac{\hat{\beta}_j - \beta_0}{\widehat{ES(\hat{\beta}_j)}}\,.$$

Según esto, tendremos:

- Si $\left| t^{\exp} \right| < t_{n-k,1-\alpha/2} \Rightarrow$ nos encontraríamos en la región de aceptación y por tanto no encontraríamos evidencia para rechazar la hipótesis nula. Podríamos, por tanto, considerar que la variable X_j no es significativa en el modelo.
- Si $\left| t^{\exp} \right| > t_{n-k,1-\alpha/2} \Rightarrow$ nos encontraríamos en la región crítica y rechazaríamos, consecuentemente, la hipótesis nula. La variable X_j sería, pues, significativa.

Para finalizar, podemos reseñar que, si se ha establecido el mismo nivel de significación α, un intervalo de confianza de β_j nos ofrece un rango de valores para dicho parámetro asumibles como hipótesis nulas, sin necesidad de llevar a cabo explícitamente los contrastes correspondientes.

Contraste de significatividad global del modelo

Consideremos nuevamente el modelo econométrico clásico de regresión lineal múltiple:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n\,.$$

El contraste de significatividad global de un modelo pretende analizar si las variables independientes son capaces de explicar de manera conjunta el comportamiento de la variable *Y*. En este caso, las hipótesis nula y alternativa se expresan como sigue:

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ &\quad \beta_3 = 0 \\ &\quad ..... \\ &\quad \beta_k = 0 \\ H_1 &: \text{algún } \beta_j \neq 0, \quad j = 2,...,k \end{aligned}$$

- Si se acepta $H_0 \Rightarrow$ ninguna variable explicativa del modelo sería relevante; es decir, "no habría" modelo: no sería significativo.
- Si se rechaza $H_0 \Rightarrow$ tendríamos que al menos una variable explicativa del modelo sería relevante; es decir, "habría" modelo: éste sería significativo.

Con la información muestral del modelo que se estudie, el estadístico de prueba experimental que se utiliza en este contraste es:

$$F^{\exp} = \frac{SCE/k-1}{SCR/n-k} = \frac{\frac{SCE/SCT}{k-1}}{\frac{SCR/SCT}{n-k}} = \frac{R^2/k-1}{(1-R^2)/n-k} \rightarrow F_{k-1,n-k}.$$

 ISBN: 978-84-694-7251-4

Según puede observarse, este estadístico sigue una distribución de probabilidad F de Fisher-Snedecor. En la *Figura 5* se muestra su función de densidad (positiva y de una única cola), así como la región de aceptación (RA) y la región crítica (RC), que están delimitadas por el valor crítico $F^{*\ 1-\alpha}_{k-1,n-k}$. Este valor se fija a partir del número de ecuaciones que tiene la hipótesis nula (que en este caso es $k-1$), de los grados de libertad del modelo ($n-k$) y del nivel de significación α que se establezca.

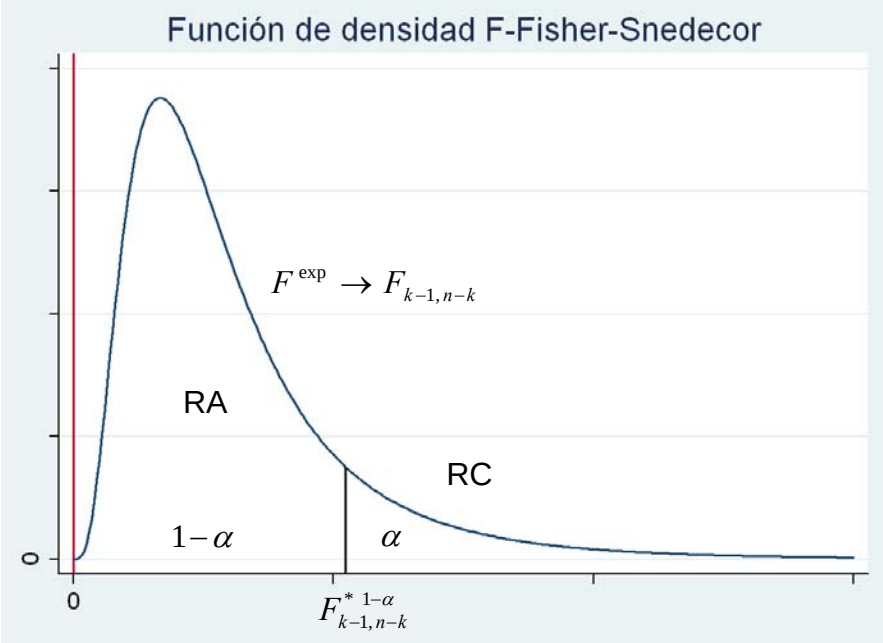


Figura 5

De acuerdo con esto:

- Si $F^{\exp} < F^{*\ 1-\alpha}_{k-1,n-k} \Rightarrow$ estaríamos en la región de aceptación y, en consecuencia, el modelo no sería significativo en su conjunto.
- Si $F^{\exp} > F^{*\ 1-\alpha}_{k-1,n-k} \Rightarrow$ nos hallaríamos en la región crítica. En este caso, el modelo sí sería globalmente significativo.

Contraste general de un conjunto de restricciones lineales

Los dos contrastes que hasta el momento se han visto en este apartado sobre los coeficientes de regresión del modelo, el de significatividad individual de los parámetros y el de significatividad global del modelo, no dejan de ser dos contrastes particulares de la generalidad de contrastes que pueden plantearse para dichos coeficientes.

En efecto, nuestro interés puede centrarse en comprobar si se cumplen una serie de relaciones (siempre lineales) entre los coeficientes, que pueden venir dadas por 1 ecuación, o por más de 1.

Sea el modelo econométrico clásico de regresión lineal múltiple:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, 2, ..., n .$$

Las hipótesis de nuestro contraste podrían expresarse matricialmente, con carácter general, de la forma:

$$\boxed{\begin{array}{l} H_0 : R\beta = r \\ H_1 : R\beta \neq r \end{array}}$$

donde $R\beta$ recoge combinaciones lineales de los parámetros β_j que conforman el vector β, siendo R la matriz de coeficientes reales de dichas combinaciones; por su parte, *r* es una matriz-columna de números reales.

En este caso general, el <u>estadístico de prueba experimental</u> que se utiliza es:

$$\boxed{F^{\exp} = \frac{(R\hat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r) \Big/ q}{SCR \Big/ n-k} \rightarrow F_{q,n-k}}.$$

Como sabemos, $n-k$ son los grados de libertad del modelo y q hace referencia al número de ecuaciones o restricciones que forman la hipótesis nula a contrastar (número de filas de R), debiendo ser linealmente independientes y verificar que: $q \leq k$.

Obsérvese que, si desarrollamos este estadístico de prueba, éste puede expresarse de forma alternativa como sigue:

$$F^{\exp} = \frac{(R\hat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r) \Big/ q}{SCR \Big/ n-k} = \frac{(R\hat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r) \Big/ q}{\hat{\sigma}_u^2} =$$

$$= \frac{(R\hat{\beta} - r)'\left[R\hat{\sigma}_u^2 \cdot (X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r)}{q} = \frac{(R\hat{\beta} - r)'\left[R\,\widehat{Var - Cov}(\hat{\beta})\,R'\right]^{-1}(R\hat{\beta} - r)}{q}.$$

En este punto merece también hacer alusión al denominado <u>estadístico de Wald</u> (que es utilizado por el programa *EViews* para llevar a cabo este tipo de contrastes). El estadístico $F_{q,n-k}$ se construye a partir del cociente de dos variables aleatorias independientes, cuyas distribuciones son del tipo chi-cuadrado con q y $n-k$ grados de libertad, respectivamente. Así:

$$W = (R\hat{\beta} - r)'\left[R\sigma_u^2 \cdot (X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r) \rightarrow \chi_q^2 \qquad \text{y} \qquad \frac{SCR}{\sigma_u^2} \rightarrow \chi_{n-k}^2 ,$$

donde *W* es el estadístico de Wald.

 ISBN: 978-84-694-7251-4

Si nos fijamos, en el estadístico *W* la varianza de las perturbaciones σ_u^2 es desconocida, por lo que suele emplearse para realizar el contraste de hipótesis (y así lo hace *EViews*) la distribución límite del estadístico de *Wald*. Ésta se puede definir de igual modo, sólo que utilizando la estimación insesgada de la varianza de las perturbaciones por MCO, esto es:

$$\hat{W} = (R\hat{\beta} - r)'\left[R\hat{\sigma}_u^2 \cdot (X'X)^{-1} R'\right]^{-1}(R\hat{\beta} - r) \to \chi_q^2 .$$

En definitiva, si como hemos visto, el estadístico experimental *F* de Fisher-Snedecor se podía escribir como:

$$F^{\exp} = \frac{(R\hat{\beta} - r)'\left[R\hat{\sigma}_u^2 \cdot (X'X)^{-1} R'\right]^{-1}(R\hat{\beta} - r)}{q} \to F_{q,n-k},$$

entonces: $F^{\exp} = \dfrac{\hat{W}}{q}$; o lo que es lo mismo: $\boxed{\hat{W} = q \cdot F^{\exp}}$.

Los siguientes ejemplos pueden ayudarnos a entender bien quiénes son los distintos elementos que intervienen en el estadístico de prueba de este contraste:

a) Supongamos el siguiente modelo: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$.

   La hipótesis nula[5] a contrastar podría ser: $H_0 : \beta_2 = \beta_3$, que si la "rehacemos" para que queden despejados los parámetros β_j, quedaría: $H_0 : \beta_2 - \beta_3 = 0$.

   En este ejemplo, tenemos que $q = 1$ y H_0 se podría expresar matricialmente de la forma:

$$(0 \quad 1 \quad -1)\cdot\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix} = 0, \quad \text{siendo:} \quad \begin{matrix} R = (0 \quad 1 \quad -1) \\ r = (0)\end{matrix}$$

b) Planteemos este otro modelo: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$.

   En este caso, la hipótesis nula a contrastar podría estar conformada por 2 restricciones, esto es, $q = 2$, siendo:

$$\begin{matrix} H_0 : \beta_2 = 3\beta_3 + \beta_4 \\ \beta_1 = 2\beta_5 + 3 \end{matrix}$$

   Si despejamos los parámetros β_j hacia el miembro de la izquierda de las ecuaciones, tenemos:

[5] La hipótesis alternativa, lógicamente, sería el incumplimiento de la hipótesis nula.

$$H_0 : \beta_2 - 3\beta_3 - \beta_4 = 0$$
$$\beta_1 - 2\beta_5 = 3$$

La hipótesis nula se podría entonces expresar matricialmente del siguiente modo:

$$\begin{pmatrix} 0 & 1 & -3 & -1 & 0 \\ 1 & 0 & 0 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \quad \text{donde:} \quad R = \begin{pmatrix} 0 & 1 & -3 & -1 & 0 \\ 1 & 0 & 0 & 0 & -2 \end{pmatrix} \quad r = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

La *Figura 6* ofrece la función de densidad del estadístico de prueba experimental, junto con la región de aceptación (RA) y la región crítica (RC), las cuales vienen delimitadas por el valor crítico $F^{*\,1-\alpha}_{q,n-k}$. Éste se obtiene a partir del número de restricciones que configura la hipótesis nula (que hemos denominado de manera general *q*), de los grados de libertad del modelo ($n-k$) y del nivel de significación α que se considere.

Así:

- Si $F^{\exp} < F^{*\,1-\alpha}_{q,n-k} \Rightarrow$ aceptaríamos la hipótesis nula que hubiésemos establecido en relación a los coeficientes de regresión del modelo $R\beta = r$.
- Si $F^{\exp} > F^{*\,1-\alpha}_{q,n-k} \Rightarrow$ rechazaríamos dicha hipótesis nula.

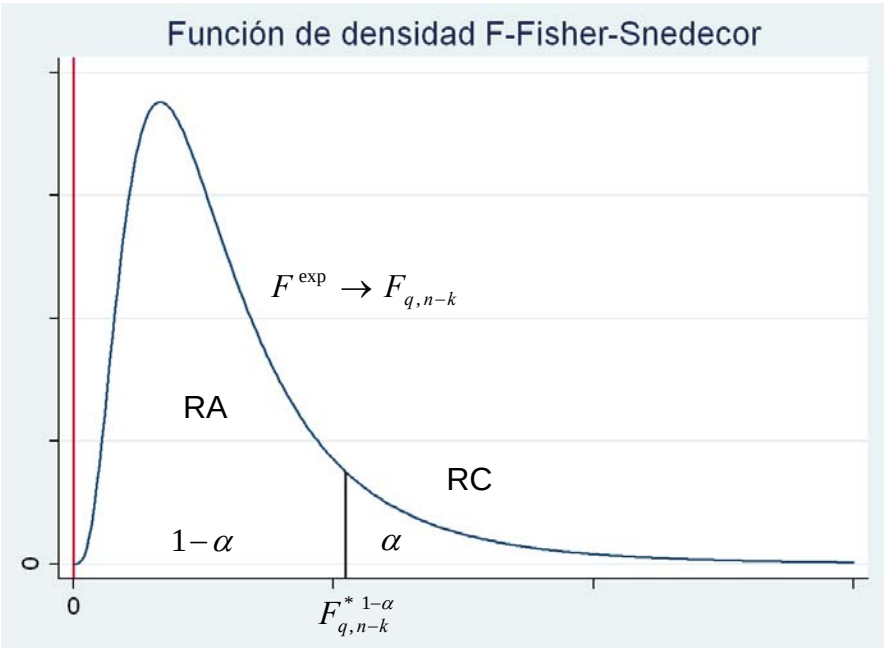


Figura 6

Para finalizar este punto, debemos resaltar que este estadístico que aquí hemos expuesto sirve para contrastar cualquier hipótesis lineal que podamos plantearnos, incluidos los contrastes de significatividad individual de los coeficientes de regresión[6] y el de significatividad global del modelo que, como ya señalamos al principio, pueden verse simplemente como casos particulares del caso general.

El modelo restringido

En este punto abordamos el estudio de los contrastes de hipótesis sobre los coeficientes de regresión del modelo desde otra perspectiva. En particular, comparamos dos modelos: uno, el original; y otro, que llamamos *modelo restringido*, que es aquél que asume como cierta la hipótesis nula y la incorpora en su propia definición.

Por ejemplo, sea el siguiente modelo: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$.

Supongamos que nos planteamos contrastar para el mismo esta hipótesis nula:

$$H_0 : \beta_4 = 0$$
$$\beta_5 = 0$$

En este caso, el modelo: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, sería el modelo restringido, ya que, si nos fijamos ha incluido en su definición las 2 restricciones indicadas por la hipótesis nula.

Por su parte, el *modelo original* lo podríamos denominar modelo no restringido.

Si abordásemos el análisis de ambos modelos a partir de este punto, podríamos obtener en cada uno de ellos toda su información característica; en particular: sus coeficientes estimados ($\hat{\beta}$ y $\hat{\beta}_r$), sus sumas de cuadrados residuales (*SCR* y *SCR*$_r$) o sus coeficientes de determinación (R^2 y R_r^2), donde el subíndice *r* hace referencia al modelo restringido frente al original o no restringido.

En este caso, podría comprobarse que el estadístico *F* de contraste general adopta las siguientes expresiones:

$$F^{\exp} = \frac{(R\hat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r)/q}{SCR/n-k} = \frac{(SCR_r - SCR)/q}{SCR/n-k} = \frac{\left(R^2 - R_r^2\right)/q}{(1-R^2)/n-k} \to F_{q,n-k}$$

[6] En el caso del contraste de significatividad individual de un parámetro β_j, si partiésemos del estadístico *F* de prueba general, éste terminaría adoptando la forma: $F^{\exp} = \frac{\hat{\beta}_j^2}{\widehat{Var}(\hat{\beta}_j)} \to F_{1,n-k}$, el cual, si nos fijamos, coincide con el cuadrado del estadístico de prueba *t*; esto es: $F^{\exp} = \left(t^{\exp}\right)^2$.

Al igual que en los contrastes anteriores, debemos nuevamente tener en cuenta la función de densidad de la distribución *F* de Fisher-Snedecor (*Figura 7*).

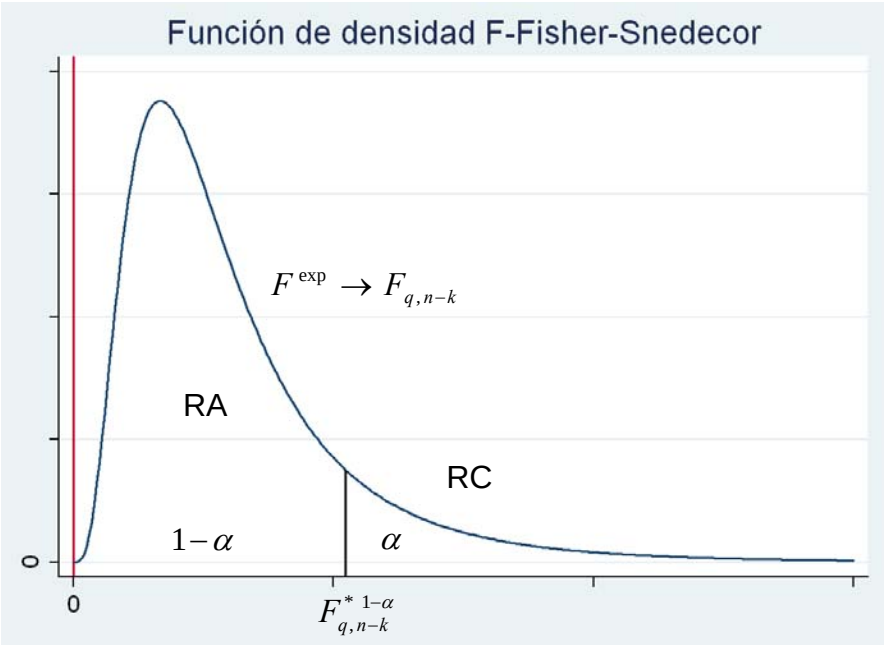


Figura 7

A la vista de esto:

- Si $F^{\exp} < F^{*\,1-\alpha}_{q,n-k} \Rightarrow$ nos encontraríamos en la región de aceptación y, por tanto, aceptaríamos la hipótesis nula que hubiésemos establecido en relación a los coeficientes de regresión del modelo.
- Si $F^{\exp} > F^{*\,1-\alpha}_{q,n-k} \Rightarrow$ estaríamos en la región crítica y rechazaríamos, pues, la hipótesis nula.

3.4. Contraste de Chow de cambio estructural.-

Un contraste de especial interés por la frecuencia con la que aparece en aplicaciones empíricas es el que se utiliza para analizar si bajo un conjunto de datos subyace una única estructura económica o modelo, o si por el contrario, se puede considerar que es divisible en dos o más submuestras y que cada una de ellas ha sido generada por estructuras distintas.

Este contraste se conoce como *test de Chow*, test de cambio estructural, o contraste de estabilidad de los parámetros y pretende analizar la hipótesis nula de ausencia de cambio estructural.

En el caso de que se esté trabajando con datos de series temporales, se utiliza habitualmente cuando se cuenta con información acerca de algún acontecimiento

relevante que se piensa que puede provocar una variación estructural en un momento del periodo muestral considerado y que, por tanto, tiene capacidad suficiente para afectar a los parámetros o coeficientes del modelo. Sería el caso, por ejemplo, de analizar cómo puede verse afectado un modelo de consumo en España entre los años 1990 y 2010 ante la entrada del euro en 2002; es decir, si este hecho supuso un cambio en la estructura de consumo.

Este contraste se utiliza también frecuentemente con datos de corte transversal; en esta ocasión, para comparar dos o más grupos de la muestra. Por ejemplo, si se está analizando un modelo que explique el salario y se quiere estudiar si su comportamiento es el mismo para hombres y mujeres o si, por el contrario, el modelo debe ser diferente para uno y otro caso. O, por poner otro ejemplo, en el caso de un modelo de consumo, si se quisiera comparar si su estructura es la misma para residentes en zona urbana o no urbana, etc.

Así pues, el planteamiento de este tipo de contraste sería el siguiente:

H_0 : Ausencia de cambio estructural
H_1 : Existencia de cambio estructural

Si nos fijamos, la hipótesis nula contempla el caso en que un mismo modelo subyace bajo el total de los datos de la muestra. Por el contrario, la hipótesis alternativa asumiría que serían precisas distintas estimaciones del modelo para cada uno de los periodos temporales o divisiones transversales (según el tipo de datos) considerados de la muestra.

Es decir, supongamos el siguiente modelo general:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, ..., n$$

Y queremos ver si (pensando que son datos temporales), ocurre en un momento determinado un cierto hecho que nos lleva a plantearnos si se produce un cambio en la estructura de dicho modelo, de forma que el problema analizado se explica mejor mediante 2 modelos diferentes: uno para la etapa previa al hecho en cuestión y otro correspondiente a la etapa posterior de dicho hecho.

La aceptación de la hipótesis nula implica que el modelo original sería válido para todas las observaciones *i* del periodo muestral completo *n*.

Por el contrario, la hipótesis alternativa supone considerar dos subperiodos en la muestra (n_1 y n_2, de forma que: $n = n_1 + n_2$), cada uno de los cuales ha sido generado por un modelo distinto. Tendríamos así:

- Subperiodo 1

$$Y_i = \beta_1^* + \beta_2^* X_{2i} + ... + \beta_j^* X_{ji} + ... + \beta_k^* X_{ki} + u_i^* \qquad \forall j = 1, 2, ..., k \qquad \forall i = 1, ..., n_1$$

- Subperiodo 2

$$Y_i = \beta_1^{**} + \beta_2^{**} X_{2i} + ... + \beta_j^{**} X_{ji} + ... + \beta_k^{**} X_{ki} + u_i^{**} \qquad \forall j = 1,2,...,k \qquad \forall i = n_1 + 1,...,n$$

Como se expuso en el apartado anterior, el modelo en el que se considera cierta la hipótesis nula sería el modelo "restringido" (en este caso, el modelo original, válido para toda la muestra); por su parte, los modelos relativos a los dos superiodos, y que se corresponden con la hipótesis alternativa, constituirían la especificación "sin restringir". De este modo, podríamos utilizar el estadístico *F* relativo a la realización de contrastes desde la óptica del modelo restringido, de manera que tendríamos:

$$F^{\exp} = \frac{\left[SCR_r - \left(SCR_1 + SCR_2\right)\right] \Big/ k}{\left(SCR_1 + SCR_2\right) \Big/ n - 2k} \rightarrow F_{k,n-2k},$$

donde SCR_r es la suma de cuadrados residuales que corresponde al modelo de un único periodo muestral y SCR_1 y SCR_2 son las de los subperiodos 1 y 2, respectivamente.[7]

Para finalizar, en el caso de que la hipótesis alternativa considere un número genérico de h subperiodos o subdivisiones muestrales, en lugar de sólo 2, la expresión general del test de Chow sería:

$$F^{\exp} = \frac{\left[SCR_r - \left(SCR_1 + ... + SCR_h\right)\right] \Big/ (h-1)k}{\left(SCR_1 + ... + SCR_h\right) \Big/ n - hk} \rightarrow F_{(h-1)k,n-hk}.$$

La forma de llevar a cabo el contraste sería análoga a la ya explicada para los estadísticos *F* ya vistos anteriormente en el presente Tema.

Es decir, atendiendo a la gráfica de la función de densidad de la distribución *F* de Fisher-Snedecor (*Figura 8*), tendríamos:

- Si $F^{\exp} < F^{*\ 1-\alpha}_{(h-1)k,n-hk} \Rightarrow$ estaríamos en la región de aceptación y asumiríamos que hay ausencia de cambio estructural; esto es, que el modelo planteado es válido para la totalidad de la muestra.
- Si $F^{\exp} > F^{*\ 1-\alpha}_{(h-1)k,n-hk} \Rightarrow$ nos hallaríamos en la región crítica y concluiríamos que hay un cambio estructural en nuestra muestra, que nos aconseja considerar diversos modelos, uno por cada subdivisión o periodo muestral tomado.

[7] Obsérvese que en este contraste la hipótesis nula contiene una ecuación por cada igualdad entre los coeficientes de regresión asociados a una misma variable en cada subperiodo; esto es: $q = k$. Igualmente, los grados de libertad asociados al denominador de este estadístico son: $n_1 - k$ para el caso del modelo del primer subperiodo y $n_2 - k$, para el segundo; por tanto, la suma de ambos hace que sea: $n - 2k$.

 ISBN: 978-84-694-7251-4

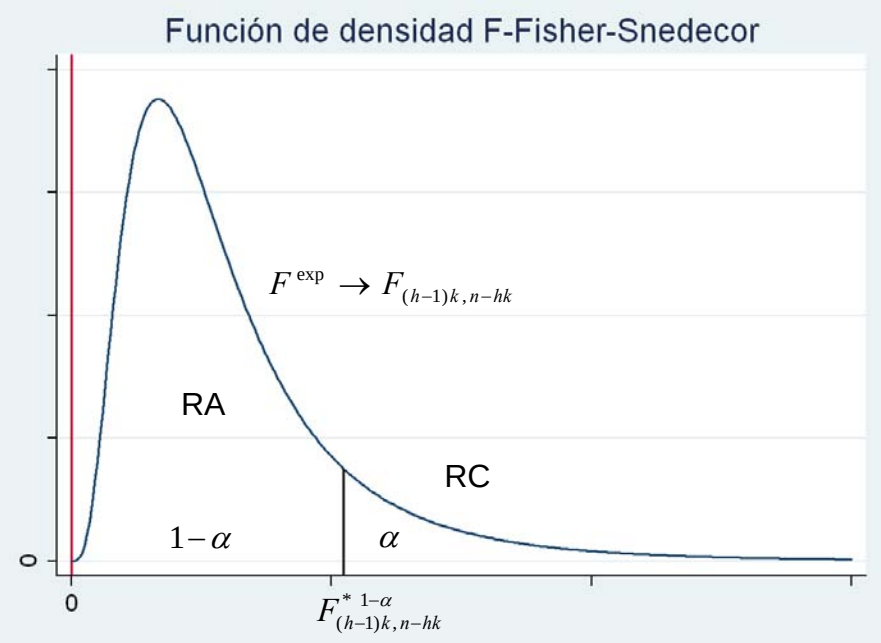


Figura 8

3.5. Predicción.-

Sea el modelo econométrico clásico de regresión lineal múltiple:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n\,.$$

Como sabemos, este modelo se puede expresar de manera matricial: $Y = X\beta + u$.

Mediante el método de MCO podemos estimar el modelo, por lo que tendremos:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n\,.$$

O bien, matricialmente: $\boxed{\hat{Y} = X\hat{\beta}}$.

A partir del modelo estimado, si dispusiésemos de una serie de valores concretos para todas y cada una de las variables explicativas X_j que conforman la matriz *X*, podríamos obtener una estimación del valor de la variable dependiente *Y*.

Si tales valores de las variables independientes fuesen extra-muestrales, es decir, valores distintos a los que integran la muestra objeto de estudio (que denotaremos de forma general $X_{j\,0}$), entonces dicha estimación de *Y* (que denotaremos por $\hat{Y}_0$) sería en realidad una predicción. Ésta podría plantearse tanto para series temporales (pensando en la obtención de un valor futuro de *Y*), como para datos transversales.

Así, si nos facilitan la serie de valores extra-muestrales de las variables independientes X_0, tendríamos:

$$\boxed{X_0 = \begin{pmatrix} X_{1\,0} \\ X_{2\,0} \\ \vdots \\ X_{k\,0} \end{pmatrix} = \begin{pmatrix} 1 \\ X_{2\,0} \\ \vdots \\ X_{k\,0} \end{pmatrix} \Rightarrow \quad \hat{Y}_0 = X_0^{'}\hat{\beta}}.$$

La estimación o predicción, sin embargo, no será completa si no viene acompañada de un intervalo de confianza. Para ello, se precisa introducir el concepto de error de predicción puntual.

El error de predicción puntual (e_0) se define como la diferencia entre el valor real de Y correspondiente a los valores dados de las variables independientes (Y_0) y la predicción obtenida a partir de los mismos ($\hat{Y}_0$); es decir:

$$\boxed{e_0 = Y_0 - \hat{Y}_0}.$$

Si desarrollamos esta expresión, tenemos que:

$$e_0 = Y_0 - \hat{Y}_0 = X_0^{'}\beta + u_0 - X_0^{'}\hat{\beta} = X_0^{'}(\beta - \hat{\beta}) + u_0\,,$$

de la cual podemos deducir cuáles son las fuentes del error de predicción, esto es, de dónde pueden provenir los errores o desviaciones que se produzcan a la hora de contrastar la realidad con la estimación. Tales fuentes son:

- Errores en la información de partida de las variables explicativas: X_0.
- Errores en la estimación de β: $\beta - \hat{\beta}$.
- Errores estocásticos procedentes de la perturbación aleatoria: u_0

Asimismo, también se puede ver que, puesto que e_0 depende de u_0 y ésta tiene naturaleza aleatoria, entonces e_0 es una variable aleatoria; además, dado que u_0 es normal, entonces e_0 sigue igualmente una distribución normal de probabilidad y, como tal, nos interesa conocer tanto su valor esperado como su varianza. Éstos resultan ser:

$$E[e_0] = 0 \qquad \text{y} \qquad Var(e_0) = \sigma_u^2 \cdot (1 + X_0^{'}(X'X)^{-1}X_0)\,.$$

Así pues, en definitiva: $\boxed{e_0 \to N\left(0; \sigma_u^2 \cdot (1 + X_0^{'}(X'X)^{-1}X_0)\right)}$

De aquí, podemos establecer un intervalo de confianza para la predicción, de forma similar al referido a los coeficientes de regresión del modelo, ya visto. En efecto, si tipificamos el error e_0, tendremos:

$$\frac{e_0}{\sqrt{\sigma_u^2 \cdot (1 + X_0^{'}(X'X)^{-1}X_0)}} = \frac{e_0}{ES(e_0)} \to N(0,1)\,.$$

 ISBN: 978-84-694-7251-4

Con ello, junto con el estadístico:

$$\frac{\hat{\sigma}_u^2}{\sigma_u^2}(n-k) \to \chi^2_{n-k},$$

podríamos generar un nuevo estadístico que seguiría una distribución de probabilidad *t*-Student, con $n-k$ grados de libertad:

$$\frac{e_0}{\sqrt{\sigma_u^2 \cdot (1+X_0'(X'X)^{-1}X_0)}} = \frac{e_0}{\sqrt{\hat{\sigma}_u^2 \cdot (1+X_0'(X'X)^{-1}X_0)}} = \frac{e_0}{\widehat{ES(e_0)}} \to t_{n-k}.$$

Este estadístico resultante vamos a denotarlo por t_{e_0}. En la *Figura 9* se representa la función de densidad que seguiría.

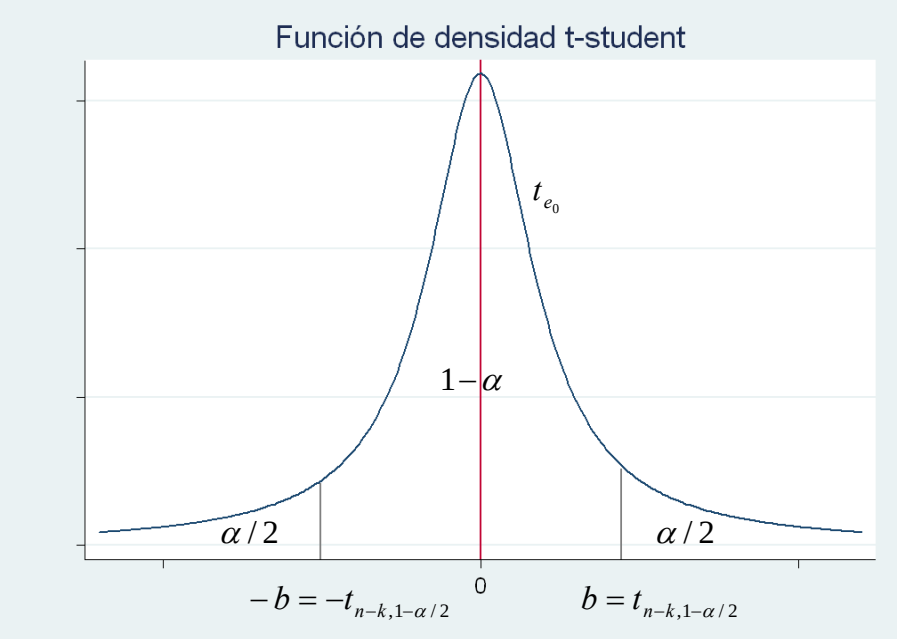


Figura 9

Partiendo de aquí, podemos deducir un intervalo de confianza para la predicción puntual Y_0:

$$P\left[-b \le t_{e_0} \le b\right] = 1-\alpha \quad \Leftrightarrow \quad P\left[-t_{n-k,1-\alpha/2} \le \frac{e_0}{\widehat{ES(e_0)}} \le t_{n-k,1-\alpha/2}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[-t_{n-k,1-\alpha/2} \le \frac{Y_0-\hat{Y}_0}{\widehat{ES(e_0)}} \le t_{n-k,1-\alpha/2}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[-t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)} \le Y_0-\hat{Y}_0 \le t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)}\right] = 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad P\left[\hat{Y}_0-t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)} \le Y_0 \le \hat{Y}_0+t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)}\right] = 1-\alpha\,.$$

En definitiva, el <u>intervalo de confianza de Y_0</u>, para un nivel de significación α es:

$$\left(\hat{Y}_0 \pm t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)}\right); \text{ es decir: } \boxed{\left(\hat{Y}_0 - t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)},\ \hat{Y}_0 + t_{n-k,1-\alpha/2}\cdot\widehat{ES(e_0)}\right)}.$$

Como sabemos, esto quiere decir que el valor real Y_0 que se corresponda con la observación de nuestra predicción se encuentra contenido en este intervalo con un nivel de confianza cifrado en el $[(1-\alpha).100]\%$.

Para concluir, podemos señalar las condiciones que deben cumplirse para que las predicciones que llevemos a cabo sean fiables; serían:

- Que la relación lineal estimada entre Y y X se mantenga estable en el futuro (si estamos trabajando con datos de series temporales) o fuera de la muestra (si estamos considerando datos de corte transversal).
- Que los coeficientes de regresión sean suficientemente estables como para que sus estimaciones actuales muestrales sean una buena aproximación a los valores obtenidos tras incorporar observaciones futuras o extra-muestrales.
- Que se conozcan los valores futuros o extra-muestrales de X, o que los modelos de predicción utilizados para su obtención sean fiables.
- Que no existan errores de especificación en el modelo estimado.
- Que el horizonte de predicción no sea muy lejano.

3.6. <u>Introducción al uso de *EViews* (II)</u>.-

Continuamos en este apartado profundizando en nuestro conocimiento del manejo de *EViews*. En particular, veremos cómo se llevan a cabo los distintos contrastes de hipótesis más comúnmente aplicados en el proceso de validación de un modelo econométrico y aprenderemos, asimismo, a realizar predicciones. Como primer paso para desarrollar esta tarea, procederemos a establecer un modelo econométrico que nos servirá así de ejemplo.

Nuestra explicación se va a estructurar siguiendo estos puntos:

- Especificación y estimación del modelo
- Contrastes de significatividad individual de las variables explicativas y contraste de significatividad global del modelo
- Contraste de normalidad de la perturbación aleatoria del modelo
- Contrastes de hipótesis sobre los coeficientes de regresión del modelo basados en el test general de restricciones lineales
- Contraste de estabilidad: test de cambio estructural de Chow
- Estimación y predicción puntual de valores de la variable dependiente del modelo

Especificación y estimación del modelo

La estimación de un modelo econométrico es un proceso interactivo que empieza con la especificación de la relación supuesta entre ciertas variables. La selección de una especificación adecuada del modelo requiere, generalmente, la consideración de varias posibilidades de estimación. Por un lado, el número y tipo de variables a incluir, junto con la forma funcional del modelo, y por otro, la estructura dinámica subyacente entre las variables para el caso de series temporales. Evidentemente, siempre tendremos la incertidumbre acerca de la idoneidad o no de esta especificación inicial. Por ello, una vez estimada la ecuación, debemos proceder a evaluar la calidad de la misma mediante la aplicación de una serie de contrastes de hipótesis que garanticen la viabilidad de la estructura estimada.

La mayor parte de los contrastes utilizados para validar un modelo se plantean bajo la consideración de una hipótesis nula, por lo que cuando se aplica uno de ellos, *EViews* proporciona el valor del test estadístico en cuestión y la probabilidad asociada al mismo (*p-valor*). Este último valor indica el nivel de significación mínimo (o si nos referimos a 1 – *p-valor*, el nivel de confianza máximo) al que se puede rechazar la hipótesis nula suponiendo que ésta fuera cierta. O bien, el nivel de significación máximo (o si nos referimos a 1 – *p-valor,* el nivel de confianza mínimo) al que se puede aceptar la hipótesis nula suponiendo que ésta fuera cierta.

Así, un valor pequeño de esta probabilidad conduce a rechazar la hipótesis nula, mientras que un valor elevado (siempre menor que 1) significa que aceptaremos la hipótesis nula. Por ejemplo, un *p-valor* comprendido entre 0,05 y 0,01, implica rechazar la hipótesis nula al 5% de nivel de significación, pero no al 1%. Con este procedimiento, *EViews* no precisaría del manejo de tablas estadísticas para efectuar contrastes de hipótesis.

Para utilizar los contrastes disponibles en *EViews* dirigidos a la evaluación de un modelo tras su estimación, vamos a resolver el Ejercicio nº 23 del Boletín de este Tema. Los datos del mismo están disponibles en el archivo ***tests.wf1***, por lo que previamente deberemos descargarlo en el *Escritorio* de nuestro PC desde el espacio reservado a la Asignatura en la plataforma de docencia virtual *WebCT*.

En dicho fichero aparecen las variables o series de datos siguientes, correspondientes al periodo 1980-2007:

- **CONSUMO**: Tasas de variación anuales del consumo privado no alimenticio en términos reales de 1992
- **PRECIO**: Tasas de variación anuales de los precios del consumo privado no alimenticio con base 1992
- **RENTA**: Tasas de variación anuales de la renta familiar disponible en términos reales de 1992
- **EMPLEO**: Tasas de variación anuales del empleo total
- **PRECA**: Precariedad en el mercado de trabajo (porcentaje de contratos temporales sobre total de contratos)
- **TIR**: Tipos de interés reales en base 1992

A partir de ellas vamos a especificar, y seguidamente a estimar, un modelo donde la variación del consumo privado no alimenticio es explicada por las demás variables.

Al abrir el fichero *tests.wf1* (y pulsar luego la opción *LABEL+/-*), obtendremos la pantalla de la *Figura 10*. Dado que el modelo que vamos a especificar con los datos disponibles (1980-2007) se utilizará luego para realizar una predicción para 2008, obsérvese en *RANGE* cómo el fichero ya está preparado en este sentido.

A continuación, procedemos al **ajuste mínimo-cuadrático del modelo**. La especificación que vamos a plantear de nuestro modelo es:

$$CONSUMO = \beta_1 + \beta_2 PRECIO + \beta_3 RENTA + \beta_4 EMPLEO + \beta_5 PRECA + \beta_6 TIR + u$$

Para realizar esta estimación se selecciona *QUICK / ESTIMATE EQUATION*. La ventana de diálogo que se abre entonces se cumplimenta como puede observarse en la *Figura 11*.

La *Figura 12* muestra los resultados obtenidos del modelo estimado. Podemos guardar éste dándole un nombre; por ejemplo: REG. Como ya sabemos, lo haremos pulsando el botón *NAME*.

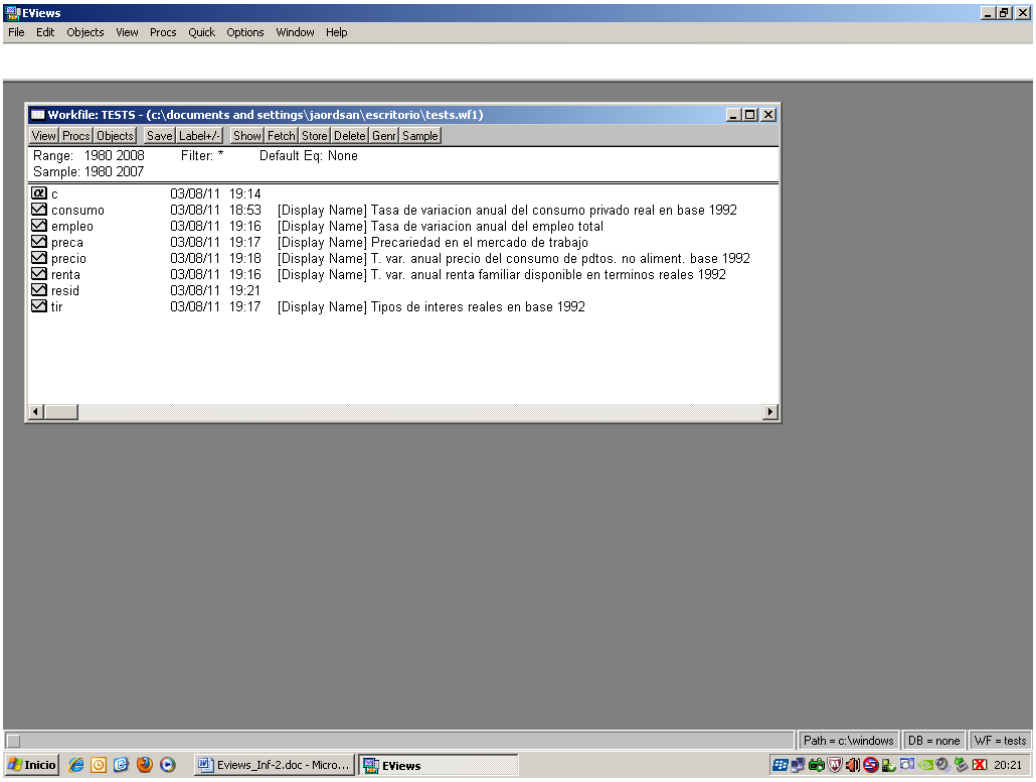


Figura 10

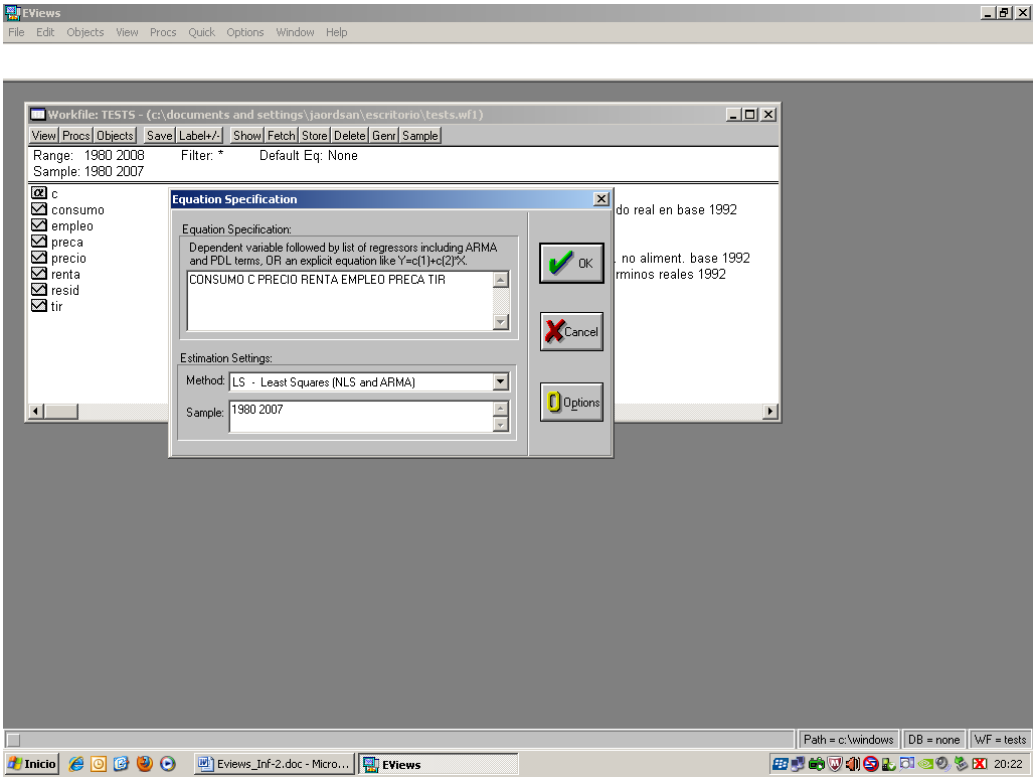


Figura 11

 ISBN: 978-84-694-7251-4

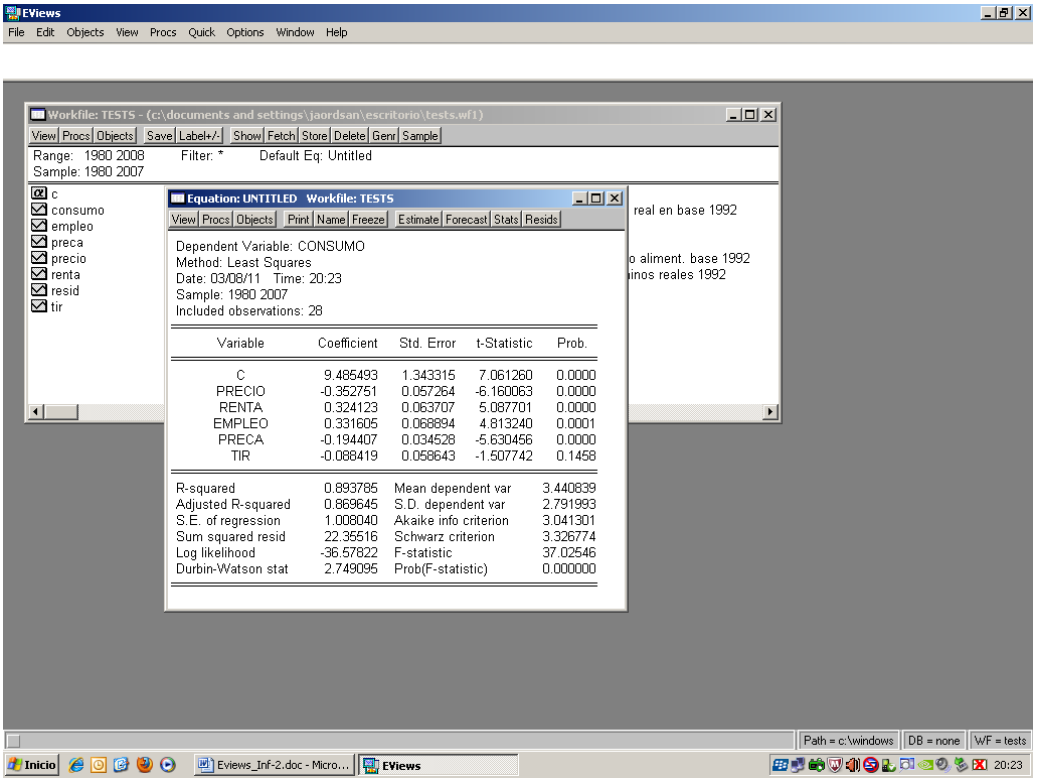


Figura 12

De estos resultados, lo primero que podemos reseñar es el signo de los coeficientes de regresión estimados, el significado de tales coeficientes, así como la bondad del ajuste muestral:

- En efecto, parece que se ha obtenido una especificación correcta del modelo en cuanto a los signos de los parámetros se refiere. Desde el ámbito teórico, un crecimiento de la renta y del empleo propiciarán un incremento del consumo no alimenticio, mientras que un incremento de los precios, de la precariedad laboral y de los tipos de interés contribuirán a retraer el consumo.
- En cuanto a los parámetros estimados, dado como están definidas las variables, en este caso éstos reflejan una aproximación al concepto de elasticidad. Así, el coeficiente asociado a la renta familiar disponible representa la elasticidad renta del consumo privado de productos no alimenticios; para nuestro caso, un crecimiento de un 1% en la renta se traduce en un incremento de un 0,324123% en el consumo de productos no alimenticios. El significado para el resto de variables es análogo.
- Respecto a la bondad del ajuste, los valores de los coeficientes de determinación (0,893785) y de determinación corregido (0,869645) pueden considerarse muy aceptables.

Contrastes de significatividad individual de las variables explicativas y contraste de significatividad global del modelo

Junto a los resultados anteriores, básicos de la estimación del modelo, la *Figura 12* también nos ofrece la información precisa para realizar los contrastes de significatividad individual de las variables explicativas y del contraste de significatividad global del modelo:

- La significatividad individual de las distintas variables explicativas del modelo puede contrastarse a través del *p-valor* asociado al estadístico *t*-Student de cada una de ellas. Según esto, todas son estadísticamente significativas a un nivel de confianza del 95% (e incluso del 99%), a excepción de TIR. Puesto que el nivel de confianza mínimo para aceptar la hipótesis nula de no significatividad de la variable de tipos de interés (TIR) es del 85,42% (la cifra complementaria al 14,58% que aparece), podemos entonces suponer que dicha variable no va a tener efectos relevantes sobre el consumo privado no alimenticio.
- En cuanto a la significatividad conjunta o global del modelo, vemos que el *p-valor* (0,000000) asociado al estadístico *F* de este contraste (37,02546) así lo evidencia; esto es, el nivel de significación mínimo para rechazar la hipótesis nula de no significatividad del modelo suponiendo que ésta fuese cierta es del 0,00%.

Contraste de normalidad de la perturbación aleatoria del modelo

Como bien sabemos, uno de los pilares fundamentales sobre los que se sustenta la construcción del modelo clásico de regresión lineal es el hecho de que la perturbación aleatoria siga una distribución de probabilidad *normal*, puesto que en ello se basa todo el desarrollo de la teoría inferencial del modelo (contrastes e intervalos de confianza).

Así pues, antes de seguir adelante debería contrastarse si, en efecto, la perturbación de nuestro modelo se comporta como una *normal*. Pero dado que la perturbación es aleatoria e inobservable por definición, el estudio de su normalidad debe hacerse a partir de una estimación de la misma; la serie de los residuos o errores muestrales del modelo constituye dicha estimación.

Por tanto, nuestro objetivo en este punto será analizar la normalidad de los residuos. Este análisis se realiza situándonos en la ventana de la ecuación estimada, donde seleccionaremos *VIEW*. De este modo se despliega un menú en el que, entre otras, tenemos las opciones siguientes para elegir:

- **Coefficient Tests**; nos facilitará los instrumentos para realizar cualquier tipo de contrastes de hipótesis nulas lineales sobre los coeficientes de regresión bien a través del estadístico de *Wald* o el de *Fisher-Snedecor*.

 ISBN: 978-84-694-7251-4

- **Residual Tests**; presenta diversas opciones destinadas a realizar un análisis exhaustivo de los residuos y con ello de la perturbación aleatoria del modelo: normalidad, heteroscedasticidad, autocorrelación…
- **Stability Tests**; ofrece la oportunidad de identificar cambios estructurales a través del *Test de Chow*, errores de especificación general con el *Test RESET de Ramsey* y analizar la inestabilidad de los parámetros utilizando los residuos recursivos.

En nuestro ejemplo, elegiremos *RESIDUAL TESTS* y luego *HISTOGRAM-NORMALITY TEST*, según aparece en la *Figura 13*.

El resultado aparece en la *Figura 14*. De esta salida, el aspecto que más nos interesa en este punto es el relativo al *contraste de Jarque-Bera*. Como ya sabemos, este contraste plantea como hipótesis nula la normalidad de la serie de datos analizada, que en este caso es la de los residuos. De acuerdo con el *p-valor* asociado al estadístico de Jarque-Bera, el nivel de confianza máximo para rechazar la hipótesis nula es del 78,19%, por lo que incluso para un 90% aceptaremos la hipótesis nula.

En definitiva, acabamos de contrastar que la serie de los residuos de la estimación, y con ello la perturbación aleatoria del modelo, sigue una distribución *normal* de probabilidad.

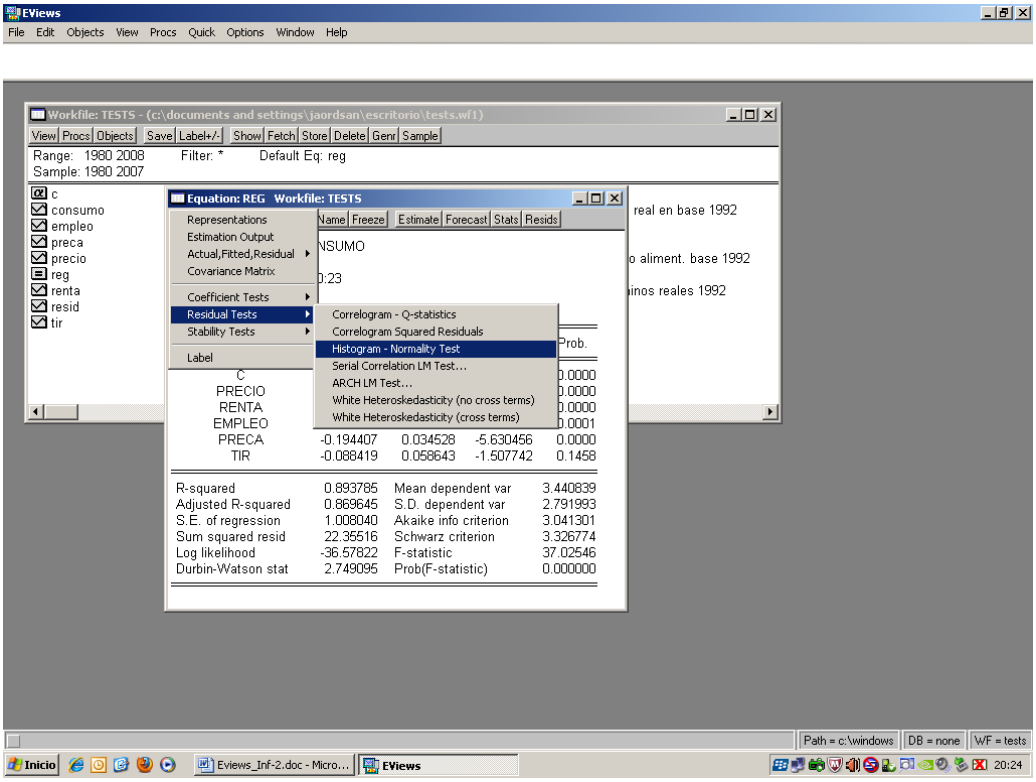


Figura 13

 ISBN: 978-84-694-7251-4

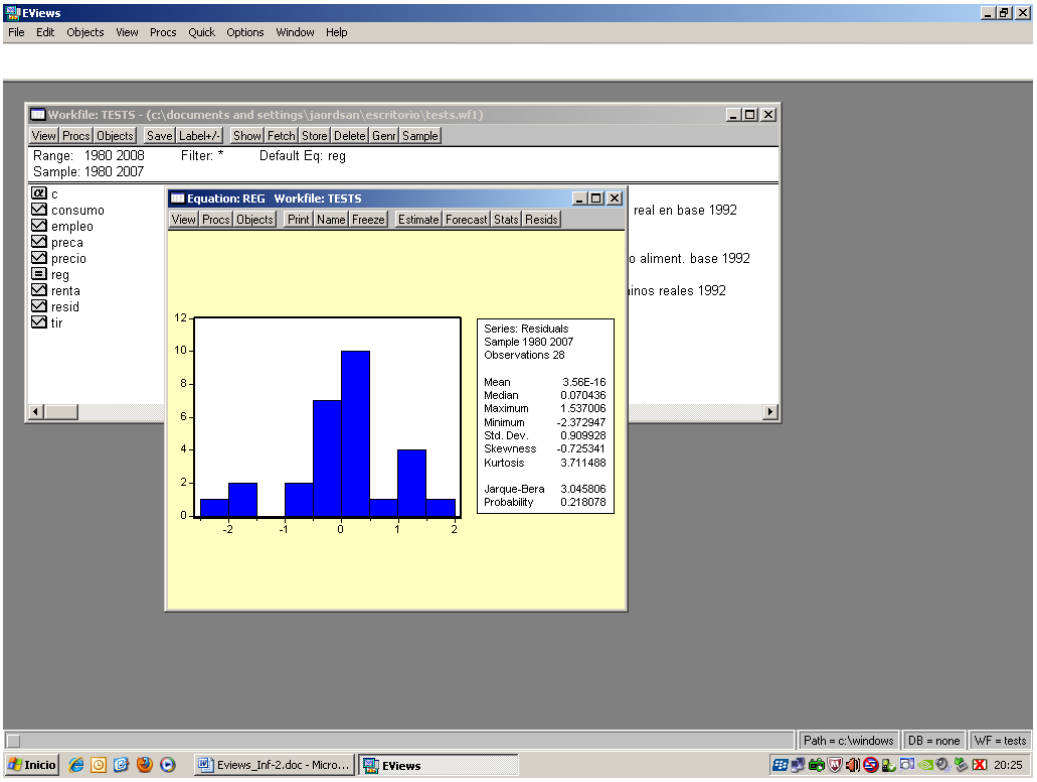


Figura 14

Contrastes de hipótesis sobre los coeficientes de regresión del modelo basados en el test general de restricciones lineales

Seguidamente vamos a realizar diversos contrastes de hipótesis lineales sobre los coeficientes de regresión del modelo.

Como ya es sabido, para ello se puede utilizar tanto un estadístico $F_{q,n-k}$ de *Fisher-Snedecor*, como el estadístico $\hat{W}$ de *Wald*, que sigue una distribución chi-cuadrado (en concreto: $\hat{W} \to \chi_q^2$). Además, se verifica que: $\hat{W} = q \cdot F$, siendo q el número de restricciones o ecuaciones que conforman la hipótesis nula a contrastar.

En nuestro ejercicio podemos plantearnos, por ejemplo, si las elasticidades de la renta y del empleo sobre el consumo son iguales. Para ello, estableceremos las hipótesis:

$$H_0 : \beta_3 = \beta_4$$
$$H_1 : \beta_3 \neq \beta_4$$

Dentro de la ventana de la ecuación REG, seleccionaremos nuevamente en la barra de menús la opción *VIEW* y después *COEFFICIENT TESTS*, donde elegiremos *WALD-COEFFICIENT RESTRICTIONS*, según se muestra en la *Figura 15*.

Aquí escribiremos la restricción referida a la hipótesis nula (*Figura 16*).

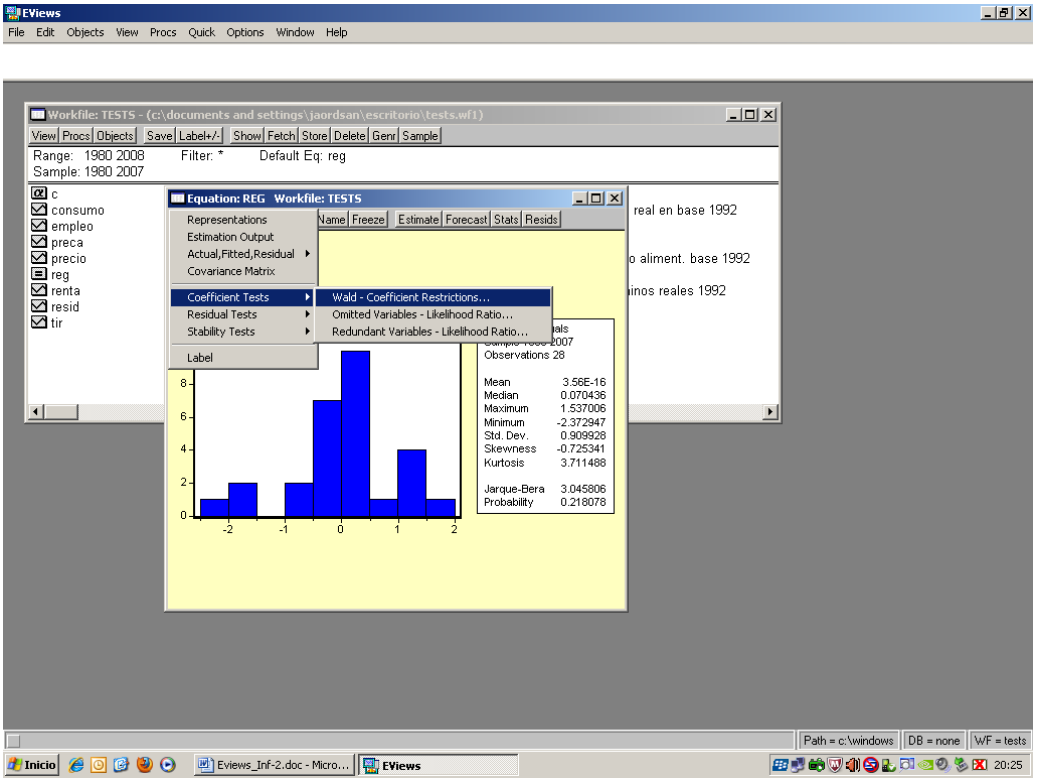


Figura 15

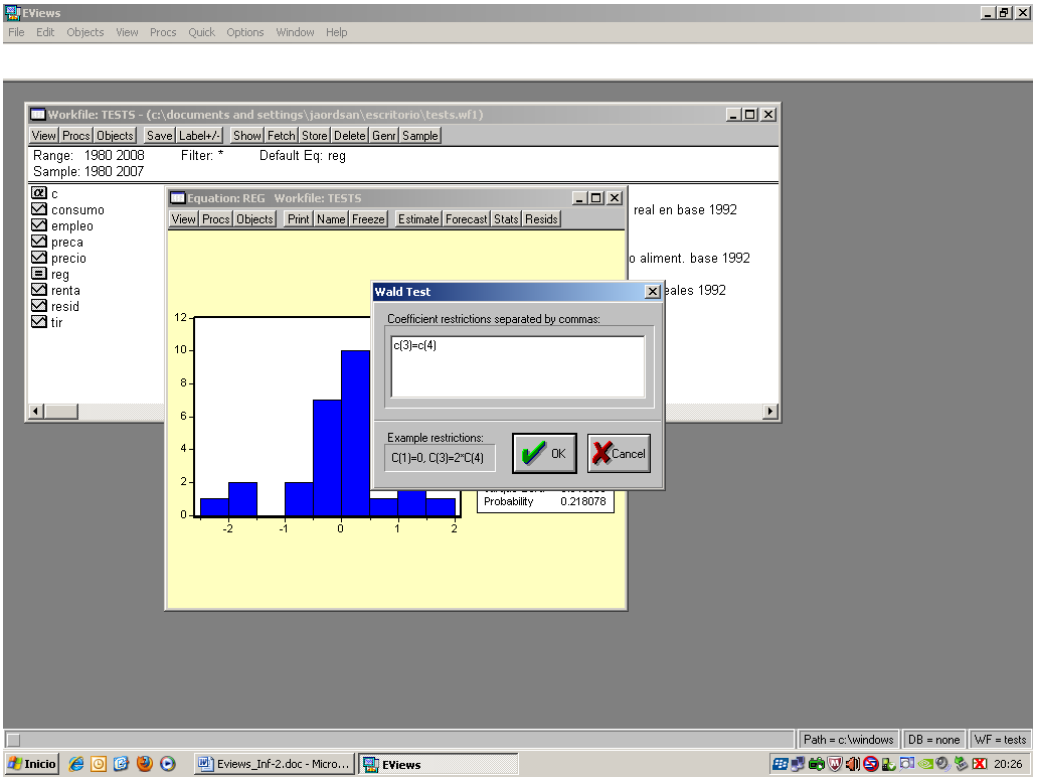


Figura 16

Tras pulsar seguidamente *OK*, obtendremos la pantalla de resultados que aparece en la *Figura 17*, donde puede verse, en primer lugar, cómo queda perfectamente indicada la

hipótesis nula que estamos verificando; en segundo lugar, el valor del estadístico F (que sigue aquí una distribución de *Fisher-Snedecor* con 1 y 22 grados de libertad) coincide con el del estadístico χ^2 *de Wald* (con 1 grado de libertad), puesto que la hipótesis nula sólo se compone en este caso de 1 restricción; y, finalmente, si atendemos a los *p-valores* asociados a cualquiera de los dos estadísticos indicados, veremos que podemos aceptar la hipótesis nula a partir de un nivel de confianza de en torno al 5,3%. En definitiva, a tenor de estos resultados podemos afirmar que ambas elasticidades son prácticamente iguales.

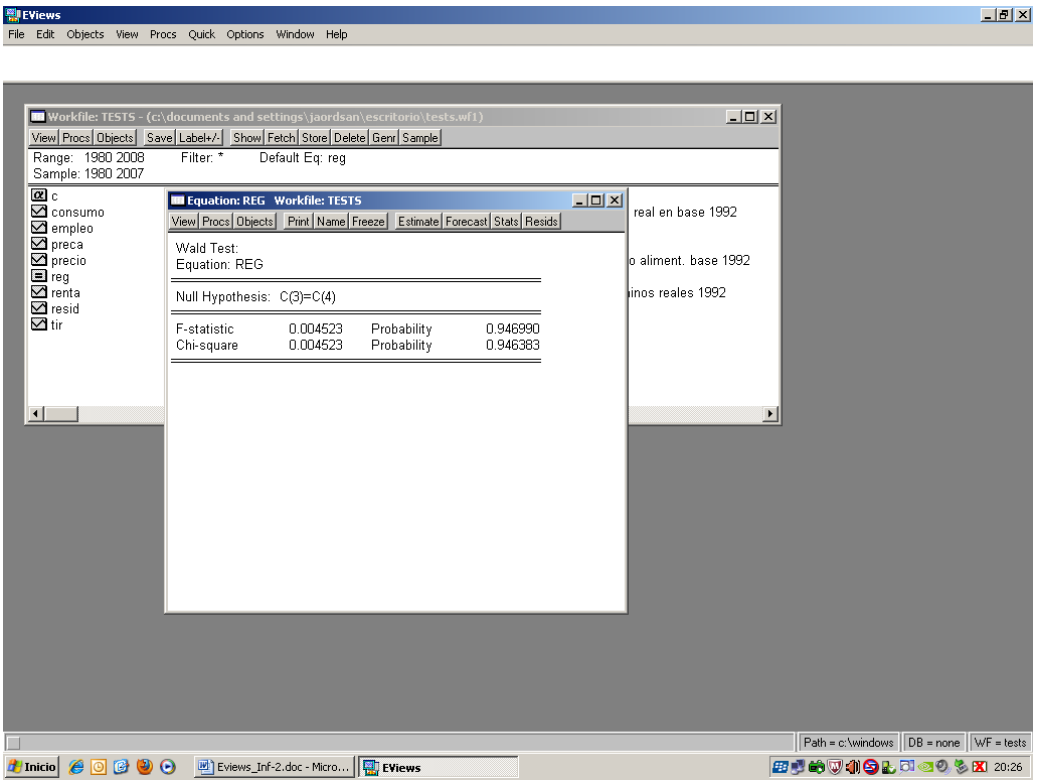


Figura 17

Para aplicar este test en el caso en que se considere más de una restricción en la hipótesis nula a contrastar, debemos separar éstas entre *comas* al indicarlas en el cuadro de diálogo del test.

Por ejemplo, podemos contrastar ahora que la suma de las elasticidades de la renta y el empleo sea igual a 1, a la vez que el valor de la primera sea el doble que el de la segunda; esto es:

$$H_0 : \beta_3 + \beta_4 = 1$$
$$\beta_3 = 2\beta_4$$
$$H_1 : \text{No se verifican a la vez ambas restricciones}$$

En la *Figura 18* aparece el resultado de este contraste: el valor del estadístico de *Wald* en este caso es el doble que el del estadístico F (pues ahora hay 2 restricciones en la

hipótesis nula) y la hipótesis nula se puede rechazar, atendiendo a cualquiera de los dos estadísticos de prueba, para todos los niveles de significación estándar en el ámbito de la Estadística (incluso del 1%).

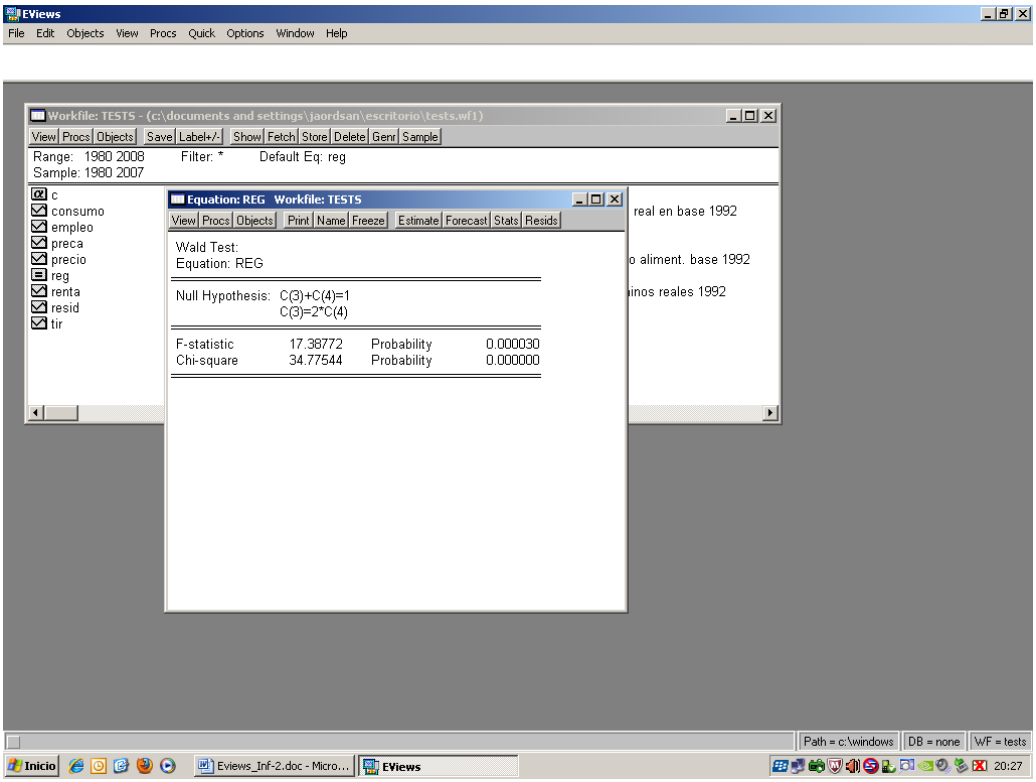


Figura 18

Contraste de estabilidad: test de cambio estructural de Chow

Para comprobar si la estructura estimada, definida por los coeficientes de regresión del modelo, es constante en el tiempo, podemos aplicar el *contraste de cambio estructural de Chow*.

Este tipo de contraste se recoge en el tercer grupo de contrastes definidos al inicio de este ejercicio, de forma que accederemos a él a partir del menú contenido en la ventana de la ecuación estimada, con *VIEW / STABILITY TESTS / CHOW BREAKPOINT TEST*.

Para realizar este contraste, es necesario definir un punto de corte de la muestra total, de forma que éste la divida en dos submuestras. Este punto es escogido a priori por el investigador, dependiendo de las circunstancias particulares de espacio y tiempo en que se muevan las variables (crisis del petróleo, etc.), así como del objetivo del análisis.

En el ejercicio que estamos desarrollando como ejemplo, vamos a comprobar si los grandes eventos del año 1992 en España (Juegos Olímpicos de Barcelona y Exposición Universal de Sevilla) tuvieron algún efecto sobre el consumo de productos no alimenticios. Así pues, seleccionamos el *Test de Chow* e indicamos como punto de

corte: 1992 (*Figura 19*). En este test, la hipótesis nula establece la ausencia de cambio estructural. El resultado final aparece en la *Figura 20*.

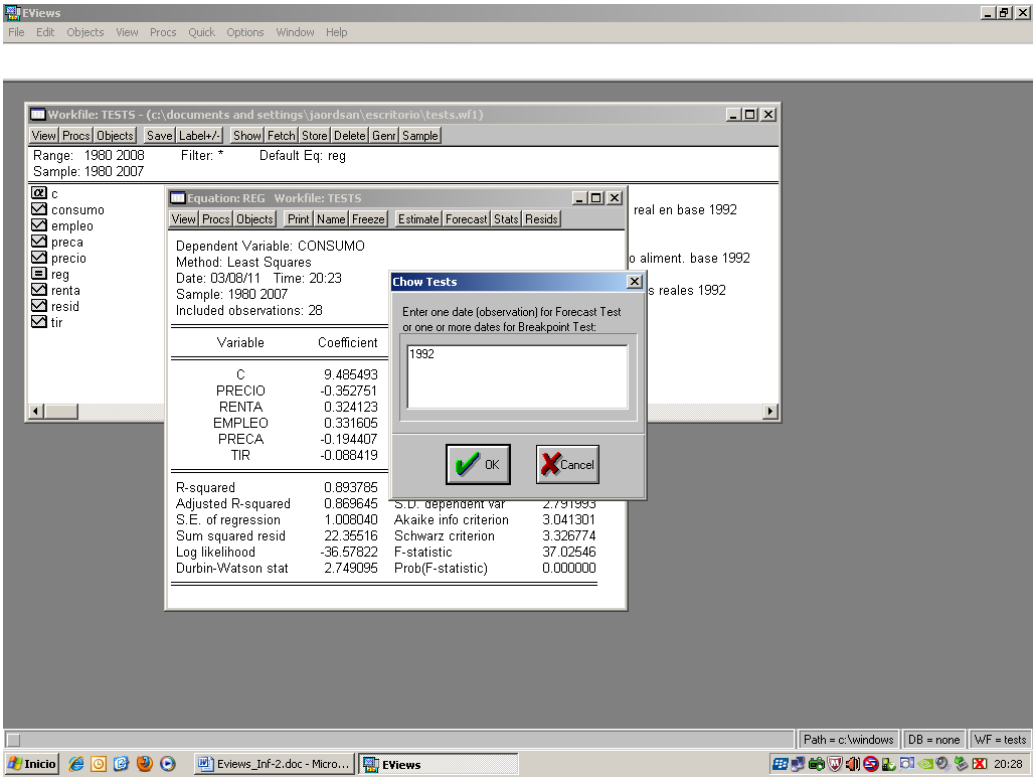


Figura 19

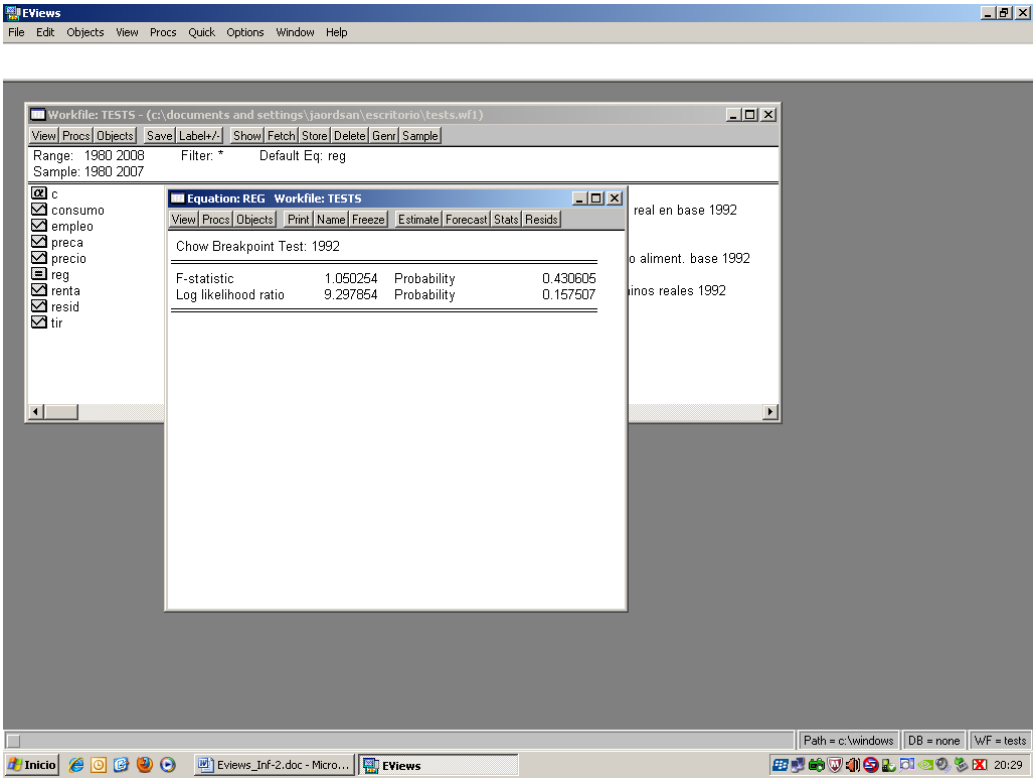


Figura 20

Aunque *EViews* nos calcule de nuevo dos estadísticos, vamos a centrarnos esta vez exclusivamente en el estadístico *F* de *Fisher-Snedecor*. Este estadístico se basa en la comparación entre la suma de los residuos al cuadrado de la regresión total (1980-2007) y las correspondientes a las regresiones de las dos submuestras (1980-1991 y 1992-2007). El elevado *p-valor* obtenido nos conduce a aceptar la hipótesis nula de ausencia de cambio estructural a un nivel máximo de significación del 43,06%; es decir, el consumo no alimenticio no presenta un comportamiento diferenciado en el transcurso de todo el periodo.

Estimación y predicción puntual de valores de la variable dependiente del modelo

En este último punto vamos a llevar a cabo una predicción para el año 2008 de la tasa de variación del consumo en productos no alimenticios, con base 1992, a partir de nuestra especificación del modelo. Para ello, se dispone de los datos correspondientes a 2008 de las variables explicativas de dicho modelo:

PRECIO	RENTA	EMPLEO	PRECA	TIR
3,2348	2,8722	2,1521	35,0152	4,2807

Lo primero que deberá hacerse es introducir las cifras indicadas para 2008 en todas y cada una de las series correspondientes. Con este fin, deberá irse abriendo cada serie y pulsar *EDIT +/-* entre sus opciones, escribiéndose entonces la cifra en la posición señalada, tal y como se refleja en la *Figura 21* para el caso de la variable PRECIO. (No debe olvidarse que los decimales en *EViews* deben escribirse en notación anglosajona, es decir, tras un punto.) Después de introducir cada cifra, pulsaremos nuevamente *EDIT +/-* para bloquear la escritura y evitar modificar alguna otra cifra por error.

A continuación se realiza la predicción del valor de la cifra de consumo utilizando la especificación lineal del modelo; para ello habrá de seleccionarse la ecuación estimada REG. Una vez abierta, se elige la opción *FORECAST*, obteniendo una ventana en la que deberemos dar un nombre a la nueva serie de los valores estimados de la variable dependiente. Por defecto, *EViews* nombra a esta serie igual que a la serie original pero añadiéndole al final una "F" (del inglés, *forecast*). En este caso, CONSUMOF. Podemos dejar este nombre, pero puede cambiarse a gusto del investigador. Asimismo deberemos elegir el rango de datos de la salida estimada. Aquí deberá elegirse 1980 2008. Con ello, las cifras de 1980 a 2007 de la serie CONSUMOF serán datos estimados, en tanto que la correspondiente a 2008 será una verdadera predicción extra-muestral. Además de ello, *EViews* permite crear la serie de errores estándar estimados de los errores de predicción puntual, que puede nombrarse como se desee (por ejemplo, ESERRORF). Por lo demás, vamos a dejar las opciones señaladas por defecto. La pantalla quedaría tal como se indica en la *Figura 22*.

 ISBN: 978-84-694-7251-4

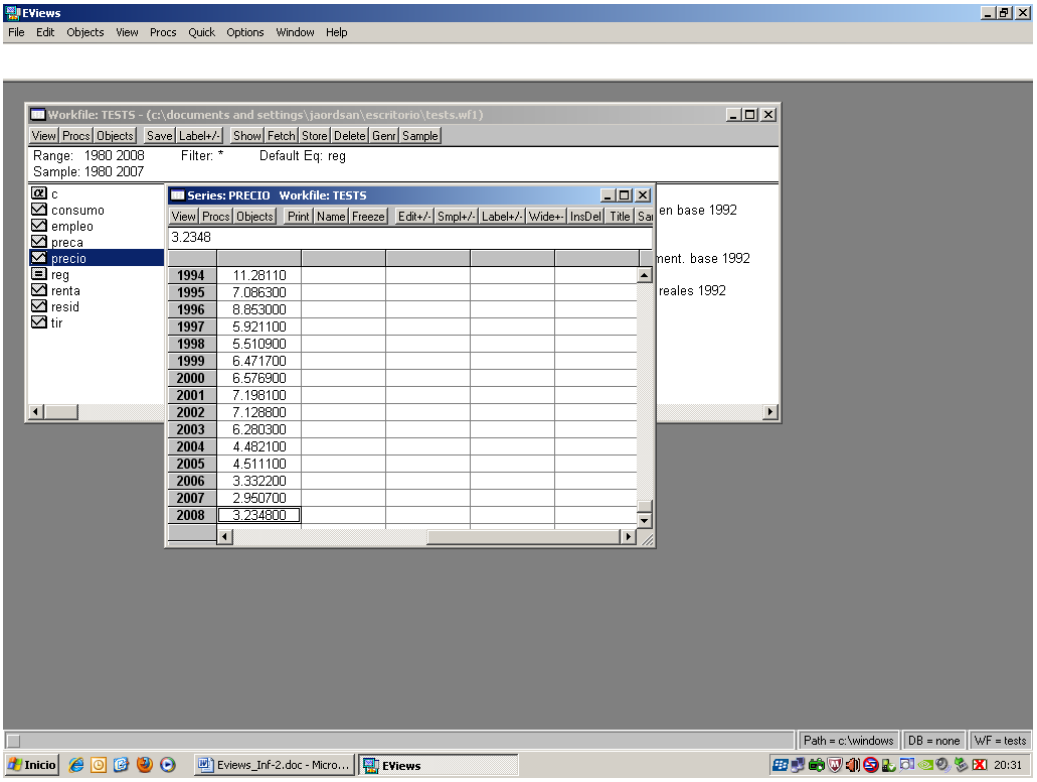


Figura 21

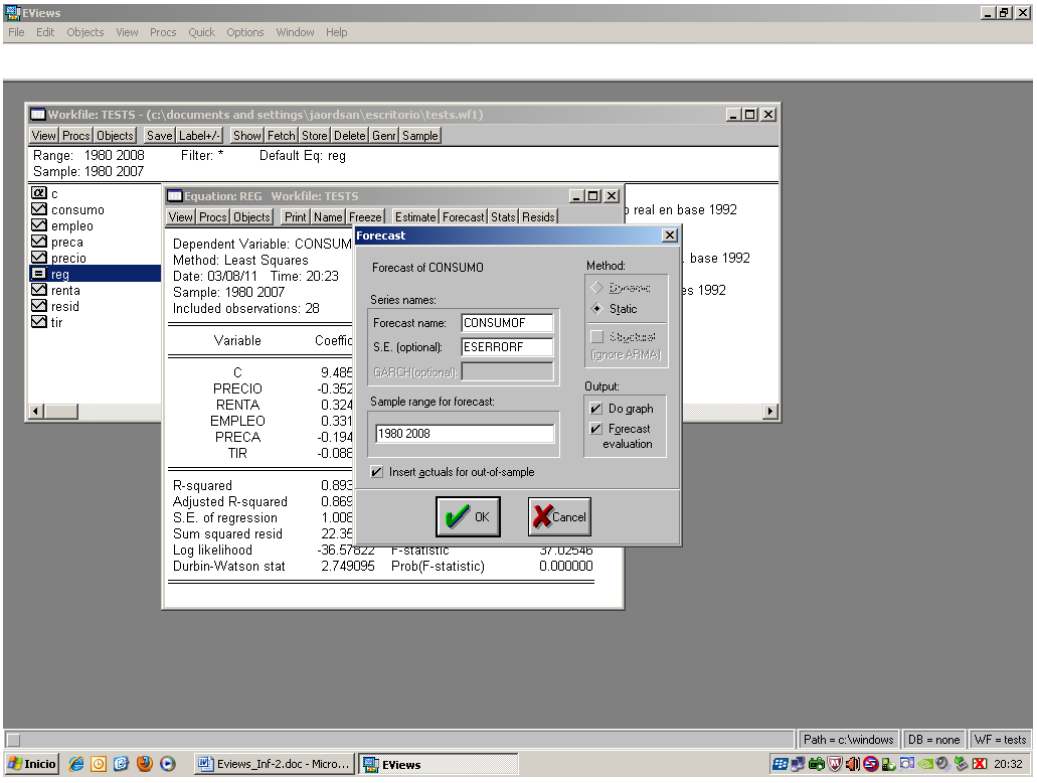


Figura 22

La *Figura 23* muestra el resultado obtenido. En ella se ofrecen algunos estadísticos de referencia para evaluar la estimación-predicción realizada (raíz cuadrada del error

cuadrático medio, error medio absoluto, etc.), así como el gráfico de la misma y un intervalo de confianza específico (para una amplitud igual a dos veces la desviación típica de la muestra; es decir, en torno al 95%).

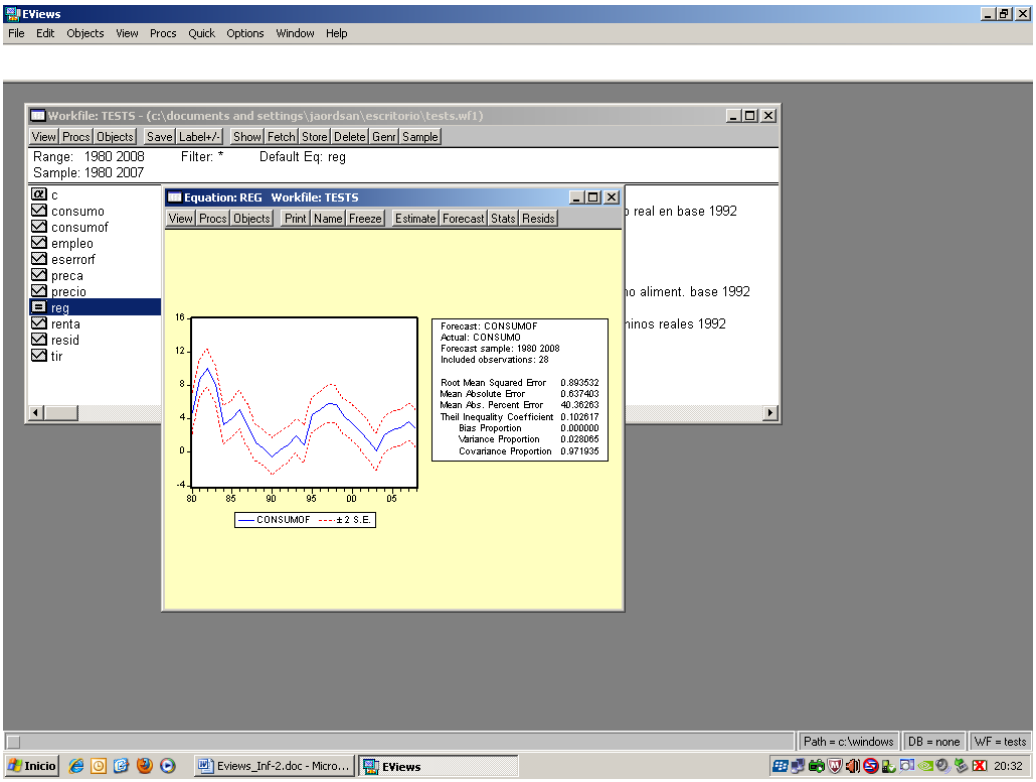


Figura 23

Igualmente, resulta interesante ver la *Figura 24*, donde se representan gráficamente para el periodo 1980-2007 la serie de datos reales de las tasas de variaciones del consumo con base 1992, conjuntamente con la de sus valores estimados a partir del modelo lineal establecido (CONSUMO y CONSUMOF, respectivamente), así como la de los residuos MCO resultantes tras el ajuste. Con ello podemos ver que dicho ajuste resulta bastante bueno y también qué observaciones son las que presentan mayores desviaciones entre el dato real y el estimado; esto es, dónde se registran los mayores residuos, siendo en este caso los correspondientes a los años 2003, 1981 y 1985. Esto se hace a través de: *VIEW / ACTUAL, FITTED, RESIDUAL / ACTUAL, FITTED, RESIDUAL GRAPH*.

Por último, podemos reseñar cómo en la ventana de trabajo puede observarse que aparecen las nuevas series de datos CONSUMOF y ESERRORF generadas. Además, si se abre la primera de ellas (CONSUMOF), se podrá comprobar que, junto con las estimaciones de los datos que van de 1980 a 2007, para 2008 aparece un nuevo dato: 2,8033, que resulta ser en este caso una predicción extra-muestral. Todo esto puede apreciarse en la *Figura 25*.

Para finalizar, podemos guardar este fichero para su uso en una sesión de trabajo posterior. Esto lo haremos yéndonos a *FILE / EXIT* en la barra principal de menús. De

este modo, podremos aprovechar este mismo modelo para analizar más adelante posibles problemas que pudiese presentar.

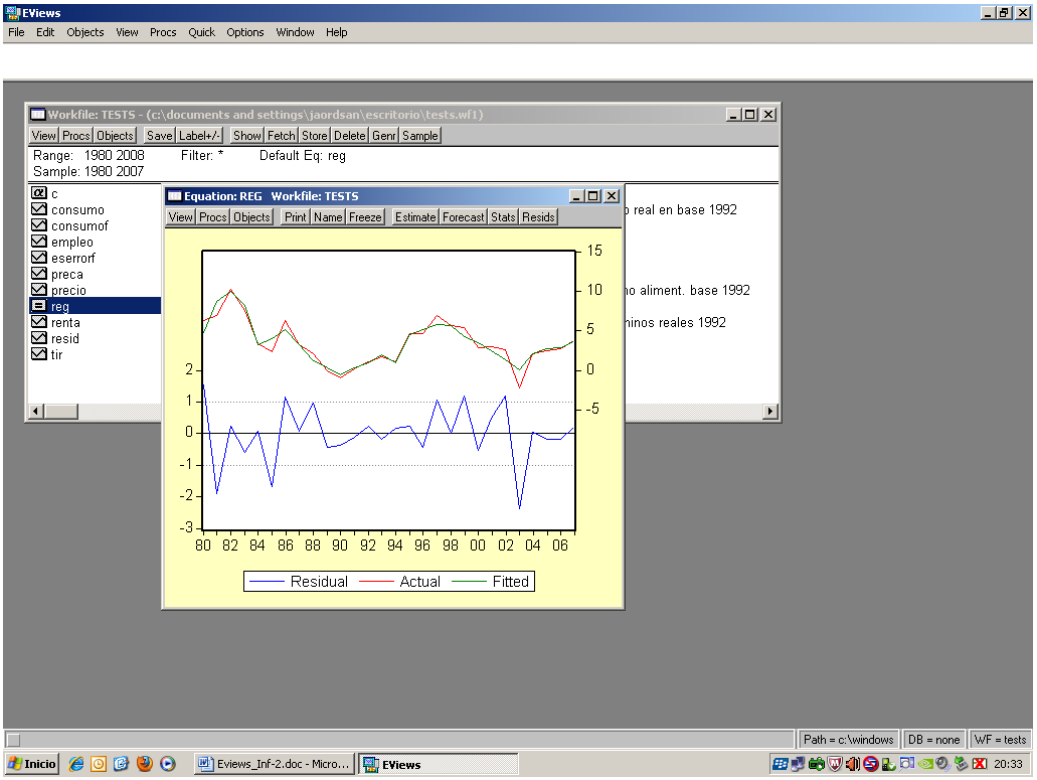


Figura 24

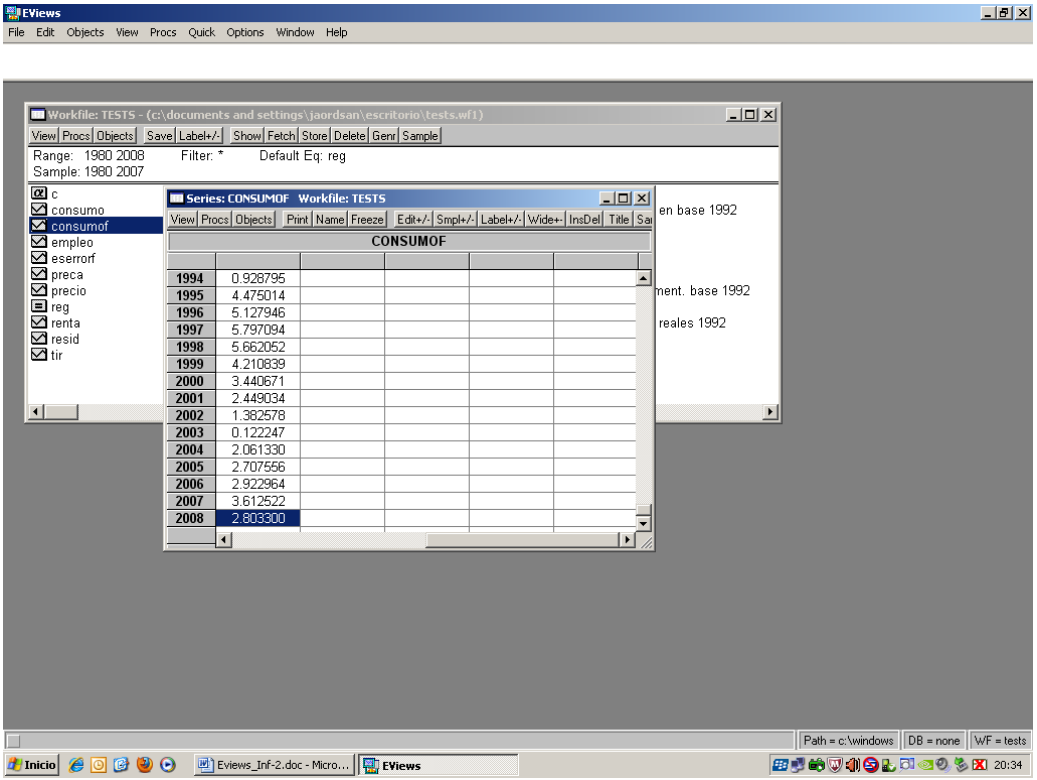


Figura 25

TEMA 4

El modelo clásico de regresión lineal: incumplimiento de supuestos

Hasta este momento hemos estado trabajando con un modelo de regresión lineal "perfecto", que posee toda una serie de propiedades sustentadas en un amplio conjunto de supuestos de distinta naturaleza, estocásticos y no estocásticos. En este Tema vamos a estudiar qué sucede en nuestro modelo clásico de regresión lineal cuando no se cumplen todos los requisitos o supuestos descritos previamente. Analizaremos distintas situaciones: problemas de errores de especificación en el modelo (que pueden provenir de diferentes causas), presencia de multicolinealidad entre las variables explicativas del modelo y existencia de heteroscedasticidad y/o autocorrelación en la perturbación aleatoria de éste.

4.1. Errores de especificación del modelo. Tests de detección.-

La especificación de un modelo constituye la primera fase, y fundamental, de todo el proceso de análisis de la realidad económica a través de la teoría econométrica. Sin embargo, en este primer paso puede que cometamos errores, originados por diferentes fuentes, que condicionen gravemente los resultados finales de nuestra modelización.

Errores de especificación del modelo

En este apartado vamos a ir analizando, de manera muy sintética, cada una de las posibles fuentes o causas que pueden dar lugar a errores de especificación de nuestro modelo.

- ***Omisión de variables explicativas relevantes***

Ante un problema de omisión de variables explicativas relevantes en la especificación del modelo:

- En general, las estimaciones MCO de los coeficientes de regresión ($\hat{\beta}$) resultan sesgadas e inconsistentes.
- La varianza estimada de la perturbación aleatoria ($\hat{\sigma}_u^2$) es sesgada.
- Las varianzas estimadas de los coeficientes de regresión ($\widehat{Var}(\hat{\beta})$) resultan ser sesgadas, sobreestimando su verdadero valor. Esto hace que la construcción de intervalos de confianza y la realización de contrastes de hipótesis no sea fiable, ya que, por una parte, la amplitud de los intervalos de confianza aumenta y, por otra,

 ISBN: 978-84-694-7251-4

disminuye la potencia de los contrastes (es decir, aumenta la probabilidad de aceptar hipótesis nulas aun siendo falsas).

- ***Inclusión de variables explicativas irrelevantes***

En el caso de que en el modelo se incluyan variables explicativas irrelevantes:

- Los estimadores MCO de los coeficientes de regresión correctamente incluidos son insesgados y consistentes. Por su parte, el valor esperado de los incorrectamente incluidos será 0, por lo que cabe prever que ante un contraste de su significatividad, se evidencie que no lo son.
- La varianza estimada de la perturbación aleatoria ($\hat{\sigma}_u^2$) es insesgada.
- Las varianzas estimadas de los coeficientes de regresión ($\widehat{Var}(\hat{\beta})$) son ineficientes; son, por lo general, mayores que lo que debieran ser. Los estimadores $\hat{\beta}$ no son, por tanto, ELIO, ya que son insesgados, pero no son los mejores.
- Los procedimientos estándares de definición de intervalos de confianza y contrastes de hipótesis siguen siendo, pese a todo, válidos en líneas generales, aunque no los mejores posibles.

- ***Adopción de la forma funcional incorrecta***

Si se elige la forma funcional equivocada, la principal consecuencia es que:

- Los coeficientes estimados por MCO ($\hat{\beta}$) pueden ser estimaciones sesgadas de los parámetros poblacionales del modelo que resultaría más adecuado en la realidad.

- ***Errores de medición***

a) Errores de medición en la variable explicada

- Los estimadores MCO de los coeficientes ($\hat{\beta}$) son insesgados.
- Sus varianzas son también insesgadas, aunque mayores que en el caso en que no hubiera este error. Así pues, los estimadores no son ELIO, pues no son eficientes.

b) Errores de medición en las variables explicativas

- Los estimadores MCO de los coeficientes ($\hat{\beta}$) son sesgados e inconsistentes.
- Éste es, por tanto, un problema más grave que el caso anterior. Se podría trabajar entonces, en lugar de con las variables verdaderas, con variables que fuesen aproximaciones de éstas (variables instrumentales o "*proxy*").

 ISBN: 978-84-694-7251-4

Tests de detección

Seguidamente se exponen dos tests, muy generalizados, que permiten detectar la presencia de algunos de los problemas de especificación en un modelo:

- ***Test de la F de comparación entre modelos restringidos y sin restringir***

Se utiliza básicamente para contrastar los errores relativos a omisión de variables relevantes o inclusión de variables irrelevantes. La hipótesis nula y el estadístico de prueba experimental en este test son:

$H_0 : \beta_m = \theta$

$$F^{\exp} = \frac{(SCR_r - SCR)/m}{SCR/n-k} = \frac{(R^2 - R_r^2)/m}{(1-R^2)/(n-k)} \rightarrow F_{m,n-k},$$

donde:

β_m es el vector de coeficientes correspondientes a las *m* variables incluidas de más u omitidas (según el tipo de error que se contraste)
n = nº de datos empleados en la modelización
k = nº de variables explicativas (incluida la ordenada en el origen) del modelo planteado de partida

- ***Test RESET de Ramsey***

Este test es un test general de mala especificación de un modelo, aplicable para detectar la omisión de variables relevantes y la elección de una forma funcional inadecuada. La hipótesis nula es que el modelo de partida está bien especificado. Sus pasos son:

1. A partir del modelo inicialmente elegido, se obtienen los valores estimados de la variable dependiente: $\hat{Y}_i$.
2. Se vuelve a estimar el modelo elegido añadiéndole como variables explicativas potencias de $\hat{Y}_i$; esto es: $\hat{Y}_i^2, \hat{Y}_i^3, \ldots$ para capturar de este modo las posibles relaciones sistemáticas existentes entre los residuos y las $\hat{Y}_i$.
3. Se calcula entonces el siguiente estadístico *F* de prueba:

$$F^{\exp} = \frac{(R^2_{NUEVO} - R^2_{ANTIGUO})/l}{(1-R^2_{NUEVO})/(n-\mathrm{m})} \rightarrow F_{l,n-m},$$

donde:

l = nº de nuevos regresores
m = nº de parámetros del nuevo modelo

El objetivo es comprobar si el nuevo modelo supone una aportación significativa (vista a través de su R^2) respecto al original.

 ISBN: 978-84-694-7251-4

4. Según esto, por tanto, si el valor de la F es estadísticamente significativo al nivel elegido, entonces podemos concluir que el modelo inicial está mal especificado.

Una ventaja del test RESET es que es fácil de aplicar, porque no requiere especificar cuál es el modelo alternativo; pero esto también puede resultar un inconveniente, ya que si se rechaza el modelo inicialmente elegido, no se tiene entonces uno alternativo para sustituirlo. Así pues, este test puede considerarse fundamentalmente como una herramienta de diagnóstico.

4.2. Multicolinealidad perfecta y aproximada: definición, detección y tratamiento.-

Definición de multicolinealidad

Consideremos, en su expresión matricial, el modelo clásico de regresión lineal múltiple $Y = X\beta + u$, que cumple las hipótesis habituales.

Como es bien sabido, la estimación de los coeficientes de regresión del modelo por el método de MCO se obtiene a través de la expresión: $\hat{\beta} = (X'X)^{-1}X'Y$, verificándose que su matriz de varianzas-covarianzas viene dada por: $Var - Cov(\hat{\beta}) = \sigma_u^2 \cdot (X'X)^{-1}$.

El concepto de *multicolinealidad* hace referencia a la existencia de relación lineal entre las variables explicativas del modelo. Esta relación puede ser de distinto grado, lo que tiene distintas consecuencias. Así, puede darse desde la multicolinealidad perfecta hasta la ausencia total de relación lineal entre las X_j.

La multicolinealidad perfecta significa que existe una relación lineal exacta entre las variables explicativas del modelo, lo que implica que: $|X'X| = 0 \Rightarrow$ No existe $(X'X)^{-1}$. En este caso, es posible obtener una estimación de una combinación lineal de los parámetros del modelo, pero no calcular $\hat{\beta}_{MCO}$ de forma única.

Cuando no hay relación lineal alguna entre las distintas variables explicativas (es decir, el coeficiente de correlación lineal entre X_j y X_m vale 0, $\forall X_j \neq X_m$), la estimación por MCO de los parámetros poblacionales del modelo puede llevarse a cabo tanto de manera conjunta a través de la expresión $\hat{\beta} = (X'X)^{-1}X'Y$, como efectuando por separado las regresiones simples de la variable explicada con cada una de las variables explicativas; los resultados coincidirían exactamente. Esto es, se podría plantear:

$$\begin{cases} Y_i = \alpha_1 + \beta_2 X_{2i} + \varepsilon_i, \text{ de donde se obtendría } \hat{\beta}_2, \\ \dots \\ Y_i = \alpha_k + \beta_k X_{ki} + \varepsilon'_i, \text{ de donde se obtendría } \hat{\beta}_k. \end{cases}$$

Y finalmente: $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \ldots - \hat{\beta}_k \overline{X}_k$.

Los supuestos anteriores son extremos y prácticamente no se dan en la realidad. Lo usual es que exista cierta relación lineal entre las variables explicativas, en un mayor o menor grado. El problema aparece cuando este grado, sin ser máximo, es elevado; es lo que se conoce como multicolinealidad aproximada (o casi perfecta). Aunque $|X'X| \neq 0$, resulta que es cercano a 0. Las consecuencias de esta situación son:

- Los $\hat{\beta}_{MCO}$ son estimadores lineales, insesgados y óptimos (en el sentido de mínima varianza); es decir, siguen siendo ELIO.
- Se trata de un fenómeno muestral: la multicolinealidad puede estar presente en una muestra y no en la población.
- Altos valores de los elementos de la matriz de $\widehat{\text{var}-\text{cov}(\hat{\beta}_{MCO})}$; esto conlleva:
  - Amplios intervalos de confianza para los parámetros poblacionales β_j, $\left(\hat{\beta}_j \pm t_{n-k,1-\alpha/2} \cdot \widehat{ES(\hat{\beta}_j)}\right)$, con la consiguiente disminución de la precisión de sus estimadores.
  - Disminución drástica de la potencia de los contrastes de significatividad individual de los parámetros[1]: $t_{\beta_j} = \dfrac{\hat{\beta}_j}{\widehat{ES(\hat{\beta}_j)}}$.
- Amplios intervalos de confianza para la predicción de valores de la variable explicada Y_0, $\left(\hat{Y}_0 \pm t_{n-k,1-\alpha/2} \cdot \widehat{ES(e_0)}\right)$, perdiendo con ello precisión la predicción.
- Los $\hat{\beta}_{MCO}$ y $\widehat{ES(\hat{\beta}_{MCO})}$ son muy sensibles a fluctuaciones muestrales.

El tratamiento de la multicolinealidad es muy desigual por parte de los económetras en la literatura. Así, si el objetivo es la predicción, en general se señala que no existen graves problemas; pero si lo que se atiende es a la significativad de las variables del modelo, entonces debe tratarse la cuestión con sumo cuidado y detenimiento.

Detección de la multicolinealidad

Para detectar la presencia de multicolinealidad en el modelo, existen diversos procedimientos:

[1] Obsérvese que al aumentar $\widehat{ES(\hat{\beta}_j)}$, disminuye el valor del estadístico t_{β_j}, con lo que aumenta la probabilidad de aceptar hipótesis nulas aun siendo falsas.

1. Observar un valor del $|X'X|$ próximo a cero. No obstante, este método no resulta definitivo, ya que puede deberse a los valores concretos de la muestra escogida. Su principal limitación es que no tiene cota superior.

2. Estar ante un R^2 elevado (y, por tanto, ante una F que indica que el modelo es globalmente significativo) y pocos estadísticos t-Student significativos asociados a las variables explicativas. Se trata de un rasgo habitual en situaciones de multicolinealidad, si bien no es del todo concluyente.

3. Constatar altos valores de los coeficientes de correlación lineal simple ($|R| \geq 0{,}8$) entre las variables explicativas. Es una condición suficiente.

4. Prueba de eliminación de variables. Este método comienza calculando R^2 para el modelo completo; si se elimina luego aquella variable que se considere más correlacionada, y resulta que el nuevo valor de R^2 apenas varía, entonces es signo evidente de que la relación de colinealidad existía.

5. Método de las regresiones auxiliares de Farrar-Glauber. Este método consiste en efectuar las regresiones de cada variable explicativa X_j en función de las restantes (denominadas regresiones auxiliares), calculando sus correspondientes coeficientes de determinación R_j^2. Para cada una de estas regresiones auxiliares se lleva entonces a cabo el siguiente contraste mediante el estadístico F de Fisher-Snedecor:

   $H_0 : R_j^2 = 0$

   $$F_j = \frac{R_j^2 / (k-1) - 1}{(1-R_j^2)/n-(k-1)} \rightarrow F_{k-2,\, n-(k-1)}.$$

   Si la regresión auxiliar es estadísticamente significativa, supondría que X_j presenta fuerte relación lineal con las restantes variables; esto es, habría multicolinealidad.

6. Factor de agrandamiento (o inflación) de la varianza, de Klein. Este coeficiente, $FAV(\hat{\beta}_j)$, se define como el cociente de la varianza observada de $\hat{\beta}_j$ y la que habría sido si X_j fuese incorrelada con el resto de variables explicativas del modelo, que de forma práctica es la inversa de $1-R_j^2$, donde R_j^2 es el coeficiente de determinación de la regresión auxiliar de X_j con el resto de variables explicativas:

   $$FAV(\hat{\beta}_j) = \frac{\text{var}(\hat{\beta}_j)}{\text{var}(\hat{\beta}_j)_0} = \frac{1}{1-R_j^2}.$$

   Cuanto mayor sea el valor de $FAV(\hat{\beta}_j)$, mayor será la relación lineal entre las variables explicativas del modelo.

 ISBN: 978-84-694-7251-4

7. Número de condición, de Belsley, Kuck y Welsch. Es el método considerado más actual; se define como:

$$n(x) = \sqrt{\frac{\lambda_{máx}}{\lambda_{mín}}},$$

donde $\lambda_{máx}$ y $\lambda_{mín}$ son los autovalores máximo y mínimo, respectivamente, de la matriz de correlaciones de las variables explicativas R_{XX}.

Atendiendo al valor de este número, se tiene que:

- $n(x) = 1 \Leftrightarrow$ hay ausencia total de correlación lineal.
- $1 < n(x) \leq 20$; esto indica escasa multicolinealidad.
- $20 < n(x) \leq 30$; esto conlleva un problema de multicolinealidad que empieza a merecer su consideración.
- $n(x) > 30$; implica la presencia prácticamente segura de multicolinealidad.

Tratamiento de la multicolinealidad

Una vez detectada la presencia de multicolinealidad, existen diversas formas de acometer su corrección. Entre ellas, pueden destacarse las siguientes:

1. Eliminación de variables explicativas del modelo. El problema de la multicolinealidad denota, en esencia, la falta de información suficiente en la muestra para permitir una estimación precisa de los parámetros individuales. En ocasiones, puede interesar eliminar algunas variables del modelo original cuando existe una alta correlación entre ellas. Sin embargo, esto tiene consecuencias; la principal es que la estimación por MCO de los parámetros deja de ser ELIO, pues dejan de ser insesgados; sin embargo, la varianza resulta ser menor. En estos casos, donde los parámetros ya no son insesgados, es el error cuadrático medio (ECM) lo que debe observarse, eligiéndose el que sea mínimo.
2. Actuaciones sobre la muestra. Dado que la multicolinealidad hace aumentar la varianza muestral de los $\hat{\beta}_{MCO}$, se puede intentar actuar sobre la muestra para disminuirla; así, se podrían introducir nuevas observaciones o bien mezclar datos de tipo transversal y temporal. Estas alternativas, sin embargo, en muchas ocasiones no resultan factibles.
3. Establecimiento de restricciones sobre el comportamiento de los parámetros poblacionales. Se puede intentar corregir el problema de la multicolinealidad utilizando toda la información extra-muestral disponible, estableciendo restricciones sobre el comportamiento de los parámetros del modelo. Con ello, se reduciría el número de parámetros a estimar.

 ISBN: 978-84-694-7251-4

4. Transformación de variables. Cuando se manejan, sobre todo, series temporales, la multicolinealidad puede reducirse transformando las variables con la aplicación de primeras diferencias, logaritmos…; sin embargo, estas modificaciones aparte de requerir una justificación económica, pueden conllevar efectos no deseados: reducción de grados de libertad, incumplimiento de alguna hipótesis básica del modelo (como introducción de heteroscedasticidad en la perturbación aleatoria), etc.
5. Técnicas multivariantes. Con frecuencia para tratar el problema de la multicolinealidad se emplean técnicas pertenecientes al análisis multivariante, tales como el análisis factorial y el de componentes principales.

4.3. Aplicación de *EViews* al análisis de errores de especificación y multicolinealidad en el modelo.-

En este apartado vamos a ver cómo analizar con *EViews* algunos de los errores de especificación que puede presentar un modelo, así como la posible existencia de multicolinealidad entre las variables explicativas del mismo. Para ello resolveremos, a modo de ejemplo, el Ejercicio nº 40 del Boletín del presente Tema, donde el modelo considerado ya se especificó en una sesión anterior de trabajo con *EViews* a partir del archivo ***tests.wf1***, disponible en el espacio reservado a la Asignatura en *WebCT*.

Al abrir este fichero (*FILE / OPEN / WORKFILE*), veremos que este fichero contenía los datos de las variables siguientes, correspondientes al periodo 1980-2007 (*Figura 1*):

- **CONSUMO**: Tasas de variación anuales del consumo privado no alimenticio en términos reales de 1992
- **PRECIO**: Tasas de variación anuales de los precios del consumo privado no alimenticio con base 1992
- **RENTA**: Tasas de variación anuales de la renta familiar disponible en términos reales de 1992
- **EMPLEO**: Tasas de variación anuales del empleo total
- **PRECA**: Precariedad en el mercado de trabajo (porcentaje de contratos temporales sobre total de contratos)
- **TIR**: Tipos de interés reales en base 1992

A partir de éstas, se especificó el modelo:

$$CONSUMO = \beta_1 + \beta_2 PRECIO + \beta_3 RENTA + \beta_4 EMPLEO + \beta_5 PRECA + \beta_6 TIR + u$$

Su estimación por MCO, mediante *QUICK / ESTIMATE EQUATION*, nos daba como resultado lo mostrado en la *Figura 2*.

 ISBN: 978-84-694-7251-4

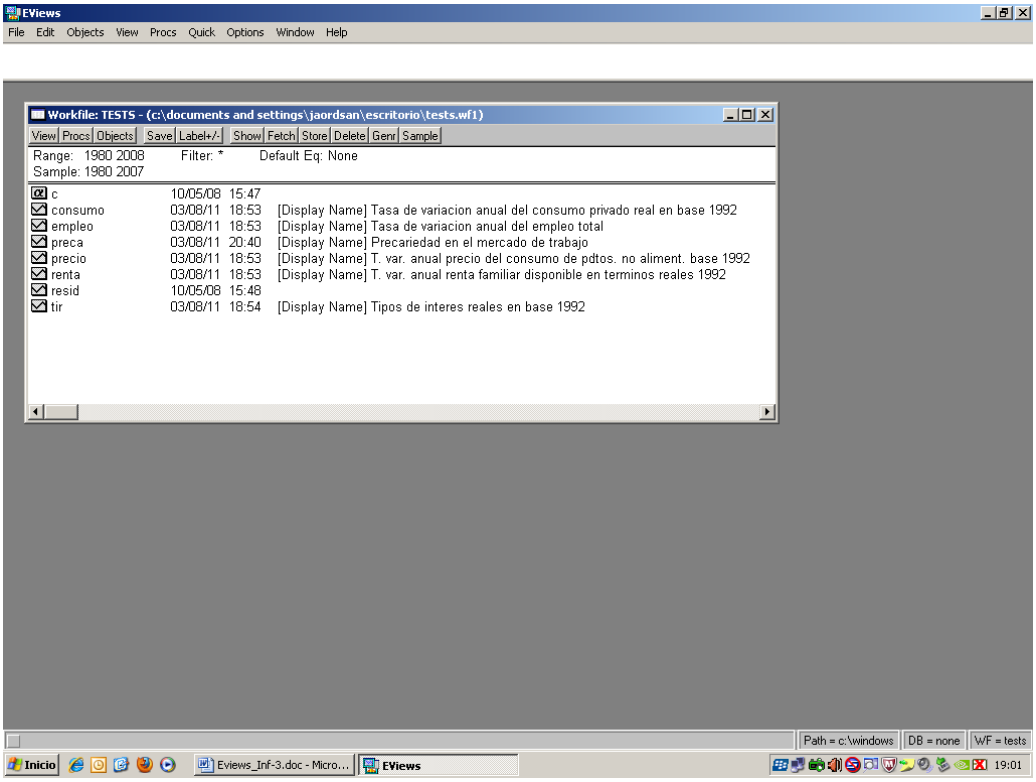


Figura 1

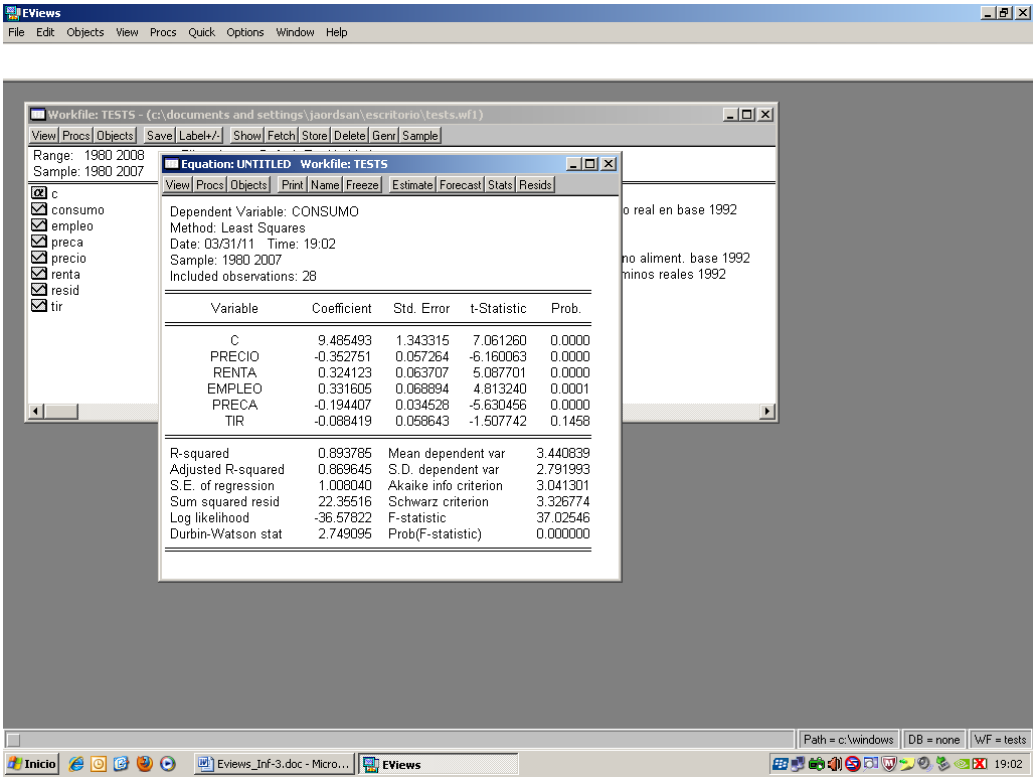


Figura 2

 ISBN: 978-84-694-7251-4

Podemos guardarlo, pulsando el botón *NAME*, con el nombre REG, por ejemplo.[2]

El análisis inicial de los resultados obtenidos nos llevó a comprobar cómo los signos de todos los coeficientes de regresión parecían correctos.

Asimismo, del estudio de la significatividad individual de las variables explicativas, a través de los *p-valores* asociados a los correspondientes estadísticos *t*-Student, podía deducirse que, con la excepción de la variable TIR, todas ellas eran estadísticamente significativas a un nivel de confianza incluso del 99%. En el caso de TIR, ésta no lo resulta siquiera para un $\alpha = 10\%$.

Respecto a la significatividad global del modelo, el *p-valor* (0,000000) asociado al estadístico *F* de este contraste (37,02546) evidenciaba que así era, para un nivel de confianza prácticamente del 100%.

En cuanto a la bondad del ajuste, el valor del coeficiente de determinación (0,893785) reflejaba que el ajuste resultaba muy aceptable. Por su parte, el valor del coeficiente de determinación corregido (0,869645), no sólo reflejaba este hecho, sino también que no existían problemas importantes de grados de libertad, dado que su valor no había sufrido un gran descenso en relación al original.

Junto a todo esto, el contraste que se hizo posteriormente sobre la normalidad de la perturbación aleatoria del modelo, a través del test de Jarque-Bera, vino a confirmar que efectivamente *u* resultaba normal.

Llegados a este punto, y antes de dar definitivamente por bueno nuestro modelo, podemos plantearnos si el modelo presenta algún tipo de error de especificación.

Según se ha podido comprobar al analizar la significatividad individual de cada una de las variables del modelo, quizás no se debería considerar la de tipos de interés (TIR). Para analizar si esta variable, incluida en la especificación inicial, es necesaria o no, podemos aplicar el *test de variables irrelevantes o redundantes*. Para llevar a cabo este test en *EViews*, dentro de nuestro modelo estimado, debemos seleccionar la opción *VIEW* / *COEFFICIENT TESTS* / *REDUNDANT VARIABLES* y escribir el nombre de la variable (TIR) en el cuadro de diálogo que surge (*Figura 3*).

La *Figura 4* nos ofrece el resultado del test, bajo la hipótesis nula de que el coeficiente de la variable seleccionada es cero, a través del estadístico *F* (y el ratio de verosimilitud o *LR* (*Log likelihood ratio*), que no consideraremos), además del resultado de realizar la regresión del modelo restringido, es decir, sin incluir la variable seleccionada. Como sabemos, el estadístico *F* compara la suma de cuadrados residuales calculada con y sin restricciones impuestas; si las restricciones planteadas son válidas, la diferencia entre los dos valores será mínima y, por tanto, el valor de *F* será pequeño (conduciendo ello a aceptar la hipótesis nula). Este estadístico tiene como grados de libertad del numerador

[2] Si guardamos el fichero en la sesión de *EViews* que hemos referido, podemos entonces recuperarlo y comenzar a trabajar en este punto.

el número de restricciones de coeficientes establecido en la hipótesis nula (en este caso, 1) y en el denominador, los grados de libertad de la regresión total (en este caso, 22).

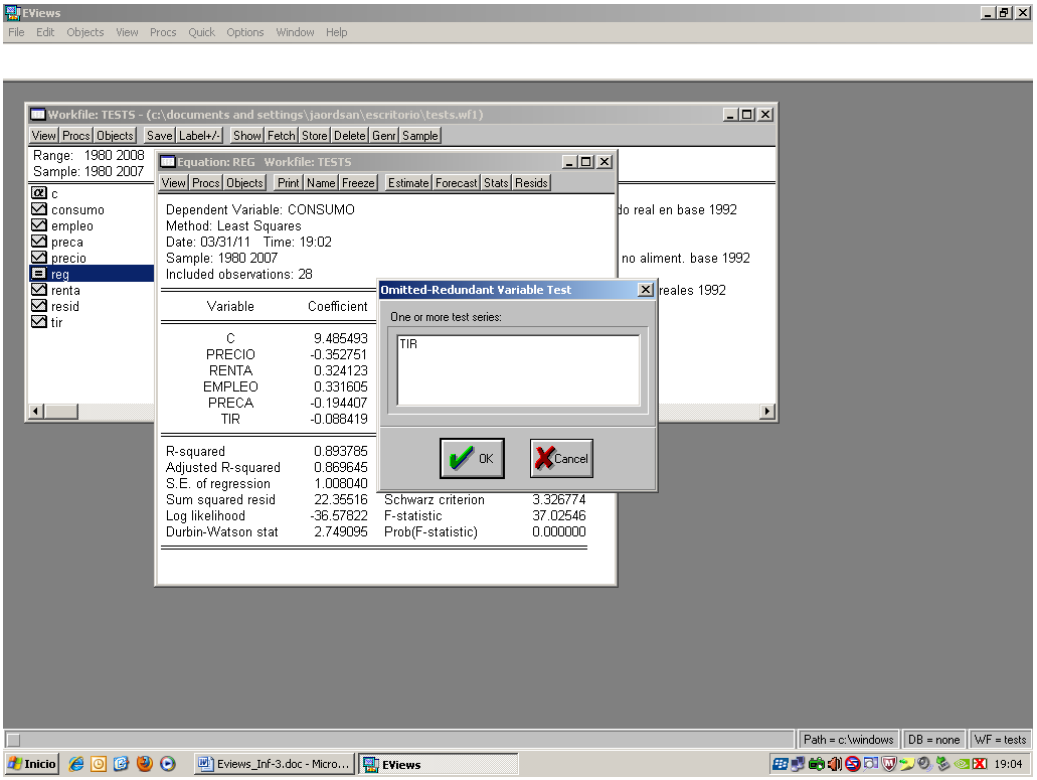


Figura 3

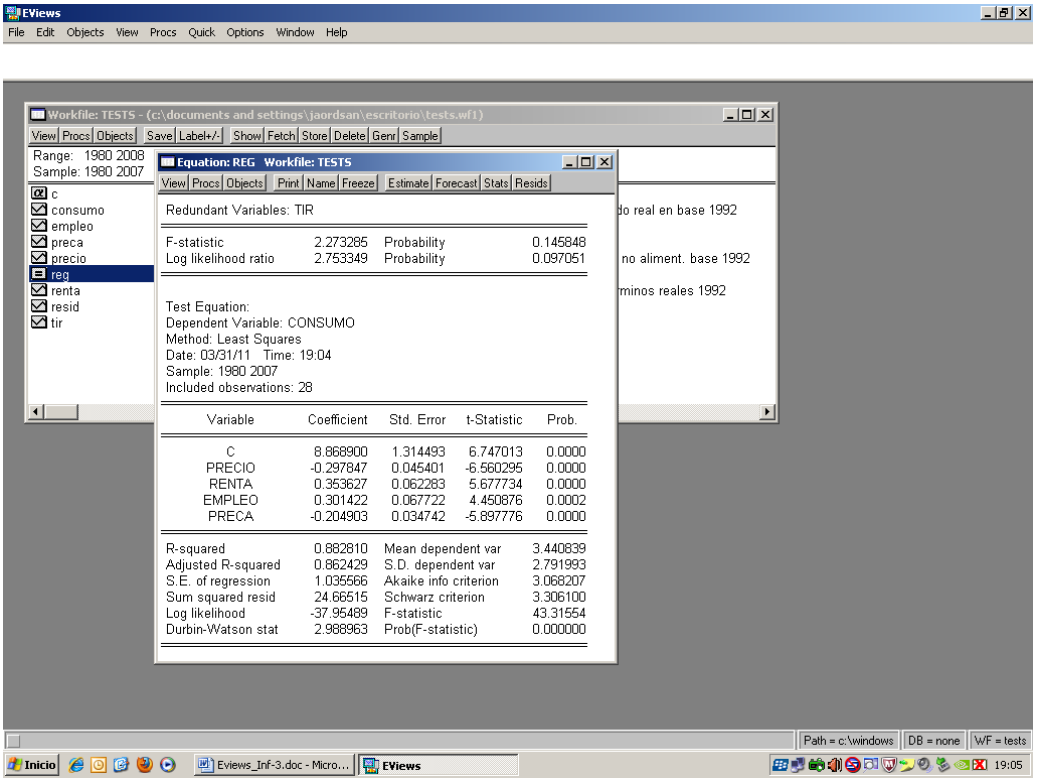


Figura 4

En este caso, los resultados obtenidos nos conducen a aceptar la hipótesis nula, pues el *p-valor* asociado al estadístico *F* nos indica que el nivel de significación mínimo al que se puede rechazar la hipótesis nula es del 14,58%, o bien el nivel de confianza máximo para rechazar dicha hipótesis es del 85,42%. Por tanto, la variable TIR no es necesaria en nuestro modelo. Si bien la nueva especificación perdería algo de bondad de ajuste (evidenciada por la comparación de los correspondientes valores del coeficiente de determinación corregido: 0,869645 frente a 0,862429), dicha pérdida no sería relevante. Así pues, podemos eliminarla de nuestra estimación. Lo haremos editando el modelo en el botón *ESTIMATE* de nuestra ecuación REG y borrando simplemente esta variable.

Tras llevar a cabo esta "depuración" de nuestro modelo, resulta que nos facilitan los datos de una nueva variable que consideramos que podría ser relevante en el mismo:

- **TIPIMP**: Tipo medio impositivo en términos reales con base 1992

Si fuese significativa, su no inclusión representaría otro tipo de error en la especificación del modelo. La omisión de una variable explicativa relevante en nuestro modelo tendría, además, consecuencias más graves que la inclusión de una variable irrelevante.

Los valores de dicha variable se encuentran a nuestra disposición en un archivo de Excel, denominado ***omitida.xls***, que se encuentra en el espacio de la Asignatura en *WebCT*. Para incorporarlo a nuestro análisis deberemos importarlo a nuestro fichero de trabajo. Como sabemos, para ello deberemos seleccionar, desde el menú principal del fichero de trabajo, la opción: *FILE / IMPORT / READ TEXT-LOTUS-EXCEL…*

La *Figura 5* muestra la pantalla del menú correspondiente a la importación de ficheros *Excel* (*Excel Spreadsheet Import*), con las opciones correspondientes seleccionadas.

Una vez importada TIPIMP, vamos a comprobar si esta nueva variable debe estar presente en el modelo; es decir, si es una variable relevante que hasta el momento hemos omitido en su especificación. Para verlo, aplicaremos el *test de variables omitidas*, que establece como hipótesis nula que la variable o variables a considerar en el nuevo modelo no son significativas. El estadístico *F* de este contraste se calcula a partir de la diferencia de la suma de cuadrados residuales de la regresión inicial (que sería la restringida) y de la regresión con las variables que se omitieron en principio. Sus grados de libertad son en este caso 1, 22.

Para la realización de este contraste, dentro de nuestro modelo REG, deberemos seleccionar *VIEW / COEFFICIENT TESTS / OMITTED VARIABLES* y seguidamente introduciremos la variable TIPIMP (*Figura 6*).[3]

[3] En este test debemos tener presente que la nueva variable que introduzcamos en el modelo ha de tener el mismo número de observaciones que las de la especificación inicial (en este caso, datos de 1980 a 2007).

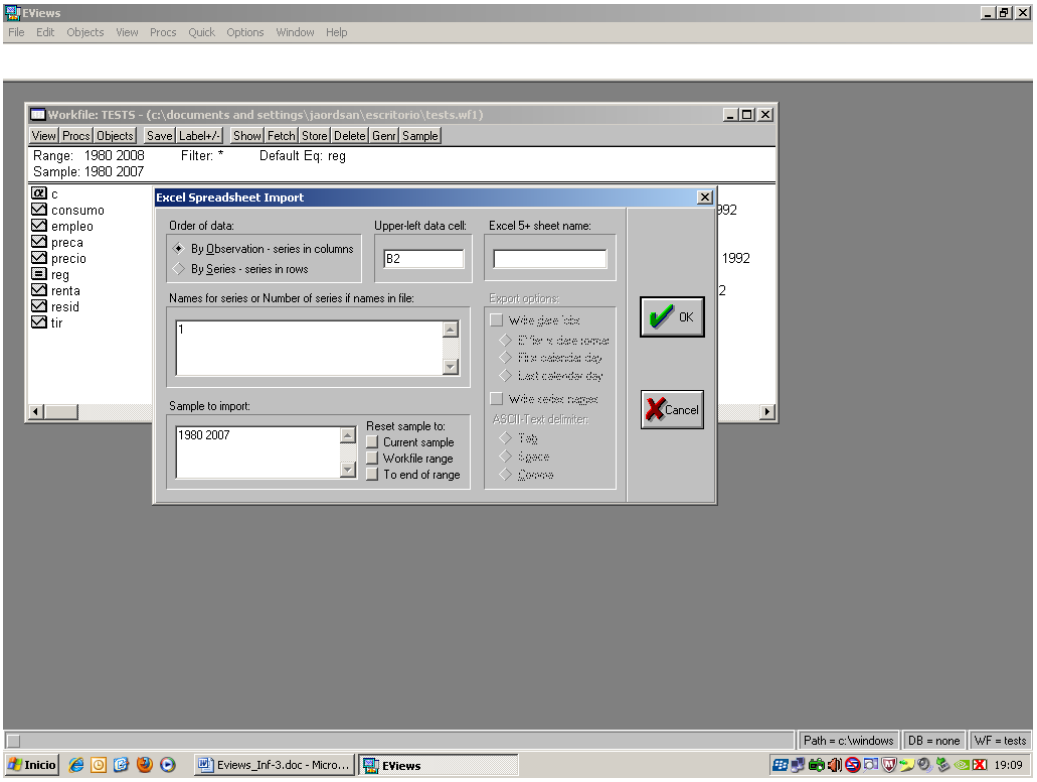


Figura 5

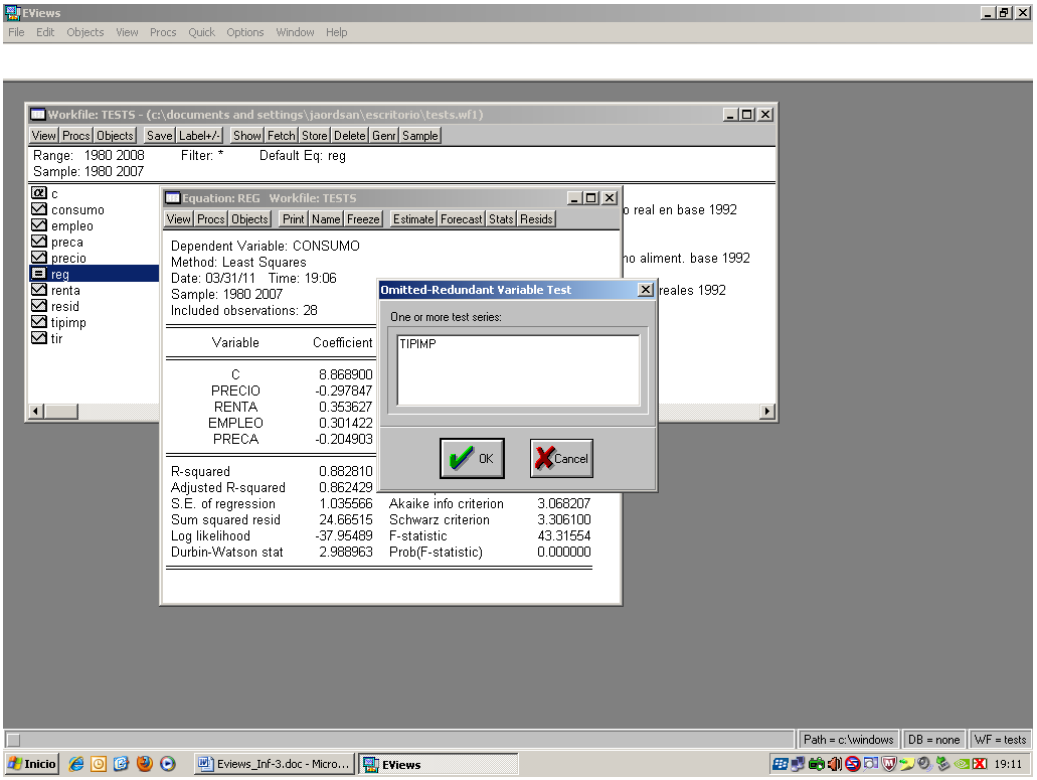


Figura 6

Los resultados se pueden ver en la *Figura 7*. Aparte de los estadísticos y *p-valores* asociados a *F* y *LR* (aunque éste último lo obviaremos), la aplicación de este test

incluye la regresión realizada añadiendo esta nueva variable en la estimación. De este modo, podemos comprobar que:

- El *p-valor* asociado al estadístico *F* es pequeño; a un nivel de significación del 5% se puede rechazar la hipótesis nula que establece que la variable TIPIMP no es significativa. Por tanto, se trata de una variable que había sido omitida.
- El signo que presenta TIPIMP es correcto (negativo), tal como era de esperar.
- La variable TIPIMP es estadísticamente significativa para un $\alpha = 5\%$.
- El coeficiente de determinación corregido ha mejorado: 0,886568.

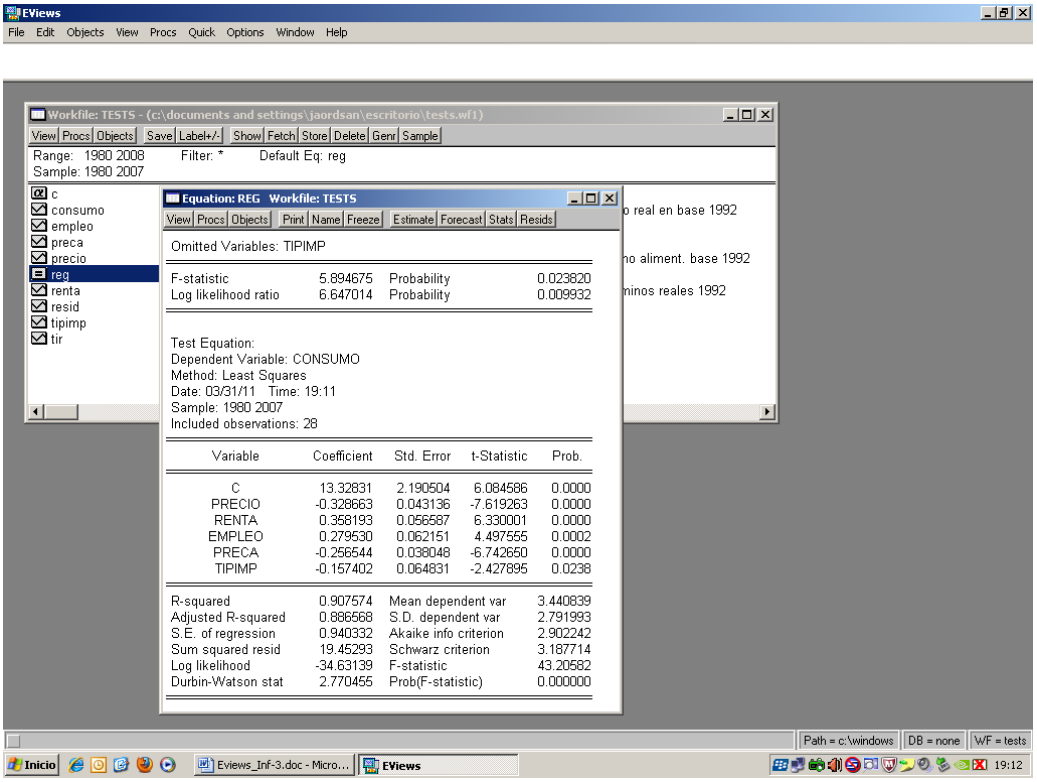


Figura 7

En definitiva, hemos visto cómo nuestro modelo debe incluir la variable TIPIMP. Para llevar a cabo de manera definitiva el nuevo ajuste, dentro de nuestro modelo REG, pulsaremos una vez más la opción *ESTIMATE* y añadiremos la variable TIPIMP. El resultado de la estimación puede apreciarse en la *Figura 8*.

Otro tipo de test destinado a detectar problemas en la especificación del modelo es el denominado *Test RESET de Ramsey*. Este test permite detectar la omisión de variables y la elección de una forma funcional inadecuada.

La realización del *Test de Ramsey* en *EViews* se hace, también dentro de la *Ventana de Ecuación*, a través de *VIEW / STABILITY TESTS / RAMSEY RESET TEST* (*Figura 9*).

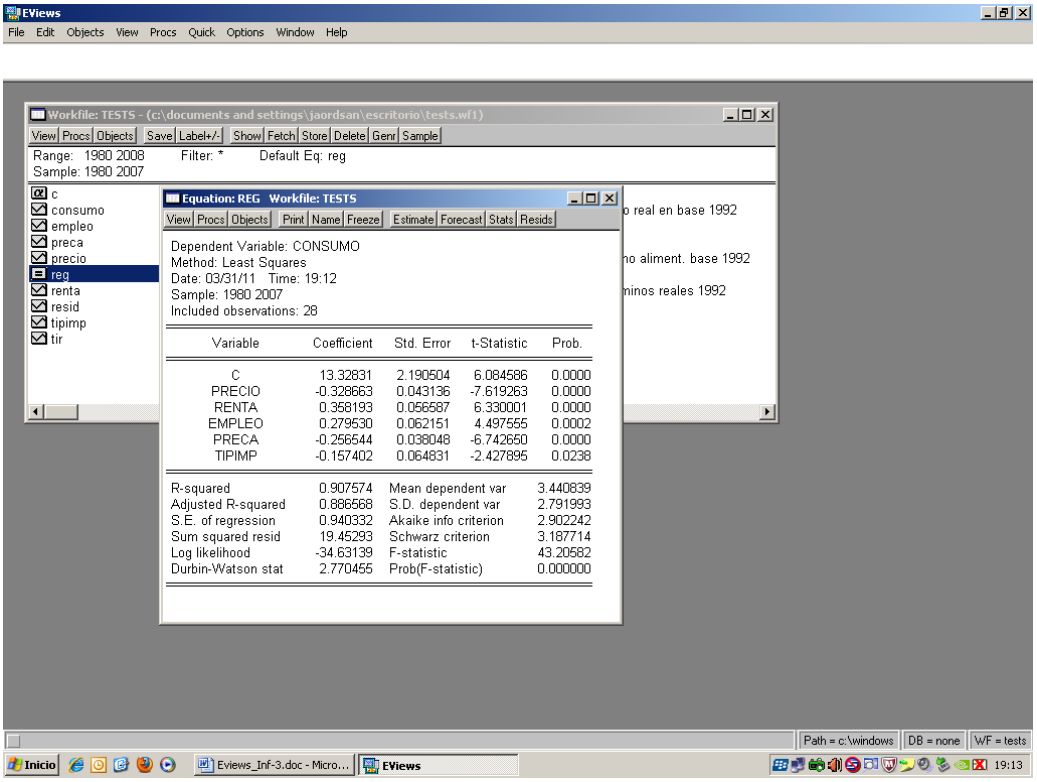


Figura 8

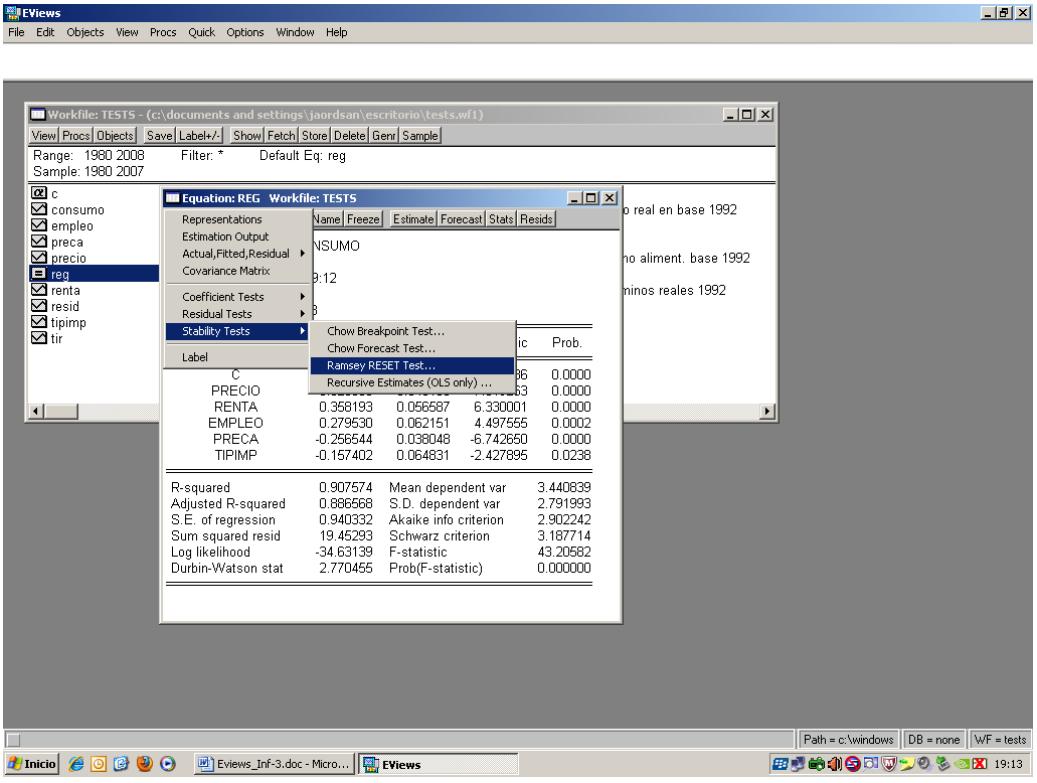


Figura 9

La base de este test reside en la comparación de la especificación inicial del modelo con una nueva que se plantea como alternativa y que añade, a las variables explicativas originales, potencias de la estimación de la variable endógena; de este modo, se pretenden capturar posibles relaciones sistemáticas existentes entre los residuos y las estimaciones de Y y que no son recogidas por el modelo inicial. En este contraste se emplea un estadístico F cuyo cálculo se basa en la diferencia entre los coeficientes de determinación del nuevo modelo y el del original. La aceptación de la hipótesis nula supone asumir que el modelo inicial resulta aceptable; por el contrario, su rechazo implica pensar que el modelo está mal especificado.

En este caso, hemos añadido a nuestra especificación 2 potencias de $\hat{Y}$: $\hat{Y}^2$ e $\hat{Y}^3$. Normalmente con este número ya resulta suficiente para obtener conclusiones (*Figura 10*).

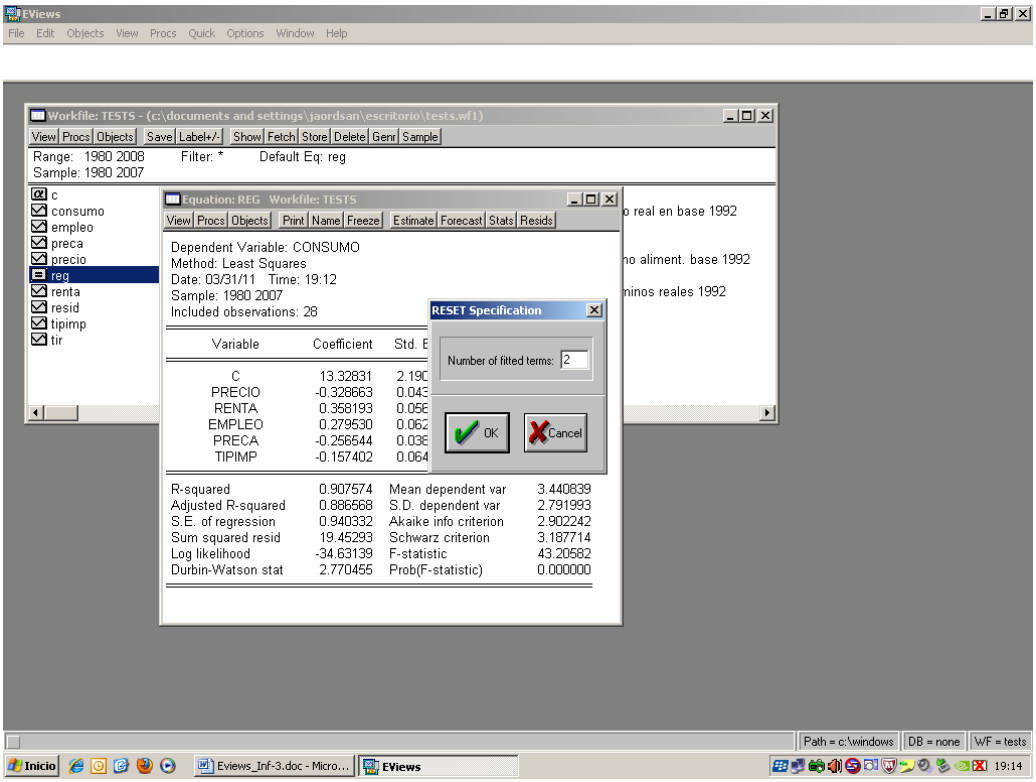


Figura 10

En la *Figura 11* aparece el resultado de este contraste, apreciándose que la hipótesis nula se acepta para un nivel máximo de significación del 56,50%. Así pues, se acepta la hipótesis nula: nuestra última especificación del modelo resulta correcta.

 ISBN: 978-84-694-7251-4

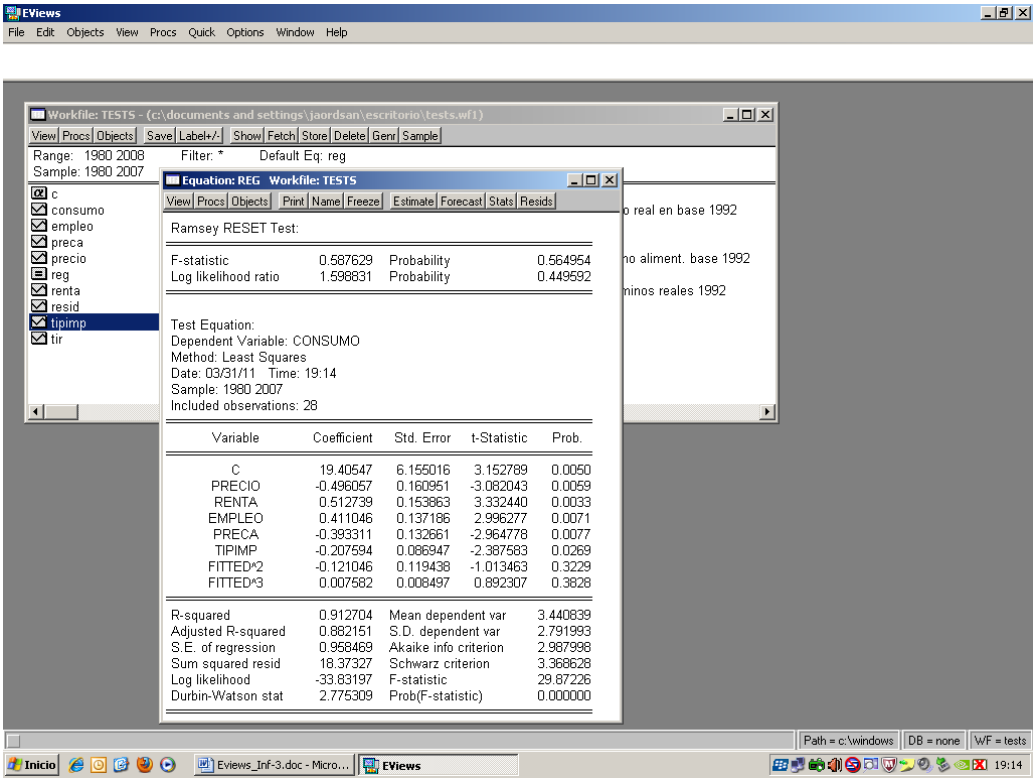


Figura 11

Por último, podemos analizar si nuestro modelo evidencia algún problema de multicolinealidad. Como sabemos, ante la presencia de ésta los coeficientes de regresión estimados por MCO siguen siendo ELIO, pero sin embargo la potencia de los contrastes de significación individual de las variables explicativas disminuyen drásticamente, lo cual puede tener consecuencias para la correcta especificación final de un modelo.

La obtención de un R^2 elevado para el modelo (y, por tanto, de significatividad global de éste manifestada a través del estadístico *F*) y simultáneamente de pocos estadísticos *t*-Student significativos de las variables explicativas, resulta un claro indicio de existencia de multicolinealidad, si bien no es del todo concluyente. Puesto que en este ejercicio no son éstas las circunstancias que se dan, ello parece sugerirnos que no tenemos problemas de multicolinealidad.

Adicionalmente, podemos emplear otro método de detección, consistente en el estudio de los coeficientes de correlación lineal simple entre las variables explicativas. Valores altos ($|R| \geq 0{,}8$) son condición suficiente, pero no necesaria, para afirmar que existe multicolinealidad en el modelo. La *Figura 12* muestra cómo obtener con *EViews* la matriz de coeficientes de correlación lineal de las variables explicativas: debemos seleccionar de una en una todas éstas (manteniendo pulsada la tecla *Ctrl*) y tras esto, pulsando el botón derecho del ratón, se elige la opción *OPEN / AS GROUP*.

Una vez aquí (*Figura 13*), en *VIEW*, podremos escoger realizar la matriz de correlaciones (*CORRELATIONS*). Al analizar los resultados de ésta (*Figura 14*), no

parece que haya problemas de multicolinealidad, pues no hay ningún valor absoluto que se sitúe por encima de 0,8. Con esto, finaliza así el presente ejercicio.

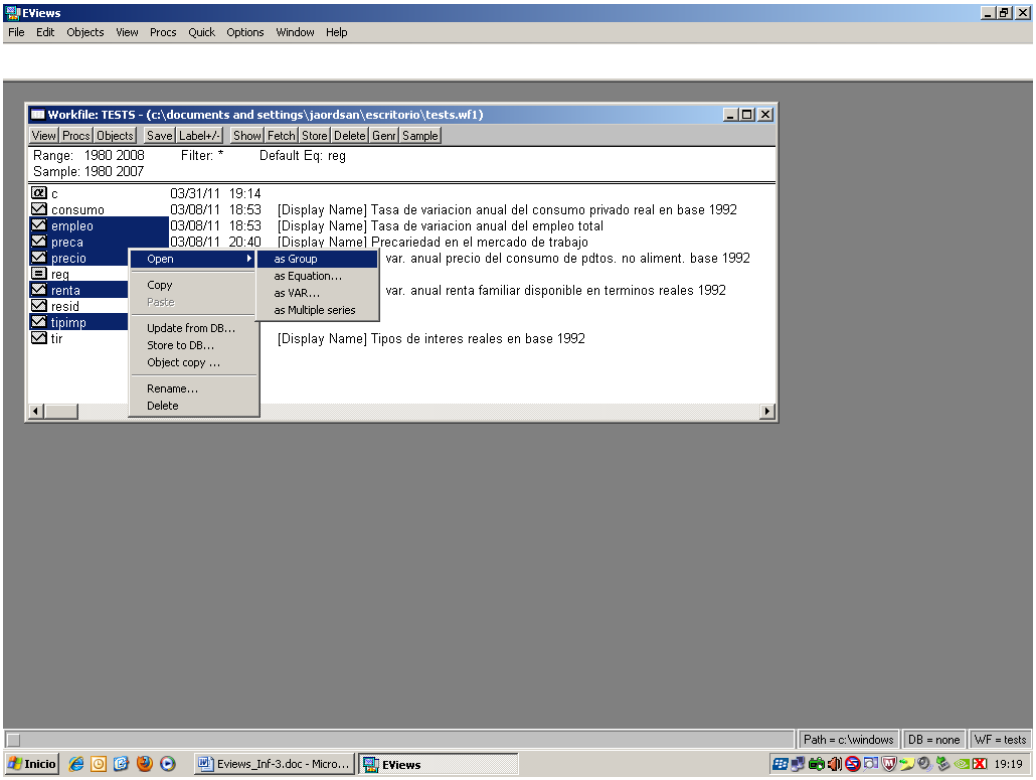


Figura 12

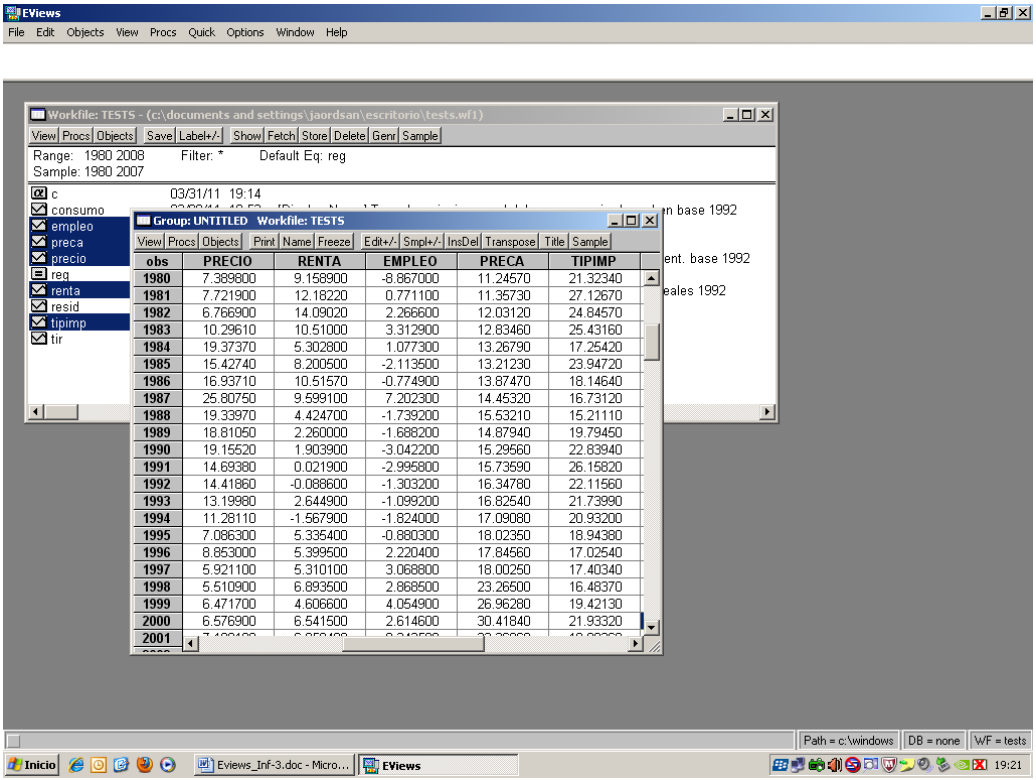


Figura 13

 ISBN: 978-84-694-7251-4

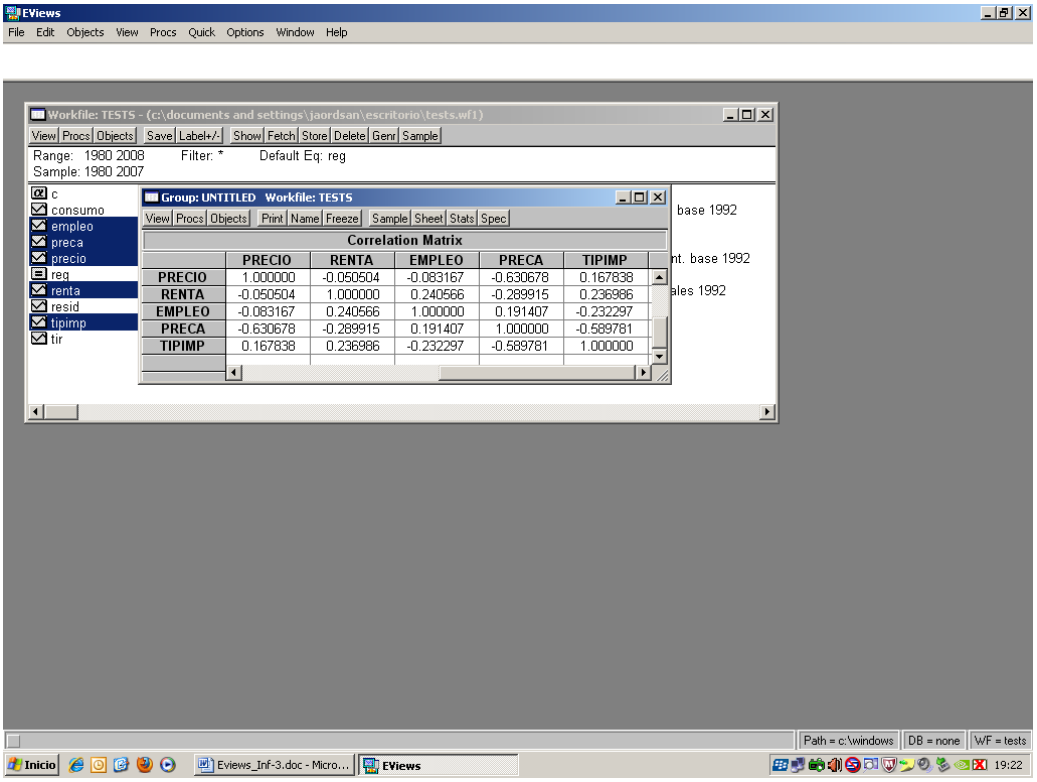


Figura 14

4.4. Heteroscedasticidad y autocorrelación. Propiedades de los estimadores MCO ante una perturbación no esférica. Estimación por mínimos cuadrados generalizados (MCG).-

Heteroscedasticidad y autocorrelación

Considérese el modelo clásico de regresión lineal general:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i \qquad \forall j = 1,2,...,k \qquad \forall i = 1,2,...,n\,,$$

o bien, mediante su expresión matricial: $Y = X\beta + u$.

Hasta este momento, todo nuestro análisis se ha basado en el hecho de que, para todas las observaciones *i* del modelo, la perturbación aleatoria ha presentado entre sus principales supuestos homoscedasticidad e incorrelación, lo cual se ha concretado en que su matriz de varianzas-covarianzas resulta escalar de orden *n* x *n*:

$$Var - Cov(u) = \begin{pmatrix} \sigma_u^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_u^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_u^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma_u^2 \end{pmatrix} = \sigma_u^2 \cdot \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{pmatrix} = \sigma_u^2 \cdot I.$$

Es decir, todos los elementos de la diagonal principal son iguales (homoscedasticidad: varianza constante para todas las observaciones) y el resto de elementos de la matriz valen todos ellos cero (incorrelación: covarianzas correspondientes a distintas observaciones iguales a cero).

En este escenario, el investigador se encuentra con $k+1$ parámetros a estimar: $\beta_1, \beta_2, \ldots \beta_k$ y σ_u^2. Como sabemos, el método de MCO nos ofrece para ello estimadores ideales: ELIO (insesgados y óptimos, además de lineales). Toda la teoría inferencial que se ha desarrollado en torno a nuestro modelo se ha basado de manera fundamental en la homoscedasticidad e incorrelación de la perturbación aleatoria.

Sin embargo, estos supuestos son sólo eso: supuestos. En la realidad, podemos encontrarnos con que alguno de ellos (o incluso los dos) no se cumpla. Cabría entonces evaluar cuáles son las consecuencias que esto tiene sobre nuestro modelo, tanto en lo que a su estimación se refiere como respecto a los aspectos inferenciales del mismo (contrastes de hipótesis e intervalos de confianza).

Así pues, por un lado podríamos tener que la varianza de la perturbación aleatoria fuese distinta de unas observaciones a otras, esto es: $Var(u_i) = \sigma_{u_i}^2$, con $\sigma_{u_i}^2 \neq \sigma_{u_j}^2, \ \forall i \neq j$.

De esta forma, se quebraría el supuesto de homoscedasticidad y estaríamos ante lo que se denomina heteroscedasticidad de la perturbación aleatoria, que es una situación característica sobre todo de modelos de series de corte transversal, si bien también puede encontrarse en series de datos temporales.

La matriz de varianzas-covarianzas de u sería en este caso:

$$Var - Cov(u) = \begin{pmatrix} \sigma_{u_1}^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_{u_2}^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_{u_3}^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma_{u_n}^2 \end{pmatrix}.$$

En esta situación estaríamos ante un importante problema de resolución matemática, de sobreparametrización, pues tendríamos $k+n$ parámetros a estimar ($\beta_1, \beta_2, \ldots \beta_k$ y $\sigma_{u_1}^2, \sigma_{u_2}^2, \ldots, \sigma_{u_n}^2$), en tanto que sólo tendríamos n observaciones. Para abordar este problema sería preciso establecer algún tipo de supuesto que permitiese, de algún modo, reducir el número de parámetros a estimar, de forma que finalmente fuese menor que n.

Ante la presencia de heteroscedasticidad, la matriz de $Var - Cov(u)$ es una matriz diagonal, que podría expresarse de la forma:

$$Var-Cov(u)=\sigma^2\cdot\begin{pmatrix} \sigma^2_{u_1}/\sigma^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2_{u_2}/\sigma^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma^2_{u_3}/\sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2_{u_n}/\sigma^2 \end{pmatrix}=\sigma^2\cdot\Omega$$

Obsérvese cómo Ω sería una matriz diagonal con un formato característico.

Por otro lado, podríamos plantearnos un escenario en el que u_i no fuese incorrelada para todas las observaciones; esto es, que existiese <u>autocorrelación</u>: $Cov(u_i,u_j)=\sigma_{ij}\neq 0$, $i\neq j$. Ello conllevaría que los elementos que no pertenecen a la diagonal principal de la matriz de $Var-Cov(u)$ no tendrían por qué ser todos cero, por lo que ésta ya no sería diagonal[4]:

$$Var-Cov(u)=\begin{pmatrix} \sigma^2_u & \sigma_{12} & \cdots & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma^2_u & \sigma_{23} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \sigma^2_u & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \cdots & \sigma^2_u \end{pmatrix}.$$

Ante la presencia de autocorrelación, el investigador ha de estimar $k+1+\dfrac{n^2-n}{2}$ parámetros, que evidentemente es un número mayor que el de observaciones muestrales n: $\beta_1,\beta_2,\ldots\beta_k,\sigma^2_u$ más los $\dfrac{n^2-n}{2}$ elementos diferentes que se hallan por encima de la diagonal principal de la matriz[5]. Por tanto, al igual que sucedía con la heteroscedasticidad, se deberá establecer algún tipo de supuesto que conduzca a reducir dicho número de incógnitas hasta que sea menor que n.

<u>La autocorrelación es una situación que se registra especialmente en modelos referidos a series temporales</u>, viniendo motivada por la existencia de ciclos y tendencias, relaciones dinámicas, etc. Igualmente, también puede estar presente en series transversales; por

[4] Nótese que al plantear autocorrelación estamos suponiendo homoscedasticidad, lo mismo que anteriormente cuando planteamos heteroscedasticidad supusimos incorrelación. Es decir, estamos considerando la quiebra de estos supuestos por separado. Si nos encontrásemos con ambas situaciones de forma simultánea, procederíamos primero suponiendo sólo una y, una vez solucionada, afrontando la otra.

[5] Los elementos que hay por debajo de la diagonal principal de la matriz son los mismos que hay por encima de la misma, dado que la matriz es simétrica: $\sigma_{ij}=\sigma_{ji}$, $\forall i\neq j$.

 ISBN: 978-84-694-7251-4

ejemplo, por haber omitido variables relevantes en la especificación del correspondiente modelo.

Véase cómo ante la existencia de autocorrelación en la perturbación aleatoria del modelo, la matriz de varianzas-covarianzas de ésta se podría expresar de modo que:

$$Var-Cov(u)=\sigma_u^2\cdot\begin{pmatrix} 1 & \sigma_{12}/\sigma_u^2 & \cdots & \cdots & \sigma_{1n}/\sigma_u^2 \\ \sigma_{12}/\sigma_u^2 & 1 & \sigma_{23}/\sigma_u^2 & \cdots & \sigma_{2n}/\sigma_u^2 \\ \vdots & \vdots & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n}/\sigma_u^2 & \sigma_{2n}/\sigma_u^2 & \cdots & \cdots & 1 \end{pmatrix}=\sigma_u^2\cdot\Omega$$

La matriz Ω tendría aquí también una forma característica, si bien distinta a la del caso de heteroscedasticidad.

En definitiva, ante la presencia de heteroscedasticidad o de autocorrelación en la perturbación aleatoria u del modelo, tendríamos que su matriz de varianzas-covarianzas se podría escribir de forma general:

$$\boxed{Var-Cov(u)=\sigma^2\cdot\Omega,\ \text{siendo}\ \Omega\neq I}.$$

Según esto, como puede apreciarse, el cumplimiento de los supuestos de homoscedasticidad e incorrelación en la perturbación da lugar a un caso particular de matriz de $Var-Cov(u)$ donde $\Omega=I$ y, además, el parámetro σ^2 se correspondería con el valor de la varianza de u, constante para todas las observaciones muestrales[6], esto es: $\sigma^2=\sigma_u^2$.

Cuando la perturbación aleatoria cumple estos supuestos, se dice que es esférica. En los casos en que incumple al menos alguno de ellos, es decir, que presenta heteroscedasticidad y/o autocorrelación, se habla de perturbación no esférica.

Propiedades de los estimadores MCO ante una perturbación no esférica

Llegados a este punto, nos planteamos qué sucede con la estimación de los parámetros del modelo por el método de MCO si la perturbación aleatoria resulta ser no esférica; esto es, nos planteamos el modelo de regresión lineal $Y=X\beta+u$, donde $Var-Cov(u)=\sigma^2\cdot\Omega$, con $\Omega\neq I$.

[6] Obsérvese que, de acuerdo con las hipótesis establecidas, esto último también sería cierto aun presentando autocorrelación la perturbación aleatoria.

En esta situación, el estimador MCO de β sigue siendo una solución del sistema de ecuaciones normales: $X'X\hat{\beta} = X'Y$, por lo que si la matriz $(X'X)$ es invertible, la solución única a dicho sistema es: $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$. Así pues, podemos seguir estimando el modelo por MCO.

El siguiente paso será comprobar si el estimador $\hat{\beta}_{MCO}$ sigue conservando sus propiedades ideales; es decir, si en este escenario sigue siendo ELIO: insesgado y óptimo (de mínima varianza). Para ello, vamos a analizar qué sucede con el valor esperado y la matriz de varianzas-covarianzas de $\hat{\beta}_{MCO}$.

Como ya sabemos, $\hat{\beta}_{MCO}$ puede expresarse como: $\hat{\beta}_{MCO} = \beta + (X'X)^{-1}X'u$.

A partir de aquí, si calculamos su valor esperado, tenemos que:

$$E(\hat{\beta}_{MCO}) = E(\beta + (X'X)^{-1}X'u) = \beta + (X'X)^{-1}X'E(u) = \beta + \theta = \beta .$$

Con lo que vemos que $\hat{\beta}_{MCO}$ sigue siendo un estimador insesgado.

En cuanto a su matriz de varianzas-covarianzas:

$$Var - Cov(\hat{\beta}_{MCO}) = E\left[(\hat{\beta}_{MCO} - E(\hat{\beta}_{MCO}))\cdot(\hat{\beta}_{MCO} - E(\hat{\beta}_{MCO}))'\right] = E\left[(\hat{\beta}_{MCO} - \beta)\cdot(\hat{\beta}_{MCO} - \beta)'\right] =$$

$$= E\left[((X'X)^{-1}X'u)\cdot((X'X)^{-1}X'u)'\right] = E\left[(X'X)^{-1}X'u\,u'X(X'X)^{-1}\right] =$$

$$= (X'X)^{-1}X'E[u\,u']X(X'X)^{-1} = (X'X)^{-1}X'Var - Cov(u)X(X'X)^{-1} =$$

$$= (X'X)^{-1}X'\sigma^2\cdot\Omega\,X(X'X)^{-1} = \sigma^2\cdot(X'X)^{-1}X'\Omega\,X(X'X)^{-1}.$$

A la vista de este resultado, deducimos que la expresión general de la matriz de $Var - Cov(\hat{\beta}_{MCO})$ resulta muy distinta a la que ya conocíamos cuando u era esférica:

$$Var - Cov(\hat{\beta}_{MCO}) = \sigma_u^2\cdot(X'X)^{-1}.$$

Nótese que si u fuese esférica, estaríamos en el caso particular de que $\Omega = I$ y $\sigma^2 = \sigma_u^2$, y entonces ambas expresiones coincidirían. Por tanto, si se utiliza ésta última en los procesos inferenciales donde resulte preciso, cuando la u no sea esférica, resultaría del todo incorrecto, pudiendo conducir a errores importantes.

En definitiva, tenemos que: $\boxed{\hat{\beta}_{MCO} \to N_k(\beta;\sigma^2\cdot(X'X)^{-1}X'\Omega\,X(X'X)^{-1})}$.

El problema que nos encontramos sin embargo con el método de estimación de MCO es que esta matriz de $Var - Cov(\hat{\beta}_{MCO})$ aun siendo correcta, no resulta la menor posible, por lo que el estimador MCO deja de ser ELIO.

 ISBN: 978-84-694-7251-4

Para finalizar, tenemos que la estimación insesgada del otro parámetro relevante en nuestro análisis, σ^2, cuando u no es esférica resulta ser:

$$\boxed{\hat{\sigma}^2_{MCO} = \frac{e'_{MCO}\,\Omega^{-1}\,e_{MCO}}{n-k}},$$

donde: $e_{MCO} = Y - \hat{Y}_{MCO} = Y - X\hat{\beta}_{MCO}$.

Estimación por mínimos cuadrados generalizados (MCG)

Según acabamos de mostrar, la estimación por MCO de los parámetros del modelo ya no resulta de mínima varianza ante la presencia de heteroscedasticidad y/o autocorrelación en la perturbación aleatoria. Esto, unido a que además la matriz de $Var-Cov\left(\hat{\beta}_{MCO}\right)$ que estemos utilizando pueda no ser la correcta, lo que invalidaría todas las conclusiones de los procesos de inferencia estadística referidos al modelo econométrico, nos hace pensar por un momento que todo lo estudiado hasta ahora sobre el modelo clásico de regresión lineal ha podido resultar en vano. Sin embargo, veremos que no es así.

En estas circunstancias, sería interesante poder transformar nuestro modelo econométrico en otro, de tal forma, que los nuevos coeficientes fuesen los mismos que los del modelo original, pero cuya perturbación aleatoria fuese esférica, es decir, resultase homoscedástica e incorrelada. De este modo, si aplicásemos el método de estimación de MCO sobre este modelo transformado, todos los supuestos de éste se cumplirían y, por el Teorema de Gauss-Markov, tendríamos que los estimadores obtenidos serían ELIO y podríamos seguir asegurando todos los resultados inferenciales derivados de los mismos.

Así pues, si partimos del modelo de regresión lineal general, $Y = X\beta + u$, la idea es poder transformarlo de manera que, si pre-multiplicamos los valores de todas las variables presentes en el modelo por una matriz de coeficientes P, cuadrada de orden *n*, es decir: $PY = (PX)\beta + Pu$, la nueva perturbación aleatoria resultante (Pu) fuese esférica.

Tendríamos, por tanto, un nuevo modelo lineal que sería:

$$\boxed{Y^* = X^*\beta + u^*},$$

donde: $Y^* = PY$, $X^* = PX$, $u^* = Pu$ y el objetivo es que u^* sea esférica, es decir, homoscedástica e incorrelada.

Bajo estas condiciones, obsérvese que, gracias a la linealidad del modelo, el vector de coeficientes β sería el mismo que el del modelo original, pero con la diferencia de que

al aplicar el método de MCO sobre este nuevo modelo transformado estaríamos en las mismas condiciones que las conocidas de un modelo clásico.

Respecto a las nuevas variables transformadas Y^* y $(X_1^*, X_2^*, \ldots, X_k^*)$ que integran X^*, cabe reseñar que éstas se obtendrían como combinaciones lineales de las variables originales Y y $(X_1, X_2, \ldots, X_k)$ que conforman X, por lo que no tendrían un significado claro.[7]

Y en cuanto a la nueva perturbación aleatoria, u^*, tendremos que su valor esperado y su matriz de varianzas-covarianzas son:

- $E(u^*) = E(Pu) = PE(u) = P\theta = \theta$
- $Var - Cov(u^*) = E[u^* u^{*\prime}] = E[Pu(Pu)'] = E[Puu'P'] = PE[uu']P' =$
  $= PVar - Cov(u)P' = P\sigma^2 \cdot \Omega P' = \sigma^2 \cdot P\Omega P'$

Puesto que lo que perseguimos es que $Var - Cov(u^*)$ sea escalar, nuestro objetivo final será ver qué matriz P debemos elegir para transformar el modelo, de tal manera que verifique que: $P\Omega P' = I$.

Sabiendo que Ω es una matriz simétrica y definida positiva, matemáticamente se puede llegar a demostrar que existe una matriz cuadrada no singular V, de tal modo que: $\Omega = VV'$. Pues bien, la matriz P que buscamos resulta ser:

$$\boxed{P = V^{-1}}.$$

Como se puede ver, esta matriz efectivamente verifica:

$$Var - Cov(u^*) = \sigma^2 \cdot P\Omega P' = \sigma^2 \cdot V^{-1} VV' \left(V^{-1}\right)' = \sigma^2 \cdot V^{-1} VV' (V')^{-1} = \sigma^2 \cdot I .$$

En resumen, cuando estemos ante un modelo $Y = X\beta + u$ con perturbación aleatoria no esférica ($Var - Cov(u) = \sigma^2 \cdot \Omega$, siendo $\Omega \neq I$), deberemos obtener a partir de Ω la matriz V que cumple que $\Omega = VV'$; tomaremos luego V^{-1} para llevar a cabo una transformación lineal del modelo original, de modo que: $Y^* = X^*\beta + u^*$, haciendo $Y^* = V^{-1}Y$, $X^* = V^{-1}X$ y $u^* = V^{-1}u$. La nueva perturbación u^* será entonces esférica, es decir, homoscedástica e incorrelada.

Sobre este modelo transformado, que cumple ya los supuestos de un modelo clásico, aplicaremos seguidamente MCO para su estimación. Este método de estimación (la

[7] Nótese que los elementos de la matriz P son simplemente los coeficientes de dichas combinaciones lineales.

 ISBN: 978-84-694-7251-4

aplicación de MCO sobre el modelo transformado) se denomina método de mínimos cuadrados generalizados (MCG). De este modo, el estimador MCG viene dado por:

$$\boxed{\hat{\beta}_{MCG} = \left(X^{*\prime} X^{*}\right)^{-1} X^{*\prime} Y^{*}}.$$

O, si lo expresamos en función de las variables originales X e Y del modelo[8]:

$$\boxed{\hat{\beta}_{MCG} = \left(X' \Omega^{-1} X\right)^{-1} X' \Omega^{-1} Y},$$

ya que:

$$\hat{\beta}_{MCG} = \left(X^{*\prime} X^{*}\right)^{-1} X^{*\prime} Y^{*} = \left(\left(V^{-1}X\right)'\left(V^{-1}X\right)\right)^{-1}\left(V^{-1}X\right)'\left(V^{-1}Y\right) =$$

$$= \left(X'\left(V^{-1}\right)'\left(V^{-1}X\right)\right)^{-1} X'\left(V^{-1}\right)' V^{-1} Y = \left(X' \Omega^{-1} X\right)^{-1} X' \Omega^{-1} Y .$$

- ***Propiedades de los estimadores MCG de los coeficientes de regresión***

1. El estimador MCG de los coeficientes de regresión satisface el sistema de ecuaciones normales del modelo:

   $$X^{*\prime} X^{*} \hat{\beta}_{MCG} = X^{*\prime} Y^{*},$$

   o bien, teniendo en cuenta las transformaciones lineales $X^{*} = V^{-1} X$ e $Y^{*} = V^{-1} Y$, esto puede expresarse en función de las variables originales X e Y:

   $$\left(X' \Omega^{-1} X\right)\hat{\beta}_{MCG} = X' \Omega^{-1} Y .$$

2. El estimador $\hat{\beta}_{MCG}$ es un vector aleatorio que sigue una distribución de probabilidad normal.

   Dado que $\hat{\beta}_{MCG}$ se puede expresar como: $\hat{\beta}_{MCG} = \beta + \left(X^{*\prime} X^{*}\right)^{-1} X^{*\prime} u^{*}$, y puesto que u^{*} es normal, al depender $\hat{\beta}_{MCG}$ de u^{*} se deduce que $\hat{\beta}_{MCG}$ es también un vector aleatorio normal.

3. Si $E\left(u^{*}\right) = \theta$, entonces $\hat{\beta}_{MCG}$ es insesgado; es decir: $E\left(\hat{\beta}_{MCG}\right) = \beta$.

4. La matriz de varianzas-covarianzas de $\hat{\beta}_{MCG}$ viene dada por:

   $$Var - Cov\left(\hat{\beta}_{MCG}\right) = \sigma^{2} \cdot \left(X^{*\prime} X^{*}\right)^{-1},$$

   o de manera alternativa, teniendo en cuenta la transformación lineal $X^{*} = V^{-1} X$, se puede expresar en función de las variables originales que conforman X:

[8] Obsérvese que, en el caso de que se trabajase con los datos originales, habría que calcular la matriz Ω^{-1}, la cual es cuadrada de orden *n* x *n*. En cambio, si se trabaja con los datos de las variables transformadas, el orden de las matrices cuya inversa debería calcularse sería sólo de orden *k* x *k*.

 ISBN: 978-84-694-7251-4

$$Var-Cov\left(\hat{\beta}_{MCG}\right)=\sigma^2\cdot\left(X'\Omega^{-1}X\right)^{-1}.$$

Como conclusión de las tres propiedades anteriores, tenemos que:

$$\boxed{\hat{\beta}_{MCG}\rightarrow N_k\left(\beta;\sigma^2\cdot\left(X^{*\prime}X^{*}\right)^{-1}\right)},\ \text{o bien:}\ \boxed{\hat{\beta}_{MCG}\rightarrow N_k\left(\beta;\sigma^2\cdot\left(X'\Omega^{-1}X\right)^{-1}\right)}.$$

5. El estimador $\hat{\beta}_{MCG}$ es ELIO de β.

   Puesto que el modelo transformado satisface las condiciones supuestas de un modelo clásico de regresión lineal, de acuerdo con el Teorema de Gauss-Markov, el estimador MCO de los coeficientes de regresión de dicho modelo transformado (esto es, el estimador MCG), resulta entonces ELIO.

   En particular, dado que $\hat{\beta}_{MCG}$ es óptimo, tenemos que:

   $$Var-Cov\left(\hat{\beta}_{MCG}\right)\leq Var-Cov\left(\hat{\beta}_{MCO}\right),$$

   pues: $\sigma^2\cdot\left(X'X\right)^{-1}X'\Omega X\left(X'X\right)^{-1}-\sigma^2\cdot\left(X'\Omega^{-1}X\right)^{-1}$ es una matriz semidefinida positiva.

- ***Propiedades del estimador MCG del parámetro*** $\boldsymbol{\sigma^2}$

1. La estimación por MCG del otro parámetro relevante del modelo, σ^2, se obtiene a partir de la expresión:

   $$\boxed{\hat{\sigma}^2_{MCG}=\hat{\sigma}^2_{u^*}=\frac{e^{*\prime}e^{*}}{n-k}=\frac{SCR^{*}}{n-k}},$$

   o en función de las variables del modelo original:

   $$\boxed{\hat{\sigma}^2_{MCG}=\frac{e'_{MCG}\,\Omega^{-1}\,e_{MCG}}{n-k}},$$

   donde:

   $$e^{*}=Y^{*}-\hat{Y}^{*}_{MCG}=Y^{*}-X^{*}\hat{\beta}_{MCG}=V^{-1}Y-(V^{-1}X)\hat{\beta}_{MCG}=V^{-1}\left(Y-X\hat{\beta}_{MCG}\right)=V^{-1}e_{MCG}.$$

2. Este estimador es insesgado; es decir: $E\left[\hat{\sigma}^2_{MCG}\right]=\sigma^2$.

- ***Coeficiente de determinación***

Una dificultad añadida que surge en el contexto de un modelo cuya perturbación aleatoria no es esférica se refiere a la utilización del coeficiente de determinación R^2 del modelo transformado como medida de bondad del ajuste. En primer lugar, dicho modelo transformado puede no tener término independiente, con lo que R^2 ya no estaría acotado entre 0 y 1. Y, en segundo lugar, tendremos que conformarnos con

medir la capacidad del modelo para explicar la variable transformada Y^*, que, sin embargo, no olvidemos que no es nuestra variable de interés, pues ésta es Y.

- ***Inferencia estadística***

Como hemos podido apreciar, el nuevo modelo transformado, obtenido tras pre-multiplicar las observaciones de las variables originales por la matriz de coeficientes lineales apropiada, no sólo tiene los mismos coeficientes de regresión que el modelo original, sino que también cumple los supuestos propios de la modelización econométrica clásica, para la cual se ha desarrollado en temas anteriores toda la teoría inferencial referida a dichos coeficientes. Por consiguiente, todos los estadísticos entonces establecidos podrán seguir siendo válidos, con la única salvedad de que en lugar de referirnos a las variables originales X e Y, deberemos hacerlo ahora a las variables transformadas X^* e Y^*, respectivamente. O de forma alternativa, si se desease seguir trabajando con las variables originales, siempre que aparezca un producto entre las matrices de datos de dichas variables, deberá considerarse en medio de ellas la matriz Ω^{-1}, de manera análoga a lo que ya se ha mostrado en expresiones anteriores de este mismo Tema.

4.5. Detección y tratamiento de la heteroscedasticidad con *EViews*.-

Como ya hemos visto, si nos centramos en el problema de la heteroscedasticidad, nuestro modelo de regresión lineal, $Y = X\beta + u$, se caracteriza porque la matriz de varianzas-covarianzas de la perturbación aleatoria *u* adopta la forma:

$$Var - Cov(u) = \sigma^2 \cdot \begin{pmatrix} \sigma_{u_1}^2 / \sigma^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_{u_2}^2 / \sigma^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_{u_3}^2 / \sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma_{u_n}^2 / \sigma^2 \end{pmatrix}.$$

Es decir, en general: $\sigma_{u_i}^2 \neq \sigma_{u_j}^2$ y $\sigma_{u_i u_j} = 0, \ \forall i \neq j$.

La detección de la heteroscedasticidad constituye una labor ardua, ya que, como sabemos, la perturbación aleatoria del modelo no es una variable directamente observable. No obstante, por el tipo de datos considerado o de análisis realizado, se pueden tener ciertas sospechas en relación a este problema; así, las series de datos de corte transversal son más proclives a presentar esta circunstancia que las series de datos temporales.

Existen distintos tipos de métodos para detectar la heteroscedasticidad en un modelo y algunos de ellos no sólo sirven para este fin, sino que también constituyen la base para ayudarnos a establecer la matriz V^{-1} que debe usarse posteriormente para transformar el modelo original y estimar luego éste por el método de MCG.

Los diferentes métodos se podrían clasificar, básicamente, en:

- Métodos gráficos
- Métodos analíticos
  - Contrastes paramétricos (basados en las hipótesis estadísticas establecidas): Park, Glesjer, White…
  - Contrastes no paramétricos (no basados en las hipótesis estadísticas establecidas; en particular, no tienen en cuenta el supuesto de normalidad de u_i): picos, Spearman…

A continuación, y con ayuda de *EViews*, analizaremos un modelo que nos servirá de ejemplo para ir viendo la aplicación de algunos de estos métodos. Una vez que evidenciemos con ellos la existencia de heteroscedasticidad, procederemos finalmente a estimar dicho modelo por MCG.

El Ejercicio que vamos a plantear es el nº 41 del Boletín de este Tema, donde se refiere que se quiere estimar la relación existente entre el Valor Añadido Bruto (VAB) y el empleo en las Comunidades Autónomas españolas y las Ciudades Autónomas de Ceuta y Melilla, que consideraremos como un único ente territorial. Con el propósito indicado, se utilizan los datos del VAB a coste de factores y el número medio de ocupados, para el año 1991. El fichero ***het.wf1*** recoge esta información, definiéndose las variables:

- **VAB**: Valor Añadido Bruto a coste de factores del año 1991 (en millones de unidades monetarias)
- **EMPLEO**: Número medio de ocupados en 1991 (en miles de personas)

Este fichero se encuentra en el espacio reservado a la Asignatura en la plataforma de docencia virtual *WebCT*. Tras descargarlo en el *Escritorio* de nuestro PC, procederemos a abrirlo con *EViews* en *FILE / OPEN / WORKFILE*.

Lo primero que haremos será estimar por MCO nuestro modelo:

$$\boxed{VAB = \beta_1 + \beta_2 EMPLEO + u}.$$

Como ya es bien sabido, para ello seleccionaremos: *QUICK / ESTIMATE EQUATION*. En el cuadro de diálogo resultante escribiremos entonces: VAB C EMPLEO, aceptando luego (*OK*) y obteniendo la *Figura 15*.

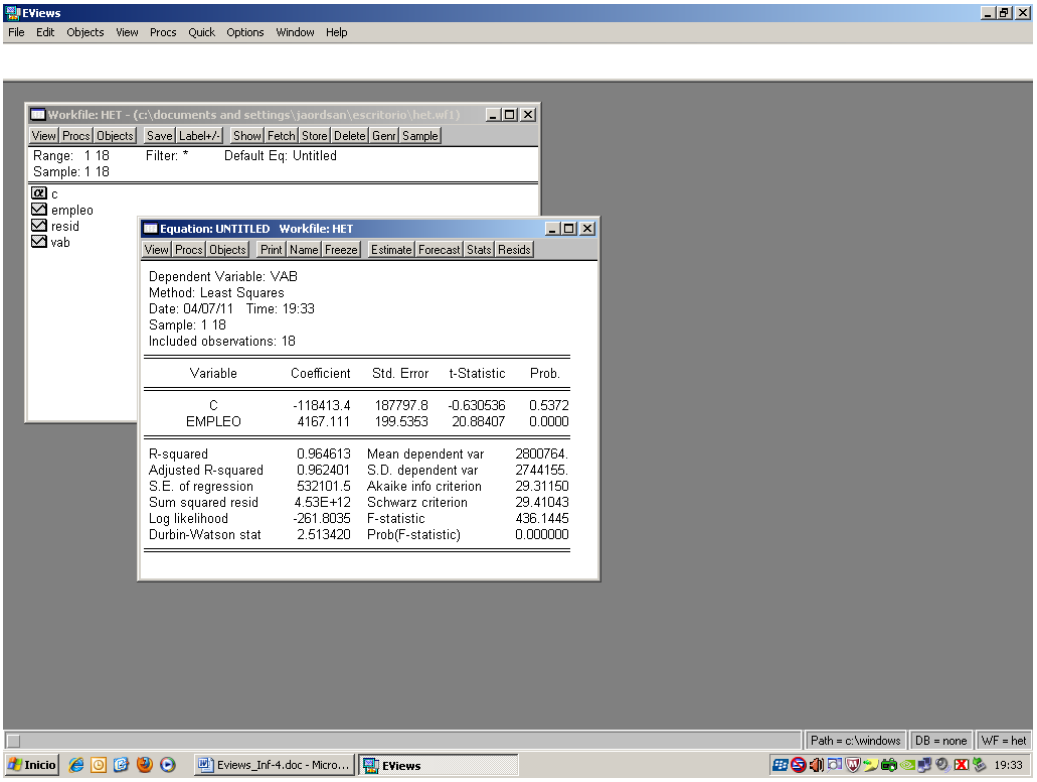


Figura 15

A continuación, seleccionamos la opción *NAME* de la *Ventana de Ecuación* para guardar dicha estimación con el nombre, por ejemplo, de AJUSTEMCO, cerrando seguidamente dicha ventana.

La observación de los parámetros, coeficientes y estadísticos conocidos obtenidos podría hacernos pensar inicialmente que el modelo resulta aceptable; sin embargo por la naturaleza de los datos, de tipo transversal, sospechamos que puede presentar problemas de heteroscedasticidad. En particular, pensamos que el comportamiento de la varianza de la perturbación aleatoria depende directamente de la variable explicativa del modelo, esto es, del EMPLEO, o bien de una transformación de ésta. Por este motivo, vamos a estudiar por distintos métodos la existencia de este posible problema.

- ***Métodos gráficos de detección***

El primer tipo de métodos que pueden utilizarse para estudiar la posible presencia de heteroscedasticidad en un modelo consiste en la realización de determinados gráficos.

En concreto, se trata de representar, de forma teórica, la varianza de la perturbación aleatoria en función de alguna variable explicativa X_j, $j = 2,\ldots,k$ del modelo (o bien de varias de ellas o, incluso, de la variable explicada): $\sigma_{u_i}^2 = f\left(X_{ji}\right), \quad \forall i = 1,2,\ldots,n$.

Sin embargo, dado que la variable aleatoria *u* no es observable (y por tanto, tampoco su varianza), una opción es tomar los cuadrados de los residuos (e_i^2) como aproximación de la varianza de u_i; es decir, plantear[9]:

$$\boxed{e_i^2 = f\left(X_{ji}\right), \quad \forall i = 1,2,\ldots,n}.$$

Para realizar estos gráficos, habría que definir primero la serie de los residuos al cuadrado.

La serie de residuos del modelo es calculada de forma automática cuando éste se estima. Sus valores se hallan en *resid*. No obstante, hay que tener presente que *resid* es un objeto donde se van guardando los valores de los residuos de la última estimación que se lleve a cabo. Dado que vamos a trabajar con la serie concreta de residuos MCO recién creada, deberemos crear ésta como una variable específica a partir de lo que hay en este instante almacenado en *resid*. Para hacer esto, seleccionaremos *GENR* en la Ventana de Trabajo y escribiremos en el cuadro de diálogo que surge (*Enter equation*): RS = RESID, según se muestra en la *Figura 16*. Tras ello, aceptaremos pulsando *OK*.

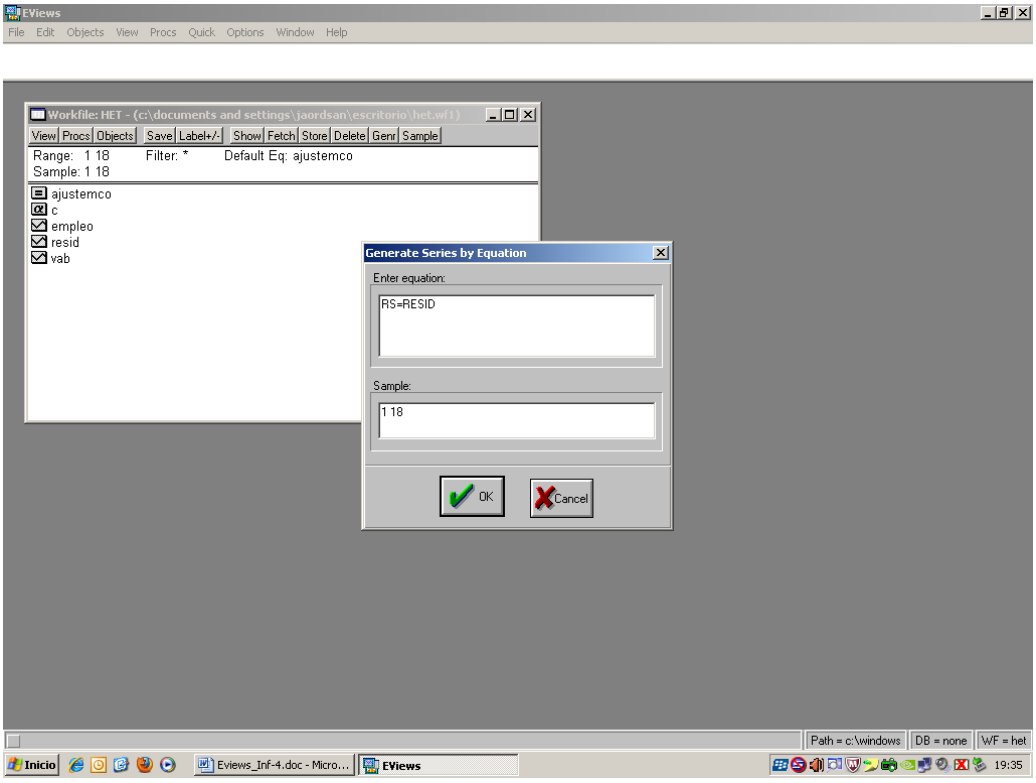


Figura 16

Una vez hecho esto, puesto que nuestra intención es trabajar ahora con los residuos al cuadrado, seguidamente generaremos dicha serie a partir de RS. Esto es, nuevamente elegiremos *GENR* y en el correspondiente cuadro de diálogo que se abre, escribiremos esta vez: RS2 = RS^2. Después pulsaremos *OK*.

[9] Junta a esta opción, existen otras posibilidades consistentes en plantear funciones para los residuos directamente, o bien, para sus valores absolutos.

A partir de aquí, podemos proceder a realizar los gráficos entre los cuadrados de los residuos y una función de la variable explicativa que consideramos que puede ser la principal generadora del problema de la heteroscedasticidad en el modelo.

En el presente ejercicio, la selección de la variable explicativa que puede causar la heteroscedasticidad no presenta problemas, pues sólo estamos considerando una: el EMPLEO. Sin embargo, si tuviésemos más, deberíamos hacer previamente esta selección, bien gracias al conocimiento económico de las variables en cuestión que estuviésemos considerando, o bien a través de la aplicación de este método a todas ellas.

La siguiente cuestión que se plantearía sería la selección de la función de X_j que habría que tomar; es decir, ¿la variabilidad de la perturbación aleatoria sigue el patrón de comportamiento de la variable X_j de forma directa, de forma inversa, de su cuadrado…? En este caso, vamos a representar gráficamente el cuadrado de los residuos únicamente en función del EMPLEO. Pero, de forma análoga, se podría hacer con otras formas funcionales: su inversa, cuadrado, etc.

Para obtener dicho gráfico, debemos elegir en la *barra principal de menús*: *QUICK / GRAPH*. Se creará de este modo una nueva ventana donde escribiremos en primer lugar la variable independiente (a representar en el eje horizontal) y luego la dependiente (a representar en el eje vertical): EMPLEO y RS2, respectivamente. Después de aceptar (*OK*), en el nuevo cuadro de diálogo que se abre elegiremos *Scatter Diagram* como tipo de gráfico. Para concluir, aceptaremos (*OK*). La *Figura 17* muestra el gráfico indicado.

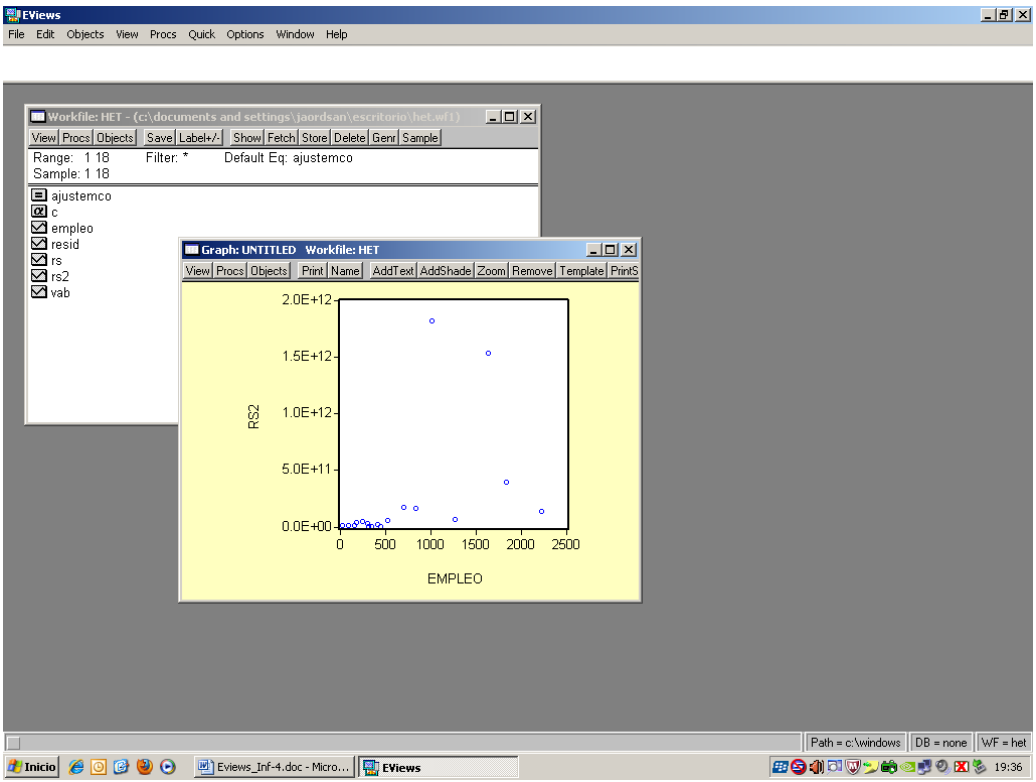


Figura 17

 ISBN: 978-84-694-7251-4

A la vista de ello, se puede comprobar cómo la variabilidad del cuadrado de los residuos es mayor conforme crecen las cifras de EMPLEO, no siendo por lo tanto constante. Así pues, parece evidente que existe una relación directa entre la varianza de los residuos (y, por consiguiente, de la perturbación aleatoria) y la variable EMPLEO, lo que parece apuntar a la existencia de heteroscedasticidad en el modelo.

Podemos guardar este gráfico llamándolo, por ejemplo, METGRAF al pulsar *NAME*.

- ***Contrastes paramétricos***

Junto con los métodos gráficos (más intuitivos que precisos), se han desarrollado un buen número de estadísticos para contrastar la hipótesis nula de igualdad de varianza u homoscedasticidad de la perturbación aleatoria correspondiente a cada observación de la muestra estudiada. Esta gran variedad se debe a que la especificación de la hipótesis alternativa de heteroscedasticidad no suele ser conocida y puede ser más o menos general. A continuación, vamos a revisar algunos de estos contrastes. En concreto, nos vamos a centrar en tres contrastes de tipo paramétrico: Park, Glesjer y White, que se caracterizan por estar basados en las hipótesis y supuestos estadísticos establecidos en el modelo.

El **contraste de Park** parte del establecimiento de una relación funcional entre los valores de la varianza de la perturbación aleatoria correspondiente a las distintas observaciones, $\sigma_{u_i}^2$, y los de la variable explicativa X_{ji}, para algún $j = 2,\ldots,k$, del tipo:

$$\sigma_{u_i}^2 = \sigma^2 \cdot X_{ji}^{\beta} \cdot e^{v_i} \qquad \forall i = 1,\ldots,n,$$

o de forma equivalente:

$$\ln \sigma_{u_i}^2 = \ln \sigma^2 + \beta \ln X_{ji} + v_i \qquad \forall i = 1,\ldots,n.$$

Dado que $\sigma_{u_i}^2$ se desconoce, Park propone utilizar como aproximación los residuos al cuadrado: e_i^2. De esta forma, finalmente considera la expresión:

$$\boxed{\ln e_i^{\,2} = \alpha + \beta \ln X_{ji} + v_i \qquad \forall i = 1,\ldots,n},$$

donde $\alpha = \ln \sigma^2$ es, simplemente, una constante.

El contraste o prueba de Park consiste, en definitiva, en realizar el ajuste MCO del logaritmo neperiano del cuadrado de los residuos respecto al logaritmo neperiano de la variable explicativa que se considere que puede ocasionar el problema de heteroscedasticidad (en el caso de nuestro problema, sólo puede ser el EMPLEO) y verificar si dicha relación es significativa, mediante la hipótesis nula:

$$\boxed{\begin{array}{l} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{array}}$$

 ISBN: 978-84-694-7251-4

Este contraste se puede realizar con el estadístico *t*-Student habitual de significatividad individual del parámetro β, o bien con el estadístico *F* de significatividad global del modelo, ya que en este caso, por ser un modelo de regresión lineal simple, la interpretación del resultado es idéntica.

El rechazo de la hipótesis nula supone admitir la existencia de heteroscedasticidad.

Sin embargo, la prueba de Park presenta de forma general algunas limitaciones:

a) Se adopta e_i^2 como una aproximación de $\sigma_{u_i}^2$.

b) Es preciso seleccionar la variable X_j posible causante de la heteroscedasticidad del modelo.

c) La perturbación aleatoria de la regresión auxiliar considerada, v_i, puede presentar problemas de heteroscedasticidad, los cuales son obviados.

Estas limitaciones nos llevan a que tengamos presente que si en el contraste se aceptase la hipótesis nula, ello no tendría por qué significar necesariamente que el modelo resulta homoscedástico; simplemente se estaría poniendo de manifiesto que la relación planteada no es significativa, pero sin embargo podría haber otras que sí lo fuesen, por lo que finalmente el modelo podría presentar heteroscedasticidad.

Para llevar a cabo este contraste, deberemos estimar por MCO la regresión auxiliar propuesta por Park y comprobar la significatividad de este ajuste. En el presente ejercicio sería la regresión lineal entre el logaritmo neperiano del cuadrado de los residuos (RS2) y el logaritmo neperiano de la variable EMPLEO (que es la única variable explicativa que tenemos, por lo que no tenemos dudas sobre su elección).

La realización de este contraste con ayuda de *EViews* comenzará con la estimación del modelo indicado según el procedimiento habitual: seleccionando *QUICK / ESTIMATE EQUATION* en la *barra principal de menús* y seguidamente, tal y como se muestra en la *Figura 18*, escribiendo en la Ventana de Especificación de la Ecuación (*Equation Specification*):

LOG(RS2) C LOG(EMPLEO).

Tras pulsar *OK*, se obtiene la pantalla del resultado de la estimación mostrada en la *Figura 19*, donde se puede comprobar cómo la regresión auxiliar planteada es significativa para un nivel de confianza máximo del 96,42%. Así pues, para un nivel de significación del 5%, la prueba de Park sugiere la presencia de heteroscedasticidad en nuestro modelo.

Para conservar el ajuste llevado a cabo, podemos seleccionar la opción *NAME* y darle un nombre a esta ecuación; por ejemplo, PARK.

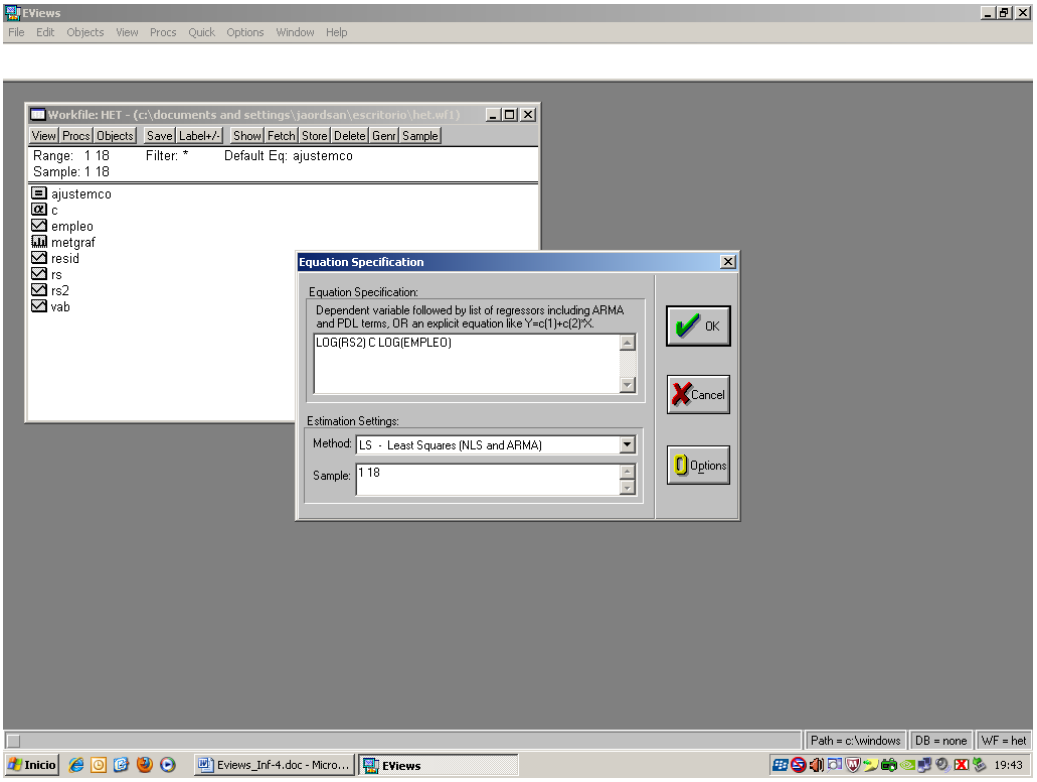


Figura 18

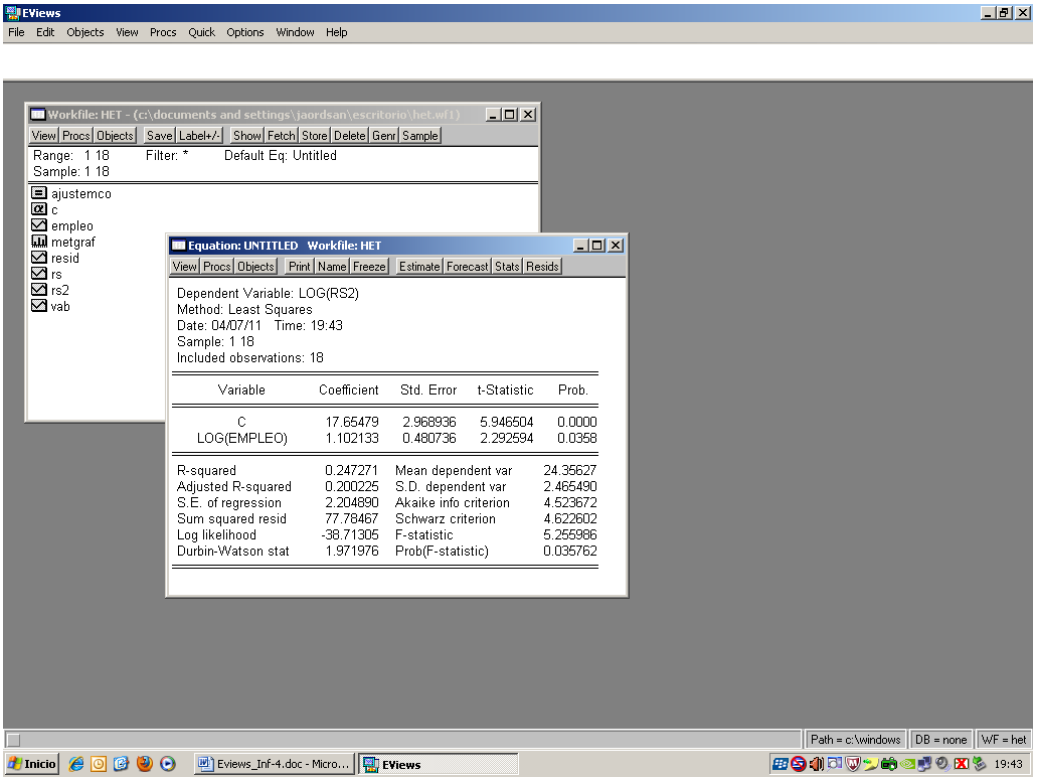


Figura 19

El **contraste de Glesjer** constituye un test de detección de la heteroscedasticidad similar en concepción al de Park. En concreto, este contraste se basa en la regresión del

valor absoluto de los residuos mínimo-cuadráticos, frente a distintas especificaciones de la variable X_j, $j = 2,\ldots,k$, que supuestamente puede estar creando los problemas de heteroscedasticidad en el modelo; esto es:

$$\boxed{|e_i| = \alpha + \beta X_{ji}^h + v_i \qquad \forall i = 1,...,n \ , \qquad \text{con } h = \{1, -1, 1/2, -1/2\}}.$$

En cada una de estas cuatro regresiones distintas, se trata de contrastar la significatividad del parámetro β. Es decir, este contraste también toma como hipótesis nula el hecho de que la perturbación sea homoscedástica, especificando, para el caso de que hubiese heteroscedasticidad, varios esquemas alternativos diferentes en un intento de determinar cuál sería la pauta de comportamiento seguida por ésta.

Al igual que ocurría con la prueba de Park, el contraste de Glesjer presenta ciertas limitaciones derivadas de su propia definición. Básicamente son:

a) Se adopta $|e_i|$ como una aproximación de σ_{u_i}.

b) Es preciso seleccionar la variable X_j posible causante de la heteroscedasticidad del modelo.

c) La perturbación aleatoria de la regresión auxiliar considerada, v_i, puede presentar problemas de heteroscedasticidad, los cuales son ignorados.

Si en alguna de las regresiones auxiliares planteadas rechazamos la hipótesis nula, eso significará que existe heteroscedasticidad en el modelo y además esto nos permite conocer su estructura, es decir, el supuesto de comportamiento que debemos tomar sobre la varianza de las perturbaciones. Por el contrario, si en ninguna de las regresiones planteadas el parámetro resulta significativo, las limitaciones anteriormente expuestas no nos permitirán asegurar que el modelo es homoscedástico, sino simplemente que los patrones de una posible existencia de heteroscedasticidad no son los planteados, pero sin embargo sí podrían ser otros.

Si en más de uno de los ajustes estimados el regresor es significativo, a la hora de elegir la estructura de heteroscedasticidad más apropiada, deberemos quedarnos con aquél que sea más significativo.

Cabe resaltar que, al utilizar el valor absoluto de los residuos como variable endógena de la regresión auxiliar de Glesjer, estamos utilizando una aproximación de la desviación típica de la perturbación y, por tanto, a la hora de determinar la estructura de su varianza tendremos que considerar el cuadrado de la relación detectada.

Volviendo al ejemplo sobre el que estamos trabajando, deberemos por tanto plantear cuatro regresiones distintas, donde en todas ellas la variable dependiente será el valor absoluto del residuo MCO y, en cuanto, a la variable independiente ésta será: el EMPLEO, la inversa del EMPLEO, la raíz cuadrada del EMPLEO y la inversa de la raíz cuadrada del EMPLEO, respectivamente en cada caso.

Para hacerlo con *EViews*, deberemos llevar a cabo los cuatro ajustes indicados y fijarnos en la significatividad estadística de cada uno de ellos. Así pues, deberemos seleccionar *QUICK / ESTIMATE EQUATION* y escribir en la Ventana de Especificación de la Ecuación[10]:

- El primero de los ajustes: ABS(RS) C EMPLEO
- El segundo: ABS(RS) C 1/EMPLEO
- El tercero: ABS(RS) C SQR(EMPLEO)
- Y finalmente el cuarto: ABS(RS) C 1/SQR(EMPLEO)

Los resultados de cada una de las cuatro estimaciones pueden verse en las *Figuras 20*, *21*, *22* y *23*, respectivamente.

Cabe decir que para llevar a cabo cada una de las especificaciones del modelo, no es necesario realizar cada vez el proceso *QUICK / ESTIMATE EQUATION*, sino que una vez estimado el primer ajuste, basta con seleccionar la opción *ESTIMATE* de la *Ventana de Ecuación* e ir cambiando la variable explicativa en cuestión.

Como se ha indicado, en cada caso hay que realizar un contraste de significatividad individual del coeficiente asociado a la variable explicativa en cuestión (o de significatividad global del modelo, ya que se trata de modelos de regresión simple).

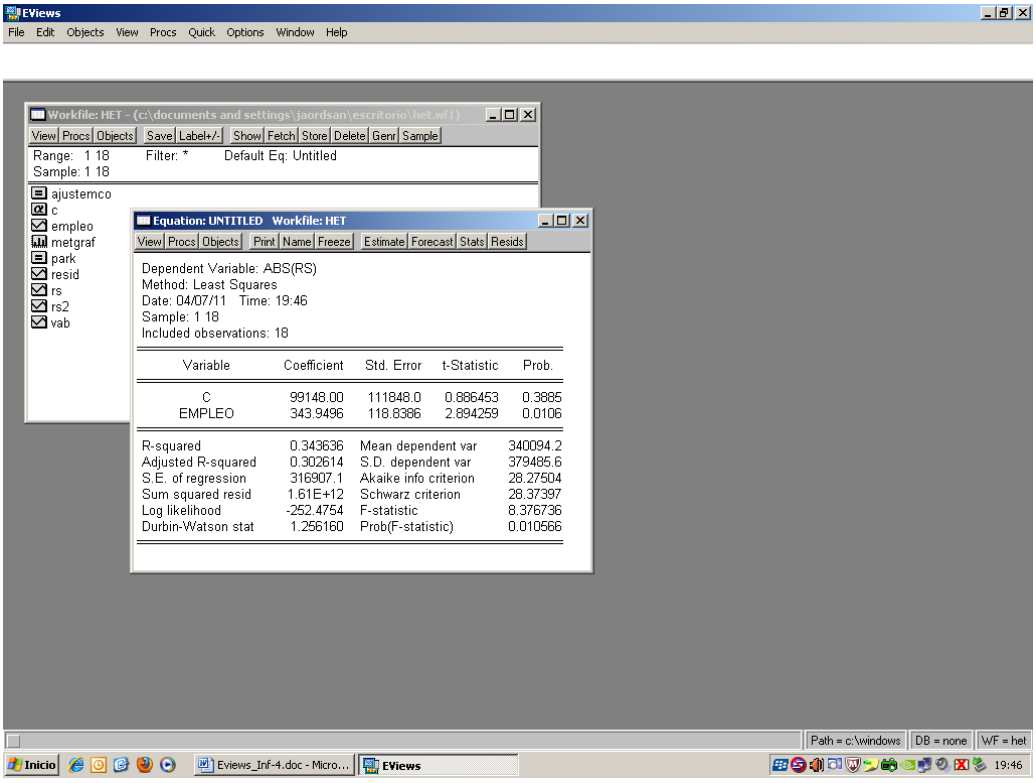


Figura 20

[10] Ha de tenerse en cuenta que la función "valor absoluto" de una variable se escribe en *EViews* de la forma: ABS(*nombre de la variable*). Por su parte, la "raíz cuadrada" se escribe: SQR(*nombre de la variable*).

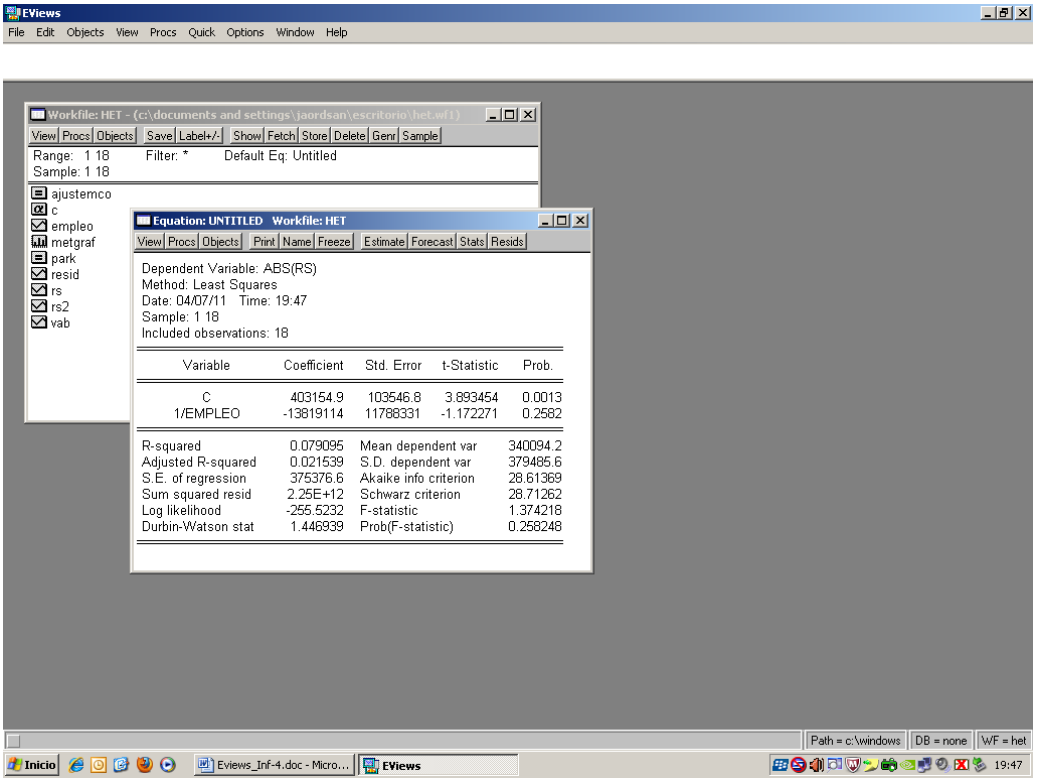


Figura 21

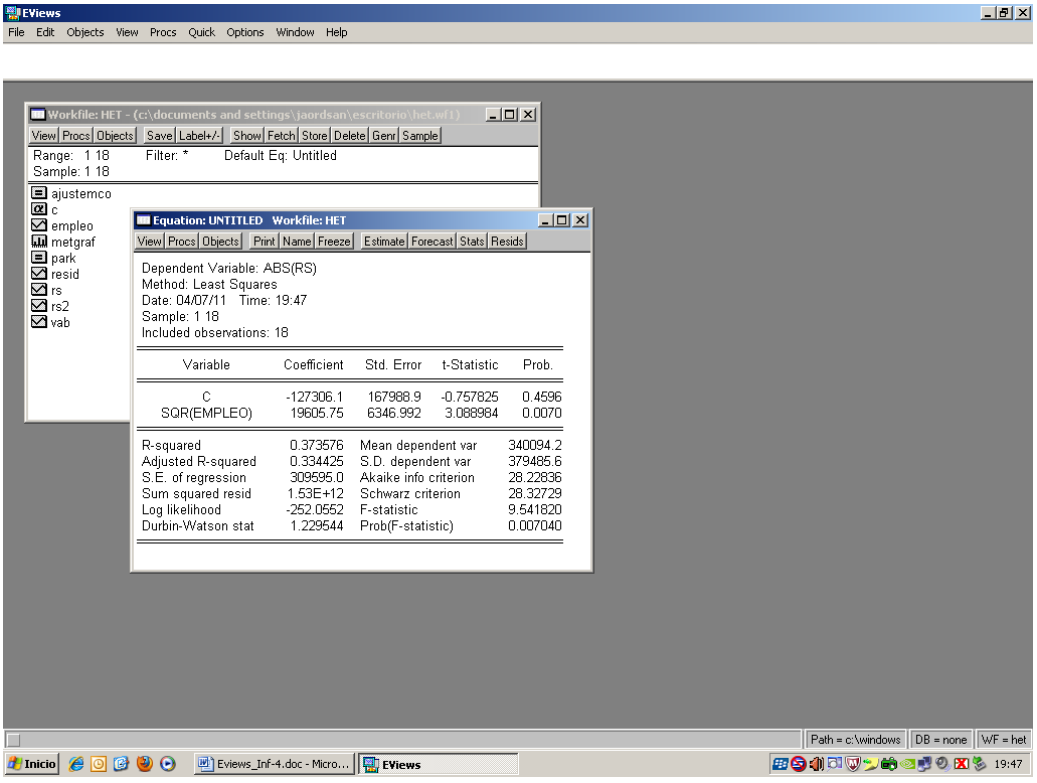


Figura 22

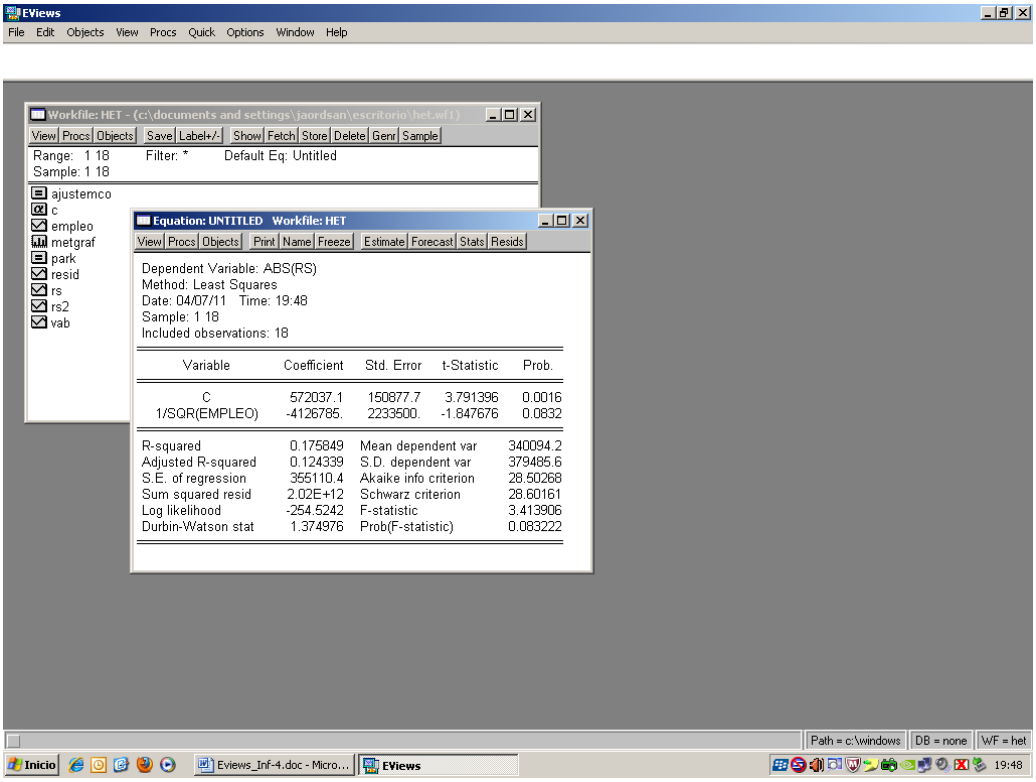


Figura 23

A la vista de los resultados obtenidos, se puede observar que, para un nivel de significación $\alpha = 5\%$, en las regresiones primera (respecto a EMPLEO) y, en mayor medida, tercera (respecto a su raíz, SQR(EMPLEO)), los estadísticos *t*–Student se sitúan en la región crítica, y, por tanto, se considera que las correspondientes variables explican el comportamiento de los residuos; es decir, se puede concluir que la varianza de la perturbación aleatoria no permanece constante a lo largo de la muestra. Y, además, vemos que éstas son las pautas de comportamiento que la heteroscedasticidad puede seguir. Así pues, vamos a guardar la regresión más significativa, la relativa a la raíz cuadrada del EMPLEO: SQR(EMPLEO), dándole en *NAME* el nombre de GLESJER.

El **contraste de White** es un contraste paramétrico más general y robusto, ya que no precisa de la elección inicial de una variable concreta del modelo de la que dependa la heteroscedasticidad bajo la hipótesis alternativa.

Este contraste se basa en la regresión de los cuadrados de los errores MCO, que se toman como aproximación de las varianzas de las perturbaciones, en función de: el término independiente, las variables independientes del modelo, los cuadrados de éstas y, de forma optativa, sus productos cruzados dos a dos. De acuerdo con esto último, *EViews* incorpora dos versiones de este contraste: una en la que se incluyen en la regresión los productos cruzados dos a dos y otra en la que éstos no se incorporan. En ambos casos, el estadístico del test de White es el producto del número de observaciones por el coeficiente de determinación de la regresión propuesta, el cual se distribuye asintóticamente, bajo la hipótesis nula, como una χ^2 con *m* grados de libertad

(donde *m* es el número de variables explicativas de esta regresión, sin contar la ordenada en el origen):

$$\boxed{\chi^2_W = n \cdot R^2 \rightarrow \chi^2_m}.$$

La interpretación del contraste reside en que si las perturbaciones fueran homoscedásticas, las variables incluidas en la regresión auxiliar no deberían tener ningún poder explicativo sobre los residuos al cuadrado y, por tanto, el valor del coeficiente de determinación debería ser muy pequeño y con ello el valor del estadístico. Por esta razón, en el caso contrario, si el valor muestral del estadístico es suficientemente alto como para que la probabilidad de rechazar la hipótesis nula, siendo cierta, sea menor que el nivel de significación que nos fijemos (por ejemplo, el 5%), rechazaremos la hipótesis nula y admitiremos la existencia de heteroscedasticidad.

Este test es el único de los contrates considerados que viene programado en *EViews*. Para realizarlo, abriremos la ecuación AJUSTEMCO haciendo doble "clic" sobre ella en el *Directorio de Objetos*. Una vez aquí, se sigue la secuencia *VIEW / RESIDUAL TESTS*. Se abrirá entonces un submenú donde se nos presentará la opción de realizar el contraste de White con o sin términos cruzados. Vamos a elegir la opción "con términos cruzados": *WHITE HETEROSKEDASTICITY (CROSS TERMS)*, como se puede ver en la *Figura 24* (aunque en nuestro ejemplo, de cualquier forma, los resultados del contraste de White van a ser iguales con o sin términos cruzados, pues el modelo sólo tiene una variable explicativa distinta del término independiente).

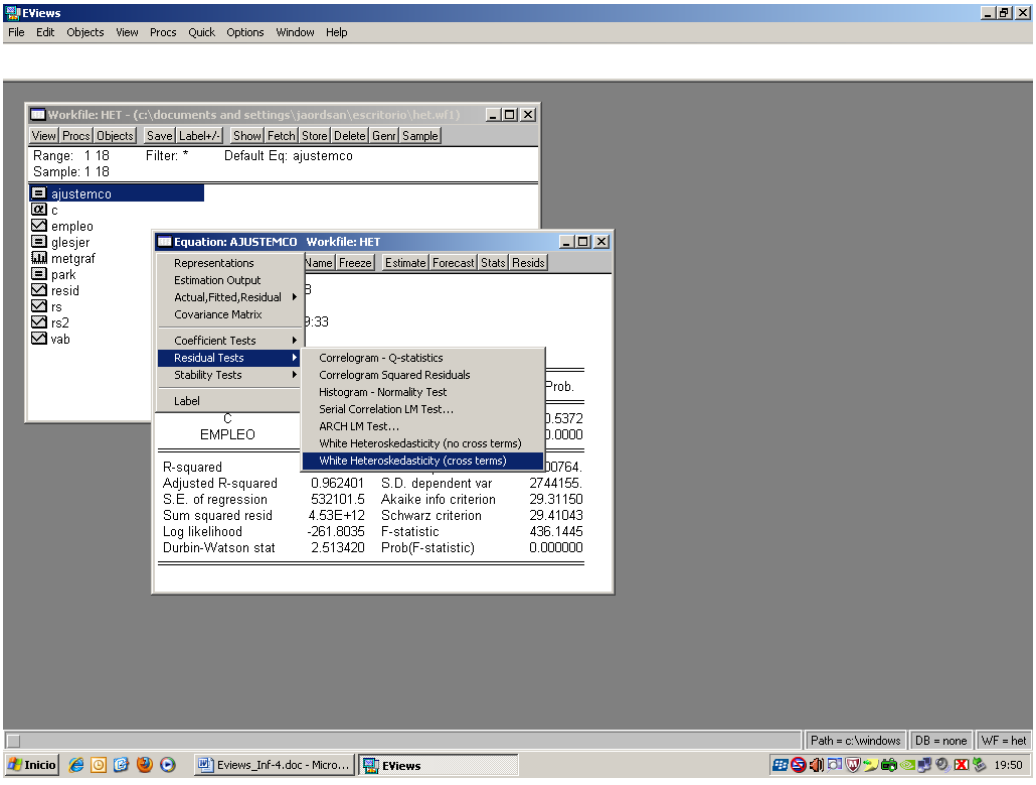


Figura 24

 ISBN: 978-84-694-7251-4

El resultado del contraste, así como la regresión auxiliar estimada, se muestran en la *Figura 25*.

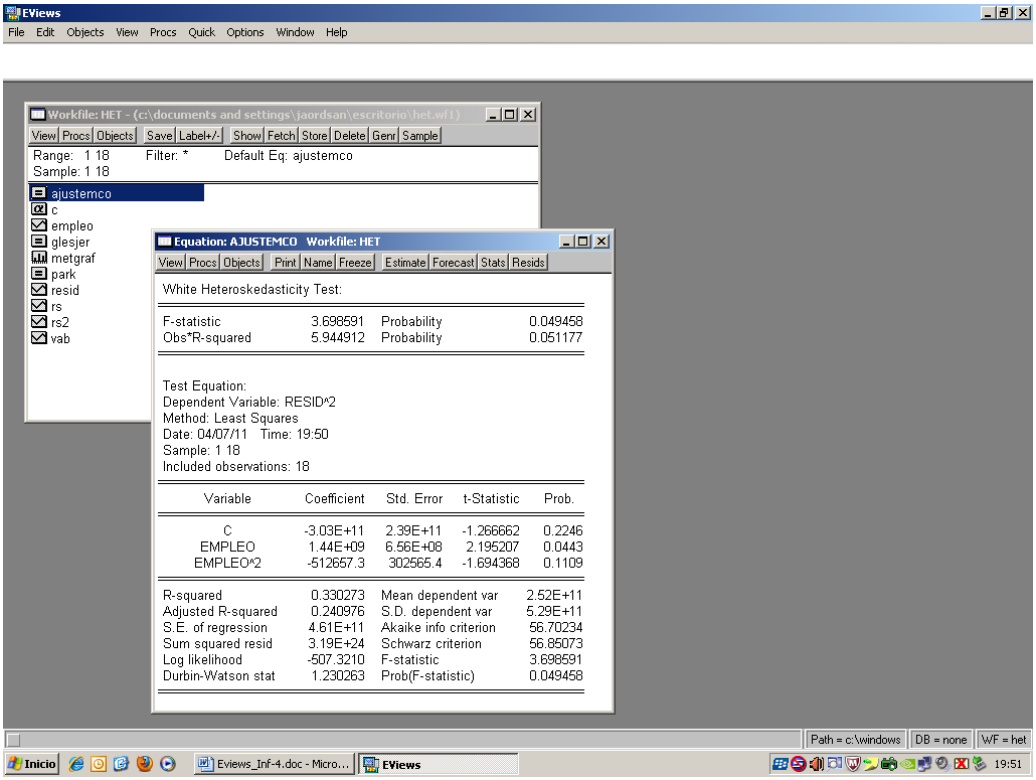


Figura 25

A la vista del *p-valor* (0,051177) que aparece asociado al estadístico de prueba de White (*Obs*R-Squared*), podemos señalar que para un nivel de significación estrictamente del 5% no se podría rechazar la hipótesis nula de homoscedasticidad. No obstante, este valor es muy próximo al 5%, con lo que dado el reducido tamaño de la muestra y el carácter asintótico de este contraste, así como por lo apuntado por todas las pruebas anteriores realizadas (gráficos y contrastes paramétricos), como conclusión final lo más prudente es asumir que pueden existir problemas de heteroscedasticidad en nuestro modelo. Por tanto, lo más apropiado es proceder a su estimación por el método de MCG, que, como bien sabemos, proporciona estimadores lineales insesgados y óptimos (ELIO) en estos casos.[11]

- ***Solución a la heteroscedasticidad: estimación del modelo por el método de MCG***

El método de estimación de los *mínimos cuadrados generalizados (MCG)* aparece como el idóneo para tratar los problemas de heteroscedasticidad, ya que hace que los estimadores del modelo resulten ELIO.

[11] En caso de duda, siempre será preferible optar por pensar que hay un problema de heteroscedasticidad en el modelo y proceder a su estimación por el método de MCG, cuyo estimador será ELIO. Nótese que, si finalmente el modelo fuese homoscedástico el estimador MCG coincidirá con el obtenido por MCO. En caso contrario, habremos cometido un error, pues el estimador MCO no sería ELIO.

Como ya sabemos, básicamente el método consiste en transformar el modelo original de una forma determinada y estimar este modelo transformado por MCO.

La existencia de heteroscedasticidad supone que la matriz de varianzas-covarianzas de la perturbación aleatoria adopta la forma general:

$$Var-Cov(u)=\begin{pmatrix}\sigma_{u_1}^2 & 0 & \cdots & \cdots & 0\\ 0 & \sigma_{u_2}^2 & 0 & \cdots & 0\\ \vdots & 0 & \sigma_{u_3}^2 & \cdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \cdots & \sigma_{u_n}^2\end{pmatrix}=\sigma^2\cdot\Omega$$

donde:

$$\Omega=\begin{pmatrix}\sigma_{u_1}^2/\sigma^2 & 0 & \cdots & \cdots & 0\\ 0 & \sigma_{u_2}^2/\sigma^2 & 0 & \cdots & 0\\ \vdots & 0 & \sigma_{u_3}^2/\sigma^2 & \cdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \cdots & \sigma_{u_n}^2/\sigma^2\end{pmatrix}.$$

A partir de la matriz Ω, se obtienen las matrices Ω^{-1} y V^{-1}, donde ésta última es tal que: $\Omega=VV'$; o de forma equivalente: $\Omega^{-1}=\left(V^{-1}\right)'V^{-1}$. Sus expresiones son:

$$\Omega^{-1}=\begin{pmatrix}\sigma^2/\sigma_{u_1}^2 & 0 & \cdots & \cdots & 0\\ 0 & \sigma^2/\sigma_{u_2}^2 & 0 & \cdots & 0\\ \vdots & 0 & \sigma^2/\sigma_{u_3}^2 & \cdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \cdots & \sigma^2/\sigma_{u_n}^2\end{pmatrix},\quad V^{-1}=\begin{pmatrix}\sigma/\sigma_{u_1} & 0 & \cdots & \cdots & 0\\ 0 & \sigma/\sigma_{u_2} & 0 & \cdots & 0\\ \vdots & 0 & \sigma/\sigma_{u_3} & \cdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \cdots & \sigma/\sigma_{u_n}\end{pmatrix}$$

Ya es sabido que, en términos matriciales, partiendo del modelo original $Y=X\beta+u$, la transformación requerida por el método de MCG consiste en pre-multiplicar sus variables por la matriz V^{-1}:

$$V^{-1}Y=\left(V^{-1}X\right)\beta+V^{-1}u\,.$$

A nivel de las observaciones individuales *i*, esto se expresaría de la forma:

$$\left(\frac{\sigma}{\sigma_{u_i}}\right)\cdot Y_i=\beta_1\left(\frac{\sigma}{\sigma_{u_i}}\right)+\beta_2\left(\frac{\sigma}{\sigma_{u_i}}\right)\cdot X_{2i}+\ldots+\beta_j\left(\frac{\sigma}{\sigma_{u_i}}\right)\cdot X_{ji}+\ldots+\beta_k\left(\frac{\sigma}{\sigma_{u_i}}\right)\cdot X_{ki}+\left(\frac{\sigma}{\sigma_{u_i}}\right)\cdot u_i\,,\ \forall i=1,2,\ldots,n$$

Obsérvese que todas las variables del modelo, tanto la explicada, como las explicativas y la perturbación aleatoria, simplemente están multiplicadas por un factor de ponderación, que podríamos denominar ω_i:

$$\boxed{\omega_i = \frac{\sigma}{\sigma_{u_i}}}.$$

La utilización de esta ponderación logra que la perturbación aleatoria resulte homoscedástica y que la aplicación entonces del método de MCO a este modelo transformado, o ponderado, dé lugar a estimadores ELIO. Todo este proceso, en conjunto, no es más que el método de MCG, o también llamado en este caso de MCP (mínimos cuadrados ponderados).[12]

El problema de esta ponderación reside en que lo habitual es que tanto σ como σ_{u_i} resulten desconocidos. Dependiendo del patrón de comportamiento que se considere que sigue la varianza de la perturbación aleatoria (o error estándar en su caso), la ponderación se concretará en una expresión diferente.

A continuación, vamos a plantear de forma sintética los supuestos más habituales que se suelen asumir sobre el comportamiento de la varianza de la perturbación y las expresiones respectivas de las ponderaciones a que dan lugar, para su utilización en el método de MCG:

1. La varianza de la perturbación aleatoria es directamente proporcional al cuadrado de una variable explicativa X_j, $j = 2,\ldots,k$.

   Es decir: $\boxed{\sigma_{u_i}^2 = \sigma^2 \cdot X_{ji}^2, \quad \forall i = 1,2,\ldots,n}$.

   En este caso, $V^{-1} = \begin{pmatrix} 1/X_{j1} & 0 & \cdots & \cdots & 0 \\ 0 & 1/X_{j2} & 0 & \cdots & 0 \\ \vdots & 0 & 1/X_{j3} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1/X_{jn} \end{pmatrix}$; esto es: $\boxed{\omega_i = \frac{1}{X_{ji}}}$.

2. La varianza de la perturbación aleatoria es directamente proporcional a una variable explicativa X_j, $j = 2,\ldots,k$.

   Es decir: $\boxed{\sigma_{u_i}^2 = \sigma^2 \cdot X_{ji}, \quad \forall i = 1,2,\ldots,n}$.

[12] Nótese que si todas las σ_{u_i} fuesen iguales a σ (es decir, si estuviésemos ante homoscedasticidad), todas las ponderaciones adoptarían el mismo valor, siendo éste igual a 1. Así pues, el método de MCO no es más que un caso particular de MCP donde las ponderaciones valen 1.

Aquí, $V^{-1} = \begin{pmatrix} 1/\sqrt{X_{j1}} & 0 & \cdots & \cdots & 0 \\ 0 & 1/\sqrt{X_{j2}} & 0 & \cdots & 0 \\ \vdots & 0 & 1/\sqrt{X_{j3}} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1/\sqrt{X_{jn}} \end{pmatrix}$; y: $\boxed{\omega_i = \frac{1}{\sqrt{X_{ji}}}}$.

3. La varianza de la perturbación aleatoria es directamente proporcional al cuadrado de la variable estimada $\hat{Y}$.

   Es decir: $\boxed{\sigma_{u_i}^2 = \sigma^2 \cdot \hat{Y}_i^2, \quad \forall i = 1,2,\ldots,n}$.

   En este supuesto, $V^{-1} = \begin{pmatrix} 1/\hat{Y}_1 & 0 & \cdots & \cdots & 0 \\ 0 & 1/\hat{Y}_2 & 0 & \cdots & 0 \\ \vdots & 0 & 1/\hat{Y}_3 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1/\hat{Y}_n \end{pmatrix}$; esto es: $\boxed{\omega_i = \frac{1}{\hat{Y}_i}}$.

Lógicamente, la dificultad a la hora de llevar a cabo un ajuste por MCG consiste en tratar de averiguar cuál es el patrón de comportamiento más adecuado que debe adoptarse, para posteriormente fijar la ponderación precisa. Una idea de ello la pueden proporcionar los contrastes de detección de heteroscedasticidad que se deben aplicar previamente.

La última de las soluciones propuestas, la que hace referencia a $\hat{Y}$, precisamente se elige cuando no resulta fácil detectar qué variable explicativa del modelo es la que condiciona la heteroscedasticidad. Tomando $\hat{Y}$, con ello se recoge el efecto de todas las variables explicativas del modelo y, en particular, el de la variable o variables que están generando esta situación.

Volviendo a nuestro ejercicio, una vez efectuados los diversos contrastes para detectar la presencia de heteroscedasticidad en nuestro modelo, podemos suponer, según los resultados que obtuvimos del test de Glesjer, que la varianza de la perturbación aleatoria es proporcional a la variable EMPLEO (recuérdese que según la especificación elegida como más significativa, el valor absoluto del residuo, que se puede considerar como una aproximación al error estándar de *u*, era proporcional a la raíz cuadrada de EMPLEO). Así pues:

$$\sigma_{u_i}^2 = \sigma^2 \cdot EMPLEO_i .$$

Por tanto, se tratará de transformar el modelo original dividiendo cada uno de sus miembros por la raíz cuadrada de esa variable, o lo que es lo mismo ponderándolos por:

$$\omega_i = 1/\sqrt{EMPLEO_i}\,.$$

Para llevar a cabo con *EViews* la estimación por MCG en este caso, en primer lugar haremos doble "clic" sobre la ecuación ajustada por MCO (AJUSTEMCO) y seleccionaremos *ESTIMATE* (*Figura 26*).

Después, en el cuadro de diálogo que aparece (que ya nos resulta familiar), pulsaremos *Options*. Seguidamente, marcaremos la opción *Weighted LS/TSLS (Unavailable with ARMA)* y escribiremos en *Weight* la ponderación concreta que vamos a utilizar[13]: 1/SQR(EMPLEO) tal y como se muestra en la *Figura 27*.

Tras pulsar *OK*, obtendremos finalmente la salida de resultados, que podremos guardar seguidamente con el nombre AJUSTEMCG, que se observa en la *Figura 28*, donde se nos ofrecen los coeficientes estimados por MCG y los valores de los estadísticos más relevantes del modelo transformado obtenido tras aplicar la ponderación correspondiente según el método de MCG (*Weighted Statistics*). Junto a ello, en la parte inferior se muestran los estadísticos derivados del ajuste MCG considerando las variables originales del modelo (no transformadas o ponderadas) (*Unweighted Statistics*).

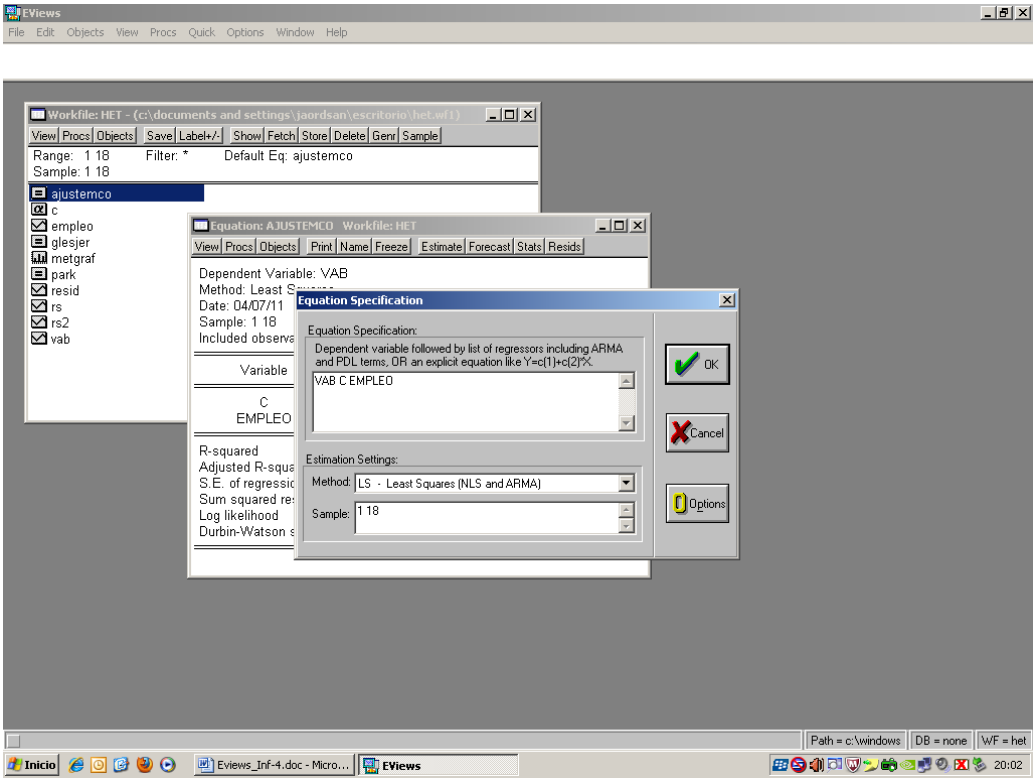


Figura 26

[13] Esta ponderación deberemos escribirla en *EViews* en letras minúsculas.

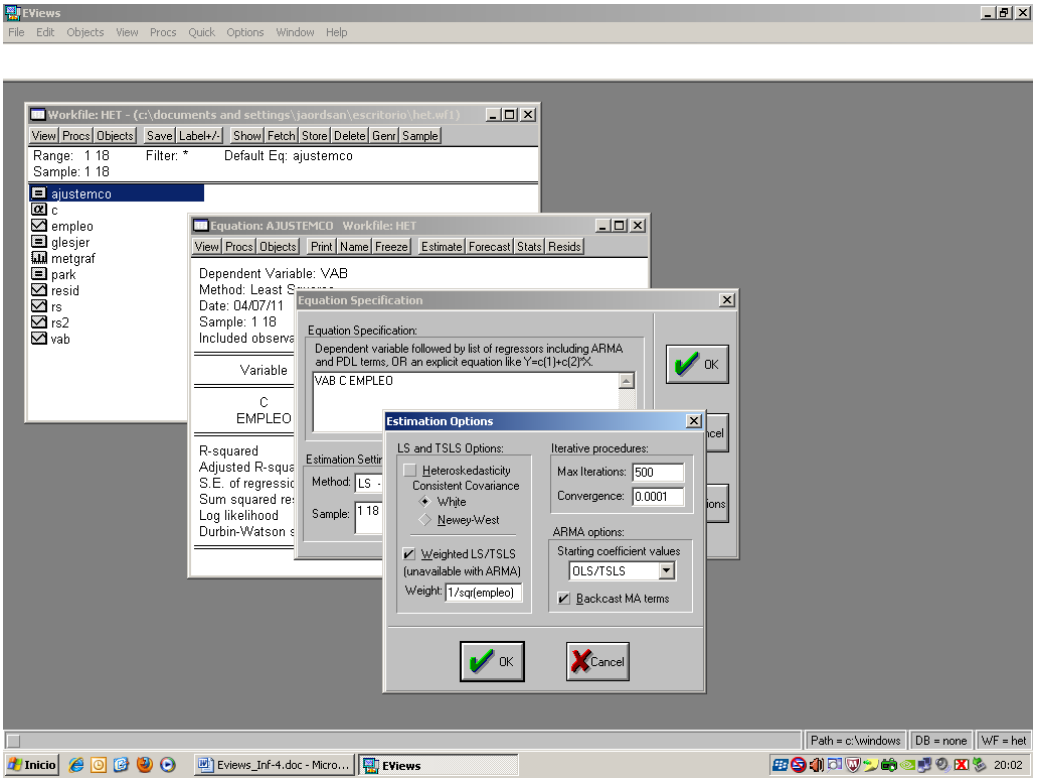


Figura 27

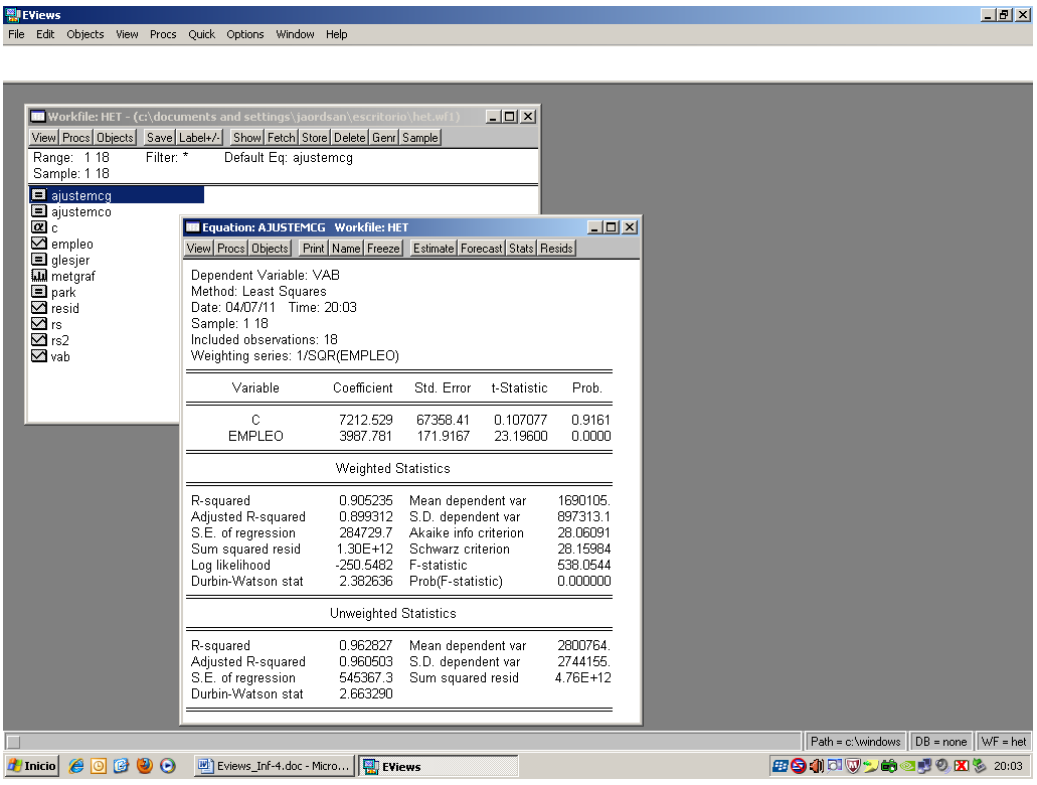


Figura 28

- ***Otra "solución" a la heteroscedasticidad: la estimación consistente de White***

Otra opción que se puede tomar para tratar la presencia de heteroscedasticidad en un modelo econométrico es la *estimación consistente de White*. Esta "solución" no da lugar a resultados óptimos, por cuanto los estimadores obtenidos por este método no son ELIO; sin embargo, esta alternativa a MCG es muy utilizada porque no precisa conocer el patrón de comportamiento que sigue la varianza de la perturbación aleatoria del modelo y, al tiempo, da lugar a resultados más correctos que los obtenidos directamente a través del método de MCO.

La estimación consistente de White halla los estimadores de los parámetros del modelo por MCO, pero adicionalmente calcula la matriz de varianzas-covarianzas de los coeficientes de regresión de manera correcta, esto es, utilizando la expresión:

$$Var - Cov\left(\hat{\beta}_{MCO}\right) = \sigma^2 \cdot \left(X'X\right)^{-1} X'\Omega X\left(X'X\right)^{-1}.$$

Así pues, este método toma en consideración el hecho de que la perturbación aleatoria u resulta no esférica; es decir: $\Omega \neq I$. Por tanto, los procedimientos de inferencia estadística pasan a ser asintóticamente válidos, es decir, aceptables para muestras grandes. Se dice entonces que las varianzas de los estimadores MCO son consistentes con heteroscedasticidad.

La estimación consistente de White puede considerarse como una vía "intermedia" entre MCO y MCG.

EViews permite obtener esta estimación de forma inmediata, puesto que tiene programado este procedimiento. Para ello, basta con proceder del modo habitual (*QUICK / ESTIMATE EQUATION*) y después de introducir la especificación del modelo, manteniendo la elección del método de MCO (*LS – Least Squares*), se pulsa el botón de *Options*, donde debe escogerse *Heteroskedasticity Consistent Covariance White* (*Figura 29*).

Tras confirmarse esta opción, se tendrán finalmente los resultados perseguidos (*Figura 30*). Si se comparan éstos con los de la *Figura 15* (la relativa en nuestro caso a la ecuación AJUSTEMCO), podrá comprobarse cómo las estimaciones MCO de los coeficientes de regresión β resultan las mismas en ambas. En cambio, los valores de los errores estándar asociados a tales coeficientes y, por consiguiente, los de los estadísticos *t*-Student y los *p*-valores asociados a ellos, resultan diferentes. En particular, en este caso los valores de los errores estándar son mayores en la estimación consistente de White. A simple vista, ello parecería peor; sin embargo, estos valores estarían correctamente calculados, ya que tienen en cuenta la presencia de heteroscedasticidad en el modelo.

Este resultado podríamos guardarlo, como siempre pulsando *NAME*, con el nombre AJUSTEWHITE.

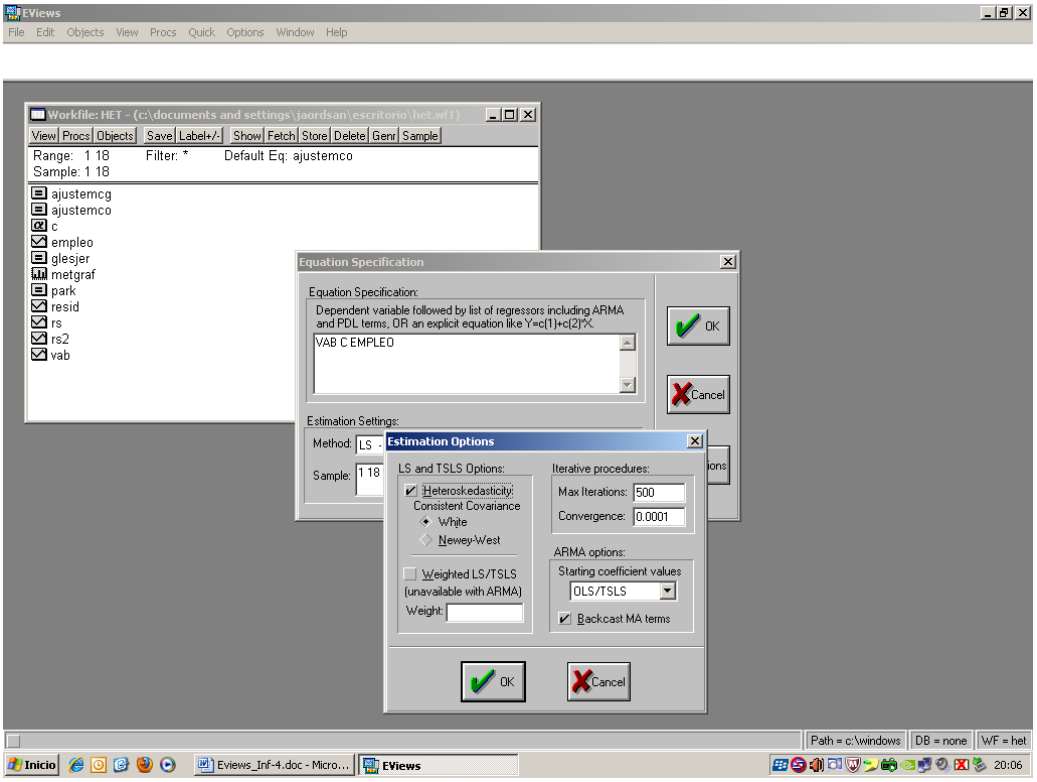


Figura 29

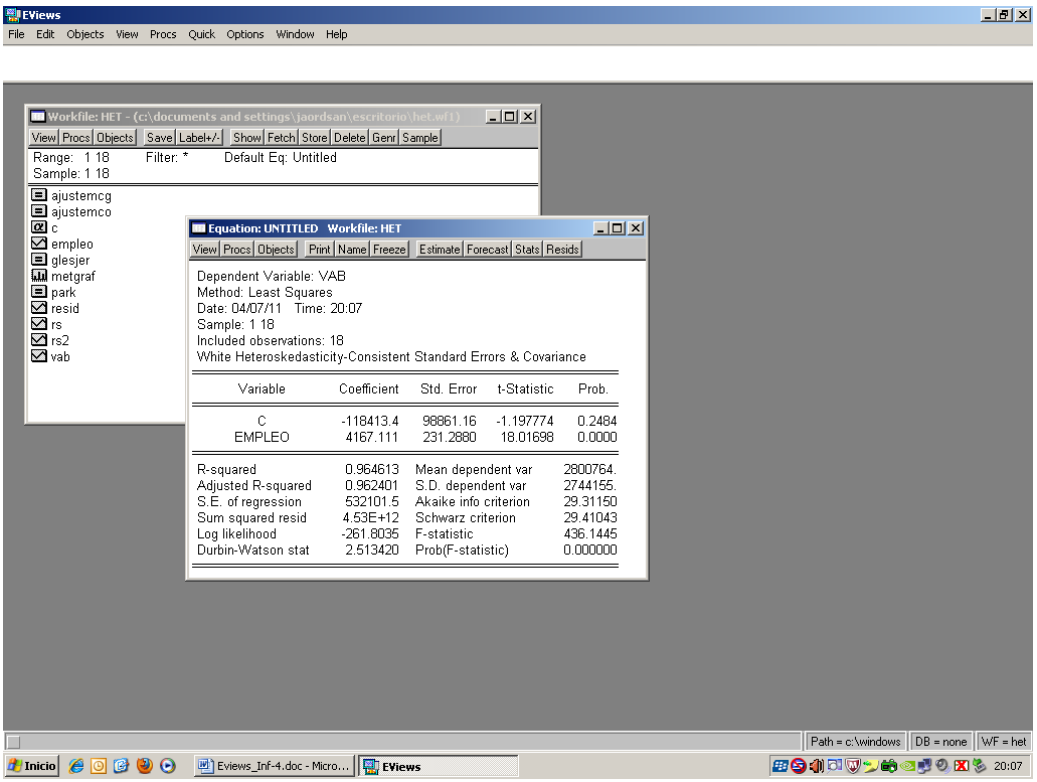


Figura 30

- ***Otra "solución" más a la heteroscedasticidad: la transformación logarítmica***

Para finalizar, cabe reseñar que otra técnica muy empleada en Econometría para abordar los problemas de heteroscedasticidad detectados en un modelo, consiste en realizar una transformación logarítmica de sus variables. Es decir, trabajar con un modelo log-log:

$$\ln Y_i = \beta_1' + \beta_2' \ln X_{2i} + \ldots + \beta_j' \ln X_{ji} + \ldots + \beta_k' \ln X_{ki} + v_i, \quad \forall i = 1,2,\ldots,n.$$

La razón de este proceder reside en que las transformaciones logarítmicas comprimen las escalas en que se miden las variables, reduciendo así la magnitud de la variabilidad del modelo; de este modo, si bien no desaparece plenamente, al menos se atenúa el problema de la heteroscedasticidad. Además, la interpretación de los parámetros resulta sencilla y usual en el ámbito de la Economía, ya que se trataría de elasticidades.[14]

En el caso que nos ocupa, para especificar un modelo de este tipo seleccionaríamos *QUICK / ESTIMATE EQUATION*, escribiendo luego en la ventana que se nos abre:

LOG(VAB) C LOG(EMPLEO)

El resultado sería el que se muestra en la *Figura 31*.

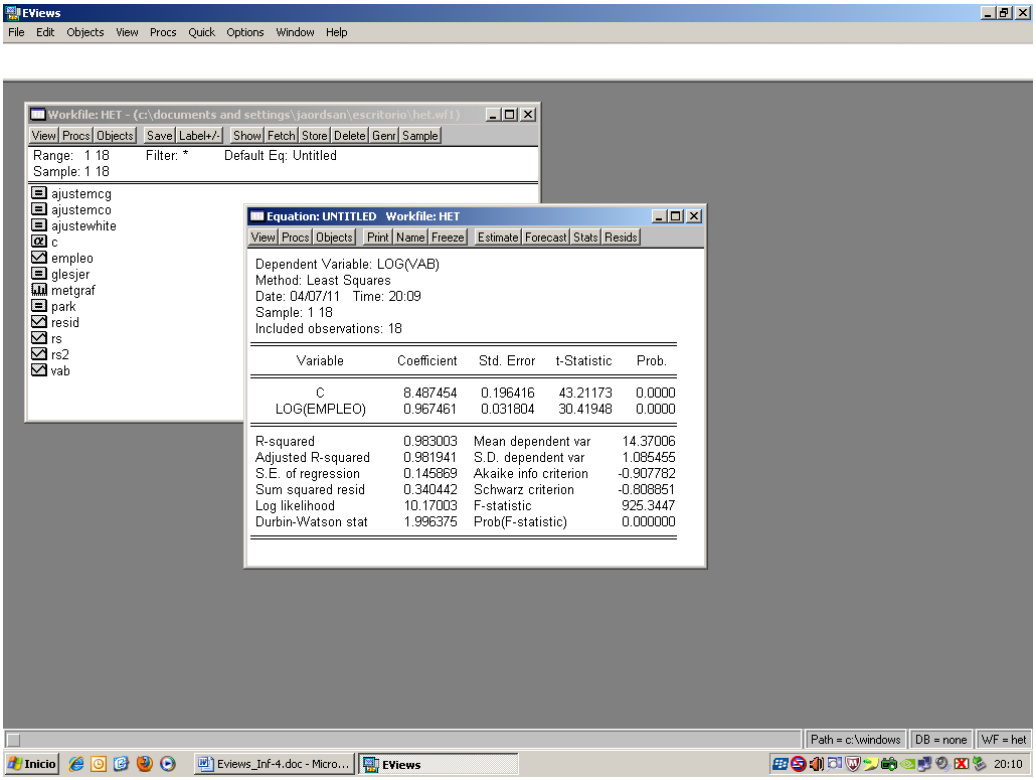


Figura 31

De este modo, concluiríamos este ejercicio. Y si lo deseamos, podemos guardar el fichero de trabajo a través de *FILE / SAVE AS*.

[14] Este método, sin embargo, no sería válido si alguna de las variables del modelo presentase valores negativos. Asimismo, debe reseñarse que resulta más eficaz cuanto mayor es el tamaño muestral.

Detección y tratamiento de la autocorrelación con *EViews*

Abordamos ahora con mayor profundidad el estudio de la autocorrelación en el modelo, consistente en la existencia de correlación en la perturbación aleatoria correspondiente a las distintas observaciones de la muestra considerada.

Según se indicó con anterioridad, la autocorrelación es un problema muy característico de los modelos de series temporales, si bien también puede estar presente con datos de corte transversal.

Al igual que ocurre ante la presencia de heteroscedasticidad, cuando hay autocorrelación las estimaciones por MCO ya no resultan ELIO y para solucionar este aspecto se utiliza como método de estimación alternativo el de MCG.

Si el modelo de regresión lineal $Y = X\beta + u$ presenta únicamente el problema de la autocorrelación, la matriz de $Var - Cov(u)$ puede escribirse de la forma:

$$Var - Cov(u) = \begin{pmatrix} \sigma_u^2 & \sigma_{12} & \cdots & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_u^2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \sigma_u^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \cdots & \sigma_u^2 \end{pmatrix} = \sigma_u^2 \cdot \begin{pmatrix} 1 & \rho_{12} & \cdots & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \vdots & \vdots & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \cdots & \cdots & 1 \end{pmatrix} = \sigma_u^2 \cdot \Omega$$

donde σ_u^2 (la varianza de u) es constante y $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_u^2} = \dfrac{\text{cov}(u_i, u_j)}{\sigma_u^2}$, $i < j$, es el coeficiente de correlación lineal entre u_i y u_j.

Como ya es sabido, la autocorrelación conlleva un problema de exceso de parámetros a estimar en el modelo: su número $\left(k + 1 + \dfrac{n^2 - n}{2}\right)$ es mayor que el número de observaciones (n), por lo que resulta imposible estimarlos. Por esta razón, se hace necesario adoptar algún tipo de supuesto que contribuya a disminuir dicho número. Se imponen así dos tipos de restricciones. Por un lado, restricciones sobre la propia hipótesis de autocorrelación y, por otro, restricciones sobre la estructura de comportamiento de la perturbación aleatoria.

- ***Restricciones sobre la hipótesis de autocorrelación***

1. Como supuesto simplificador, ya señalado desde el principio, estableceremos la no existencia de heteroscedasticidad en la perturbación aleatoria.
2. La covarianza de la perturbación aleatoria correspondiente a distintas observaciones depende únicamente de la distancia entre ellas, es decir:

$$\text{cov}(u_i, u_{i+s}) = \gamma_s, \quad s = 1, 2, \ldots, n-1, \quad i = 1, 2, \ldots, n-s.$$

De esta forma, la matriz de $Var-Cov(u)$ puede expresarse de la forma:

$$Var-Cov(u)=\sigma_u^2\cdot\Omega=\sigma_u^2\cdot\begin{pmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1}\\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2}\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & 1 & \rho_1\\ \rho_{n-1} & \cdots & \cdots & \rho_1 & 1\end{pmatrix},$$

donde σ_u^2 es la varianza (constante) de la perturbación y $\rho_s=\dfrac{\gamma_s}{\sigma_u^2}$ es el coeficiente de correlación lineal entre dos perturbaciones cuya distancia entre ellas es s.

Los elementos no diagonales distintos de la matriz Ω se reducen así a $n-1$ y, en consecuencia, el número total de parámetros a estimar pasa a ser $k+n$. A pesar de la disminución experimentada, ésta no resulta aún suficiente, puesto que el número de parámetros sigue siendo superior al número de observaciones n. Por este motivo, deben establecerse hipótesis adicionales sobre la estructura de comportamiento de u.

- ***Restricciones sobre la estructura de comportamiento de la perturbación aleatoria***

Las restricciones sobre el comportamiento de la perturbación aleatoria dan lugar al planteamiento de dos modelos distintos muy frecuentes: los *modelos autorregresivos* (AR) y los *modelos de medias móviles* (MA).

Un **modelo autorregresivo de orden *p*, AR(p)**, es aquél en el que la perturbación aleatoria u_i presenta una componente sistemática de *p* retardos, que pretende captar la incidencia de momentos anteriores en el tiempo, y una componente estrictamente aleatoria ε_i.

Analíticamente, estos modelos se expresan de la forma siguiente:

$$\text{AR(1): } u_i=\phi_1 u_{i-1}+\varepsilon_i,\quad i=2,\ldots,n$$

$$\text{AR(2): } u_i=\phi_1 u_{i-1}+\phi_2 u_{i-2}+\varepsilon_i,\quad i=3,\ldots,n$$

$$\vdots$$

$$\text{AR(p): } u_i=\phi_1 u_{i-1}+\phi_2 u_{i-2}+\cdots+\phi_p u_{i-p}+\varepsilon_i,\quad i=p+1,\ldots,n.$$

La componente aleatoria ε_i se dice que es *ruido blanco*; esto significa que es una variable aleatoria que satisface las siguientes condiciones[15]:

$$E(\varepsilon_i)=0,\ \ \forall i=1,\ldots,n$$

[15] Obsérvese que la perturbación aleatoria u que cumple las hipótesis establecidas en el modelo de regresión lineal clásico es *ruido blanco*.

$$\text{var}(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall i = 1,\ldots,n$$
$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j, \ i, j = 1,2,\ldots,n.$$

Si nos centramos en un modelo autorregresivo de orden 1, el parámetro ϕ_1 se corresponde con el coeficiente de correlación lineal entre u_i y u_{i-1}. Por esta razón, podemos cambiar su notación por ρ_1, o más abreviadamente, por ρ. De este modo, tendríamos que un modelo de este tipo se escribe habitualmente de la forma:

$$\boxed{\text{AR(1): } u_i = \rho u_{i-1} + \varepsilon_i, \quad i = 2,\ldots,n}.$$

La matriz de $Var - Cov(u)$ en un AR(1) adquiere una forma particular que reduce notablemente el número de parámetros a estimar.

Si calculamos la varianza de u, resulta que[16]:

$$\sigma_u^2 = Var(u_i) = Var(\rho u_{i-1} + \varepsilon_i) = \rho^2 Var(u_{i-1}) + Var(\varepsilon_i) + 2\rho\, Cov(u_{i-1}, \varepsilon_i) = \rho^2\sigma_u^2 + \sigma_\varepsilon^2.$$

Despejando de la expresión anterior, se tiene que:

$$\boxed{\sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}}, \text{ debiendo ser } |\rho| < 1.$$

En cuanto a las covarianzas, se obtienen las siguientes relaciones:

$$\text{cov}(u_i, u_{i+1}) = \rho\sigma_u^2, \quad i = 1,2,\ldots,n-1$$
$$\text{cov}(u_i, u_{i+2}) = \rho^2\sigma_u^2, \quad i = 1,2,\ldots,n-2$$
$$\vdots$$
$$\text{cov}(u_1, u_n) = \rho^{n-1}\sigma_u^2$$

de donde se deduce que los coeficientes de correlación lineal ρ_s entre perturbaciones que distan s retardos, vienen dados por:

$$\boxed{\rho_s = \frac{\text{cov}(u_i, u_{i+s})}{\sigma_u^2} = \frac{\rho^s\sigma_u^2}{\sigma_u^2} = \rho^s, \quad s = 1,2,\ldots,n-1}.$$

Y por tanto, se tiene finalmente que:

$$Var - Cov(u) = \sigma_u^2 \cdot \Omega = \frac{\sigma_\varepsilon^2}{1-\rho^2} \cdot \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \rho \\ \rho^{n-1} & \cdots & \cdots & \rho & 1 \end{pmatrix},$$

[16] Nótese que $Cov(u_{i-1}, \varepsilon_i) = 0$, por depender u_{i-1} de ε_{i-1} y ser ε_i *ruido blanco*.

que quedará determinada conociendo σ_ε^2 y ρ.

De esta manera, en un modelo AR(1) el número de parámetros a estimar se ha reducido en última instancia a $k+2$: $\beta_1, \beta_2, \ldots, \beta_k, \sigma_\varepsilon^2, \rho$.

Un **modelo de media móvil de orden *q* , MA(q)**, se caracteriza porque la perturbación aleatoria u_i presenta una componente aleatoria ε_i (que se supone *ruido blanco*) y un comportamiento sistemático en función de *q* retardos de ésta, ε_{i-s}, $1 \le s \le q$.

Analíticamente, el esquema de comportamiento de la perturbación en un modelo de media móvil es, según el orden:

$$\text{MA(1): } u_i = \alpha\,\varepsilon_{i-1} + \varepsilon_i, \quad i = 2, \ldots, n$$

$$\text{MA(2): } u_i = \alpha_1\varepsilon_{i-1} + \alpha_2\varepsilon_{i-2} + \varepsilon_i, \quad i = 3, \ldots, n$$

$$\vdots$$

$$\text{MA(q): } u_i = \alpha_1\varepsilon_{i-1} + \alpha_2\varepsilon_{i-2} + \cdots + \alpha_q\varepsilon_{i-q} + \varepsilon_i, \quad i = q+1, \ldots, n$$

Centrándonos en un modelo de medias móviles de orden 1, es decir:

$$\boxed{\text{MA(1): } u_i = \alpha\,\varepsilon_{i-1} + \varepsilon_i, \quad i = 2, \ldots, n},$$

se puede ver que la varianza de u adquiere como expresión:

$$\sigma_u^2 = Var(u_i) = Var(\alpha\,\varepsilon_{i-1} + \varepsilon_i) = \alpha^2\,Var(\varepsilon_{i-1}) + Var(\varepsilon_i) + 2\alpha\,Cov(\varepsilon_{i-1}, \varepsilon_i) =$$

$$= \alpha^2\sigma_\varepsilon^2 + \sigma_\varepsilon^2 = (\alpha^2 + 1)\,\sigma_\varepsilon^2.$$

Así pues: $\boxed{\sigma_u^2 = \alpha^2\sigma_\varepsilon^2 + \sigma_\varepsilon^2 = (\alpha^2 + 1)\,\sigma_\varepsilon^2}$.

En cuanto a las covarianzas:

$$\operatorname{cov}(u_i, u_{i+1}) = \alpha\,\sigma_\varepsilon^2, \quad i = 1, 2, \ldots, n-1$$

$$\operatorname{cov}(u_i, u_{i+s}) = 0, \quad s = 2, \ldots, n-1, \quad i = 1, 2, \ldots, n-s.$$

Se dice de este modelo que "no tiene memoria", en el sentido de que cuando la distancia en la perturbación de dos observaciones es mayor que 1, la covarianza entre ellas es nula. Los coeficientes de correlación correspondientes serán entonces, nulos:

$$\boxed{\rho_s = 0, \quad s = 2, \ldots, n-1}.$$

Por otro lado, para observaciones consecutivas se tiene que:

$$\boxed{\rho_1 = \rho = \frac{\operatorname{cov}(u_i, u_{i+1})}{\sigma_u^2} = \frac{\alpha\,\sigma_\varepsilon^2}{\sigma_u^2} = \frac{\alpha\,\sigma_\varepsilon^2}{(1+\alpha^2)\,\sigma_\varepsilon^2} = \frac{\alpha}{1+\alpha^2}}.$$

 ISBN: 978-84-694-7251-4

La matriz de $Var-Cov(u)$ puede escribirse, consiguientemente, como:

$$Var-Cov(u)=\sigma_u^2\cdot\Omega=\sigma_u^2\cdot\begin{pmatrix}1&\rho&0&\cdots&0\\ \rho&1&\rho&\cdots&0\\ \vdots&\ddots&\ddots&&\vdots\\ \vdots&&\ddots&1&\rho\\ 0&\cdots&\cdots&\rho&1\end{pmatrix}=(1+\alpha^2)\sigma_\varepsilon^2\cdot\begin{pmatrix}1&\frac{\alpha}{1+\alpha^2}&0&\cdots&0\\ \frac{\alpha}{1+\alpha^2}&1&\frac{\alpha}{1+\alpha^2}&\cdots&0\\ \vdots&&\ddots&\ddots&\vdots\\ \vdots&&&1&\frac{\alpha}{1+\alpha^2}\\ 0&\cdots&\cdots&\frac{\alpha}{1+\alpha^2}&1\end{pmatrix}$$

que quedará determinada conociendo: σ_u^2 y ρ; ó equivalentemente: σ_ε^2 y α. De nuevo, pues, hemos reducido el número de parámetros a estimar a $k+2$.

Para detectar la posible existencia de autocorrelación, se pueden aplicar distintos métodos. En particular, podemos citar:

- Métodos gráficos: representaciones de los residuos y correlogramas
- Contrastes analíticos: Durbin-Watson, Breusch-Godfrey…

A continuación, vamos a describir los métodos indicados, aplicándolos de manera práctica con ayuda de *EViews* sobre un modelo que especificaremos para tomar como ejemplo.

Así, vamos a plantear la estimación, a través de un modelo de regresión lineal, del consumo público de cierto país en función de su Producto Interior Bruto a precios de mercado (Ejercicio nº 42 del Boletín del presente Tema).

Con este fin usaremos la información referida al periodo 1998-2010 que nos proporciona el fichero ***aut.wf1***, disponible en el espacio virtual de la Asignatura en la *WebCT*, relativa a las variables:

- **CP**: Consumo público (en millones de unidades monetarias)
- **PIB**: Producto Interior Bruto a precios de mercado (en millones de unidades monetarias)

Tras abrir el fichero en *EViews* a través de la sucesión de comandos *FILE / OPEN / WORKFILE*, procederemos a estimar mediante *QUICK / ESTIMATE EQUATION* el modelo:

$$\boxed{CP=\beta_1+\beta_2 PIB+u}.$$

Con este propósito, en el usual cuadro de diálogo resultante deberemos escribir:

CP C PIB

Tras pulsar *OK*, se obtienen los resultados que muestra la *Figura 32*.

La observación de todos los coeficientes, expresiones y estadísticos obtenidos podría hacernos pensar que el modelo resulta aceptable; sin embargo, por la naturaleza de los datos, de corte temporal, sospechamos que puede haber problemas de autocorrelación en la perturbación aleatoria.

Vamos a guardar este ajuste con el nombre AJUSTEMCO, por ejemplo, seleccionando la opción *NAME* de la *Ventana de Ecuación*. De este modo, podremos tener así acceso a la estimación cuando lo deseemos.

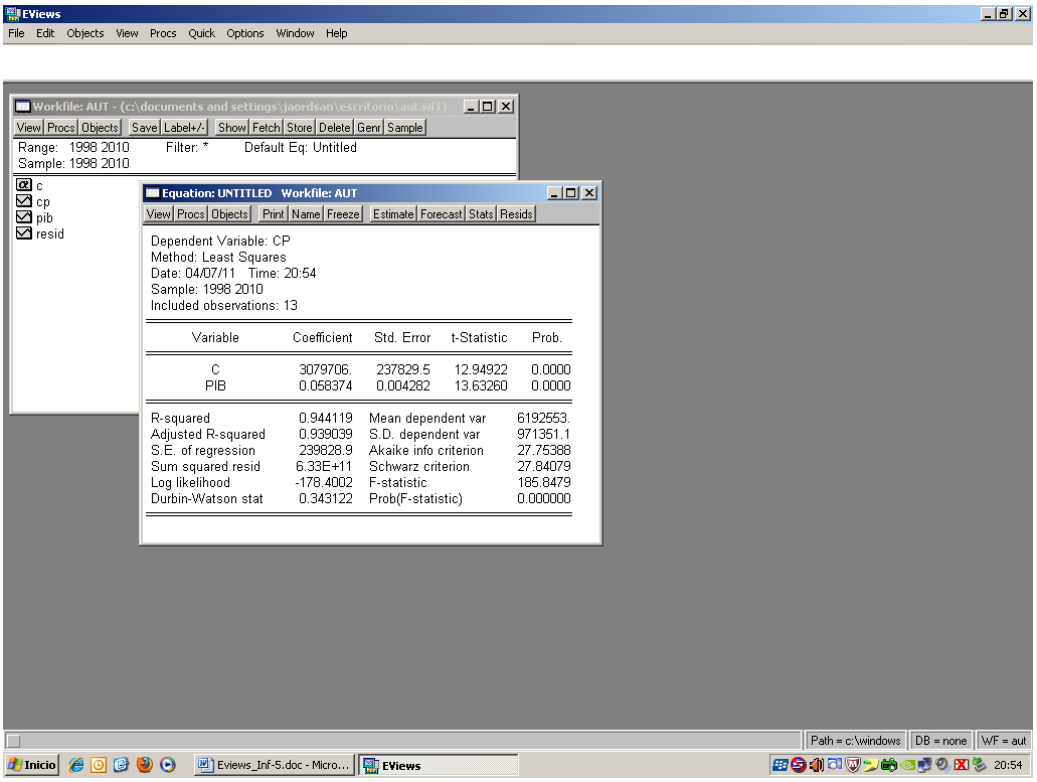


Figura 32

- ***Métodos gráficos de detección***

El análisis de la posible presencia de autocorrelación en un modelo debería basarse, teóricamente, en el estudio del comportamiento de la perturbación aleatoria. Pero dado que ésta no es observable, en la práctica se analiza la serie de los residuos MCO que, como sabemos, es su estimación.

Los métodos gráficos más usuales de detección de la autocorrelación consisten, por un lado, en representar los residuos, bien a lo largo del tiempo, o bien en relación a los del periodo anterior; y, por otro, en dibujar sus correlogramas. A continuación, veremos cada uno de ellos de forma más concreta.

Previamente, vamos a crear en *EViews* una serie de datos que contenga los residuos de la estimación AJUSTEMCO. Recordemos que *EViews*, al hacer el ajuste, incluye automáticamente en la variable *Resid* los residuos del modelo. Pero, si se lleva a cabo otra estimación, estos valores serán reemplazados por los residuos del último modelo

estimado. Puesto que nos interesa trabajar con los residuos de nuestro modelo AJUSTEMCO, optamos entonces por convertirlos en un objeto específico. De este modo, a nuestra serie de residuos MCO la llamaremos, por ejemplo, RS. Para hacer esto, elegimos la opción *GENR* en la Ventana de Trabajo y escribimos luego en la ventana que se abre (*Enter equation*): RS = RESID. Tras aceptar (*OK*), tendremos en el *Directorio de Objetos* de nuestro fichero de trabajo el nuevo objeto RS.

Llegados a este punto, ya estamos en condiciones de llevar a cabo los métodos gráficos para detectar la existencia de autocorrelación.

En primer lugar, analizaremos cuál es el comportamiento gráfico que se observa al representar los residuos en relación con el tiempo. Cuando se observan rachas de residuos con el mismo signo, suele ser indicativo de autocorrelación. Si hay pocos cambios de signo, la autocorrelación será positiva; por el contrario, cuando hay muchas rachas, se tiene evidencia de autocorrelación negativa.

Veamos qué sucede en nuestro ejemplo al aplicar este método gráfico con *EViews*. Para ello, nos iremos a la *barra principal de menús* y elegiremos: *QUICK / GRAPH*. Al abrirse la ventana correspondiente, escribiremos la serie a representar: *RS*. Tras hacer "clic" en *OK*, seleccionaremos *Line Graph* en tipo de gráfico. Después, en *SHOW OPTIONS*, optaremos por *Symbols only* en el apartado *Line Graphs*. Para concluir, pulsaremos *OK*.

Podemos ver el resultado en la *Figura 33*, que podríamos grabar con el nombre GRAF1.

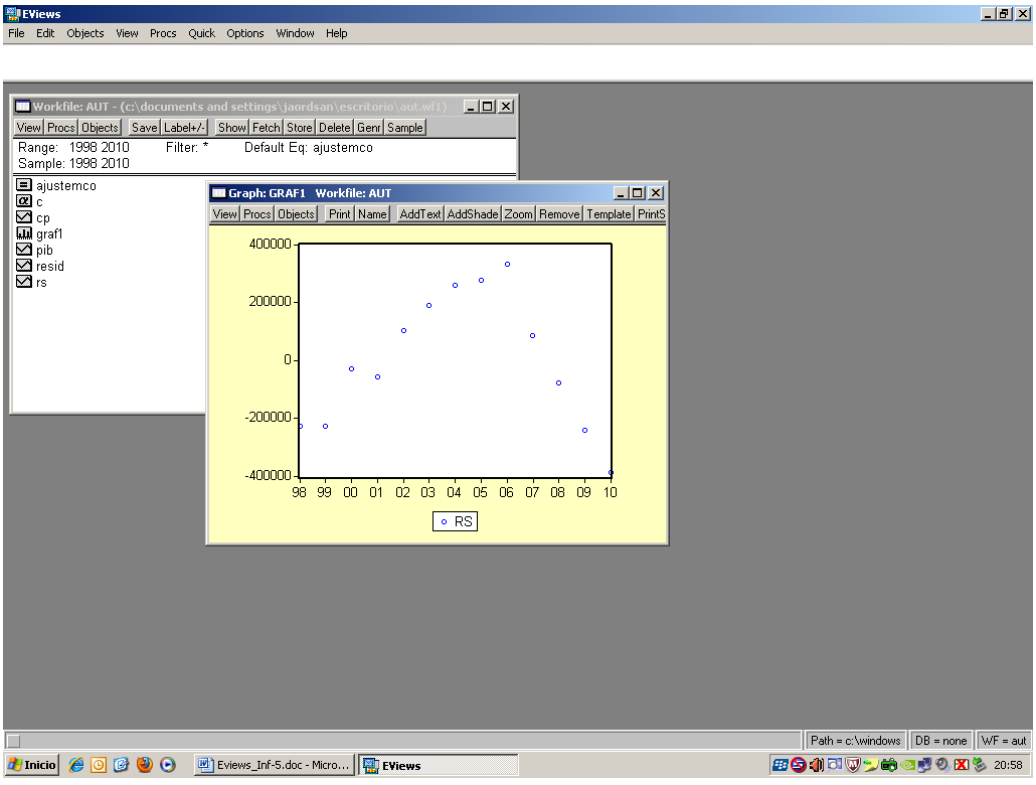


Figura 33

En el periodo 1998-2001 hay una racha de residuos negativos; le sigue una racha de residuos positivos correspondientes al periodo 2002-2007; y, finalmente, otra racha de residuos negativos en el periodo 2008-2010. Podemos por tanto intuir la existencia de autocorrelación y, además, como hay únicamente dos cambios de signo, sería positiva.

Otro procedimiento gráfico interesante consiste en la representación de los residuos frente a los del periodo anterior. Este método es útil para detectar, al menos, la existencia de autocorrelación que sigue un esquema AR(1). Éste será el caso si se observa una relación lineal clara entre ambas variables, ya que indicaría que la perturbación aleatoria u_i es una función lineal de la perturbación aleatoria u_{i-1}, tal como formula el modelo AR(1). Además, podremos indicar si se trata de autocorrelación positiva o negativa, dependiendo del signo de la pendiente de la recta que ajusta estos puntos.

Para hacer esta gráfica con *EViews*, seleccionamos en el menú principal *QUICK / GRAPH*, y escribimos las series que deseamos representar, RS(-1) y RS, recordando que en primer lugar debe ir aquélla que queramos posicionar sobre el eje *X*. Obtendremos el gráfico que aparece en la *Figura 34*, bajo la selección previa del tipo de gráfico *Scatter Diagram*. La nube de puntos resultante, que podemos nombrar como GRAF2, parece indicar claramente la existencia de una relación lineal directa entre ambas variables, sugiriendo la posible presencia de autocorrelación positiva, siguiendo, al menos, un esquema AR(1).

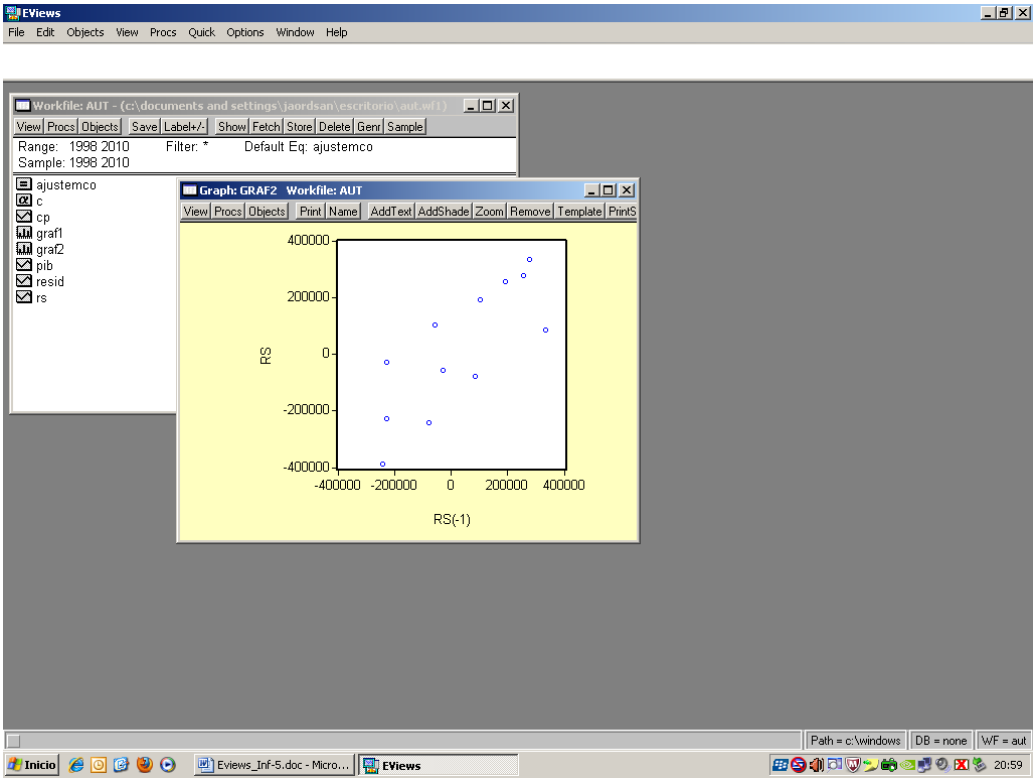


Figura 34

Otro método gráfico habitual para evidenciar la posible existencia de autocorrelación es la realización del *correlograma* de los residuos.

Aquí podemos encontrar dos variantes. La primera se refiere a la representación gráfica de la llamada *función de autocorrelación* (FAC), es decir, de los coeficientes de correlación (ρ_s) entre distintas observaciones de la serie de los residuos en función de la distancia o retardo *s* que hay entre ellas. La otra posibilidad consiste en trabajar con los coeficientes de correlación parcial (ϕ_s), que miden la correlación entre dos observaciones de dicha serie en función de los retardos, sin tener en cuenta la influencia de las demás observaciones, obteniéndose así la representación de la denominada *función de autocorrelación parcial* (FACP).

El comportamiento de la FAC y la FACP se puede resumir en el siguiente cuadro:

	Función de Autocorrelación (FAC)	Función de Autocorrelación Parcial (FACP)
AR(*p*)	Muchos coeficientes no nulos que decrecen con el retardo como mezcla de exponenciales y senoides	*p* primeros coeficientes no nulos y el resto cero
MA(*q*)	*q* primeros coeficientes no nulos y el resto cero	Muchos coeficientes no nulos que decrecen con el retardo como mezcla de exponenciales y senoides

En la práctica, se representan los correlogramas de los residuos y no de la serie de la perturbación aleatoria, que en teoría es lo que debería hacerse. Esto da lugar generalmente a que los correlogramas estimados no sigan claramente el comportamiento teórico esperado. Por esta razón, las consideraciones prácticas a tener en cuenta serían:

- La identificación del proceso se lleva a cabo con la FAC para los AR y la FACP para los MA. Si la FAC presenta un mayor número de coeficientes significativos en los primeros retardos, estaríamos ante un proceso AR; en cambio, si esto ocurre en la FACP, sería un proceso MA.
- La identificación del orden del modelo se realiza en los AR con la FACP y en los MA con la FAC, y vendrá dado por el número de coeficientes que se muestren significativos en la función correspondiente.
- Un coeficiente significativo para un retardo que no corresponda a los primeros valores de retardos, en principio, no será relevante.

Para llevar a cabo los correlogramas, el modo de proceder con *EViews* es seleccionar, en la *Ventana de Ecuación* de nuestro modelo AJUSTEMCO, la opción *VIEW / RESIDUAL TESTS / CORRELOGRAM – Q-STATISTICS*, indicando el número de retardos que se quieren incluir (*Lags to include*). Por defecto, *EViews* nos plantea 11 retardos (*Figura 35*). Generalmente este número es suficiente, por lo que nos quedaremos con esta propuesta.

El resultado se muestra en la *Figura 36*, donde podemos apreciar que hay un mayor número de coeficientes significativos (o próximos a serlo para un $\alpha = 5\%$) en la FAC que en la FACP, lo que nos indica que estamos antes un modelo AR para la perturbación aleatoria.

Para determinar el orden, nos fijamos entonces en los coeficientes de la FACP que se salen del intervalo de confianza señalado con líneas verticales de trazo discontinuo (para un 95% de nivel de confianza). Vemos que únicamente el primero de ellos está fuera de dicho intervalo, lo que nos permite pensar que estamos ante un modelo AR de orden 1, al igual que nos sugerían los otros gráficos representados anteriormente.

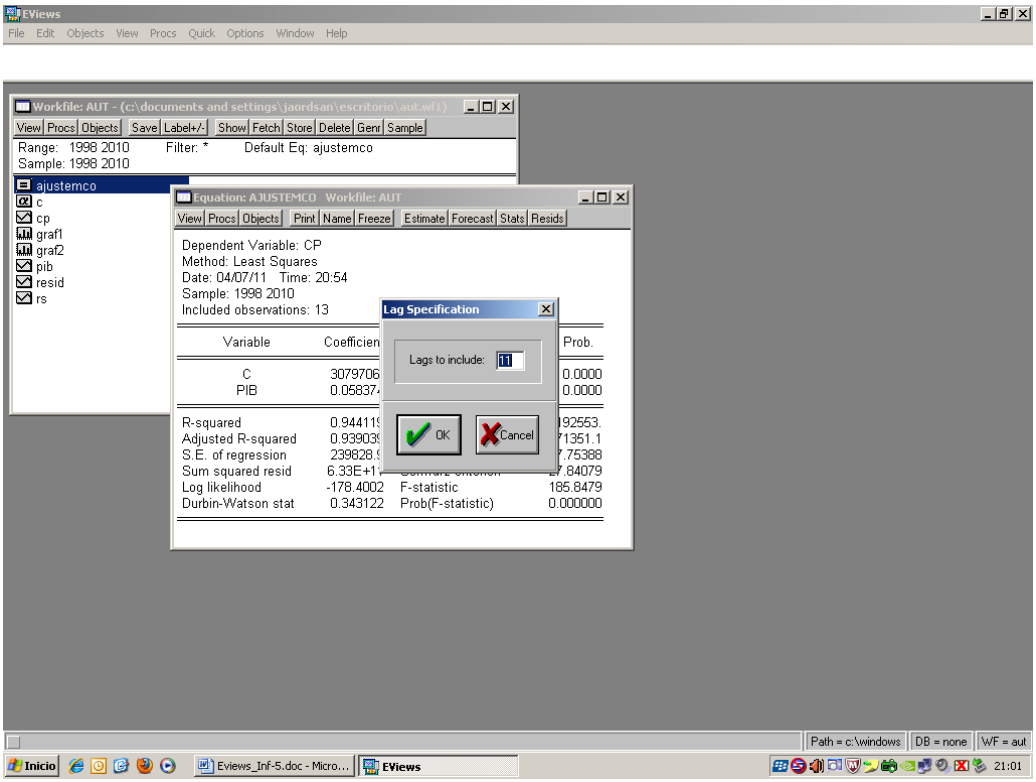


Figura 35

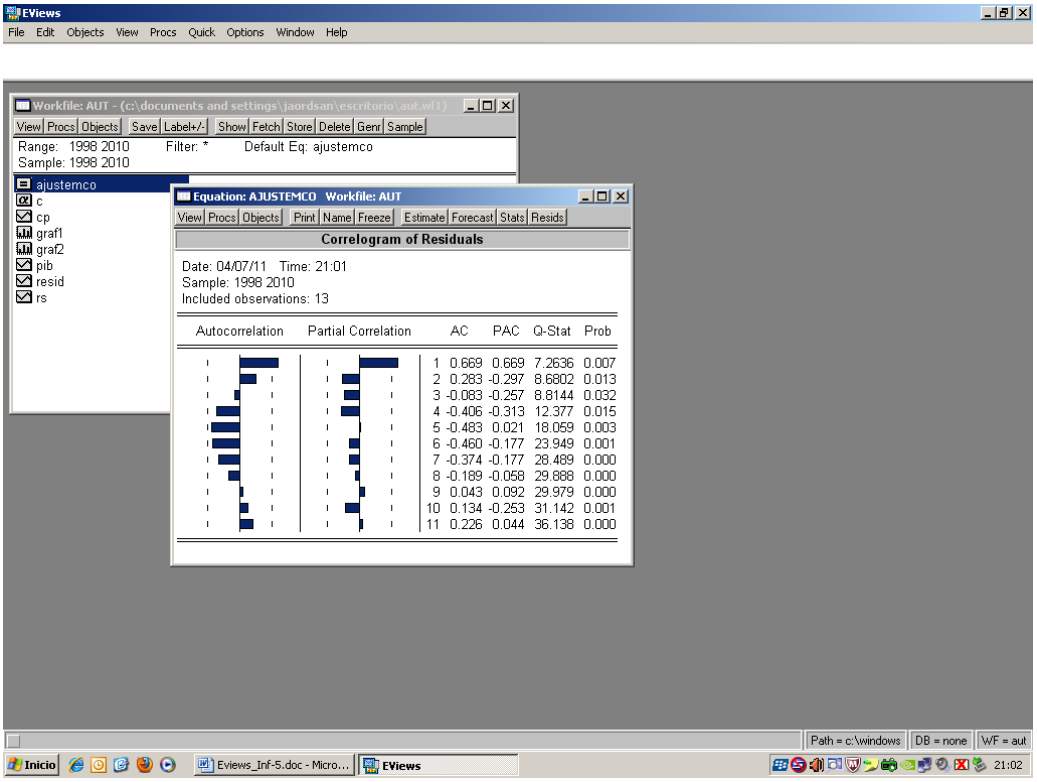


Figura 36

- ***Contrastes analíticos***

Además de los métodos gráficos, existen diversas posibilidades de tipo analítico para contrastar la existencia de autocorrelación. La hipótesis nula que se establece en todos los casos es la ausencia de autocorrelación y la diferencia entre unos contrastes y otros radica en la hipótesis alternativa que se formula. A veces, éstas son muy generales y únicamente nos indican la existencia de autocorrelación, mientras que en otros casos, no sólo plantean la existencia de autocorrelación, sino también el esquema concreto de la misma que está presente en el modelo.

A continuación, vamos a revisar dos de estos contrastes.

El **contraste de Durbin-Watson** permite comprobar la existencia de autocorrelación de tipo AR(1). Recordemos que ésta responde al siguiente esquema: $u_i = \rho u_{i-1} + \varepsilon_i$, $i = 2,\ldots,n$, donde ε_i es ruido blanco.

La hipótesis nula que se plantea en el contraste de Durbin-Watson es:

$$H_0: \text{Ausencia de autocorrelación de tipo AR(1) } (\rho = 0)$$

La hipótesis alternativa puede tener dos formulaciones distintas:

$$H_1: \text{Autocorrelación positiva de tipo AR(1) } (\rho > 0)$$

$$\text{o bien, autocorrelación negativa de tipo AR(1) } (\rho < 0)$$

El estadístico de Durbin-Watson se define, a partir de los residuos, como[17]:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}.$$

Asimismo, se puede demostrar que $DW \approx 2(1-\hat{\rho})$, siendo $\hat{\rho} \approx \frac{\sum_{i=2}^{n} e_i \cdot e_{i-1}}{\sum_{i=2}^{n} e_i^2}$.

Como $-1 \leq \hat{\rho} \leq 1$, se deduce que $0 \leq DW \leq 4$. A partir de aquí, se pueden extraer las siguientes conclusiones:

a) Cuando $\hat{\rho} \approx 1$ (que refleja autocorrelación positiva plena en la perturbación aleatoria para las distintas observaciones de la muestra), el estadístico *DW* tomará un valor muy próximo a 0.

b) Cuando $\hat{\rho} \approx -1$ (que refleja autocorrelación negativa plena en la perturbación aleatoria para las distintas observaciones de la muestra), el estadístico *DW* tomará un valor muy próximo a 4.

c) Cuando $\hat{\rho} \approx 0$ (que refleja ausencia de autocorrelación en la perturbación aleatoria para las distintas observaciones de la muestra), el estadístico *DW* tomará un valor muy próximo a 2.

Este contraste tiene, sin embargo, algunas limitaciones:

1. Sólo permite contrastar la existencia de autocorrelación de tipo AR(1).
2. No es válido si entre las variables explicativas del modelo se encuentra la variable endógena retardada.
3. No es estrictamente válido si el modelo no tiene ordenada en el origen.
4. La distribución del estadístico de Durbin-Watson, *DW*, depende de la matriz de datos X de las variables explicativas del modelo, lo que hace imposible obtener una tabla de valores críticos única a partir de la cual se puedan efectuar contrastes de hipótesis independientemente de la muestra utilizada.

Este último problema fue resuelto por Durbin y Watson con la obtención de cotas inferior (d_L) y superior (d_U) sobre el conjunto de todas sus posibles distribuciones de probabilidad, para cada nivel de significación y cada tamaño muestral. De esta manera, gráficamente se pueden representar las distribuciones de probabilidad de dichas cotas,

[17] Algunos autores inician el sumatorio de la expresión del denominador en $i = 1$.

las cuales son independientes de X, aunque sí dependen del nivel de significación, del número de variables explicativas y del tamaño muestral.

La representación gráfica de estas distribuciones (*Figura 37*) nos permite distinguir distintas zonas, de modo que:

- Si $DW < d_L$, entonces existe autocorrelación positiva.
- Si $DW > d_U$, entonces no existe autocorrelación positiva.
- Si $d_L < DW < d_U$, nos encontramos en una zona de duda, si bien está claro que no existe autocorrelación negativa.
- Si $DW > 4 - d_L$, entonces existe autocorrelación negativa.
- Si $DW < 4 - d_U$, entonces no existe autocorrelación negativa.
- Si $4 - d_U < DW < 4 - d_L$ nos encontramos en una zona de duda, aunque está claro que no existe autocorrelación positiva.

En el caso de que el estadístico *DW* caiga en alguna de las zonas de duda, una forma conservadora de proceder sería actuar como si existiese autocorrelación aunque no la hubiese, en lugar de lo contrario.

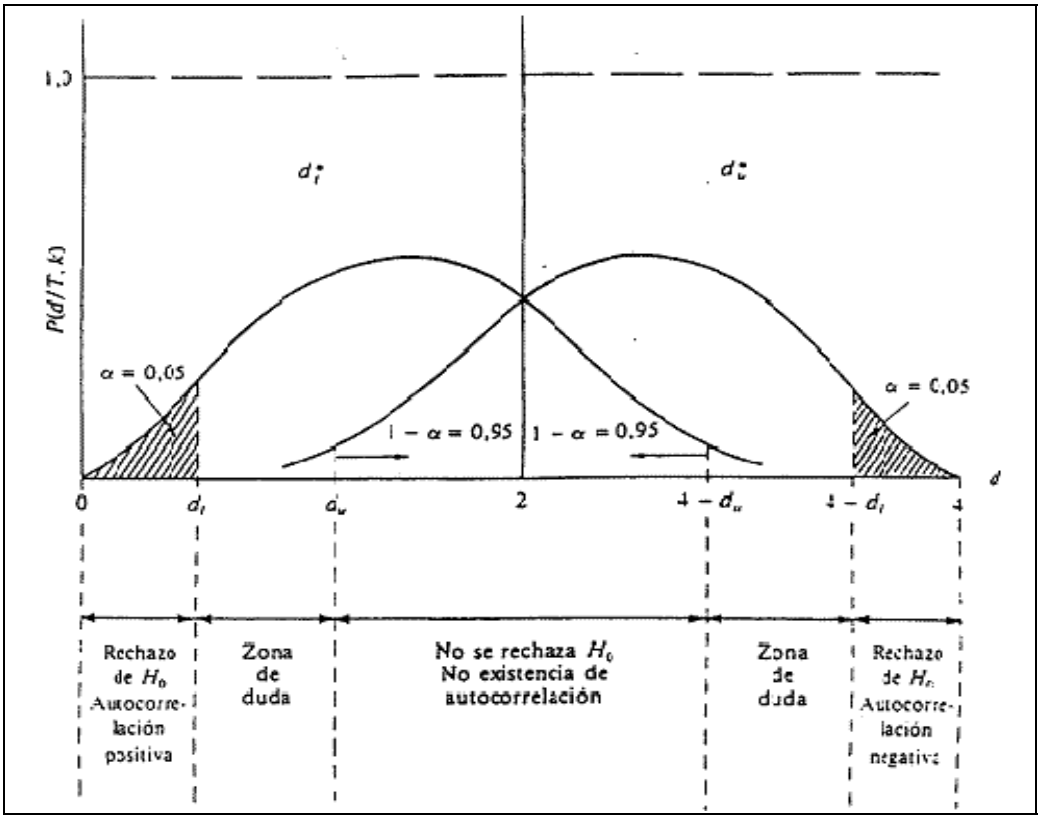


Figura 37

 ISBN: 978-84-694-7251-4

Cuando trabajamos con *EViews*, el estadístico de Durbin-Watson aparece directamente entre los estadísticos de la salida de resultados básicos (*Estimation Output*) que se muestran al hacer la estimación.

Así, en la salida de la regresión AJUSTEMCO (*Figura 32*) se puede encontrar el valor del estadístico Durbin-Watson (*Durbin-Watson stat*): 0,343122. Para un nivel de significación del 5% las cotas inferior y superior con un tamaño muestral igual a 13 son 1,01 y 1,34, respectivamente (estas cotas se buscan en las tablas correspondientes). Estamos entonces en el caso en que $DW < d_L$, con lo que, además de rechazar la hipótesis nula de no autocorrelación, podemos decir que la autocorrelación presente es positiva y de tipo autorregresivo de orden 1.

El **contraste de Breusch-Godfrey** permite contrastar la existencia de autocorrelación de forma más general que el contraste de Durbin-Watson, puesto que es válido tanto para procesos AR como MA y para cualquier orden de autocorrelación. Tiene además la ventaja, con respecto al contraste de Durbin-Watson, de que se puede aplicar incluso cuando entre las variables explicativas está presente la variable endógena retardada.

Las hipótesis nula y alternativa que se plantean en este contraste son:

H_0 : Ausencia de autocorrelación de orden $\leq p$

H_1 : Autocorrelación de orden $\leq p$ (AR(p) ó MA(p)).

Este contraste parte de la estimación por MCO del modelo analizado en cuestión. Seguidamente se estima una regresión auxiliar de sus residuos en función de *p* retardos de éstos y de las variables explicativas del modelo (pudiendo, incluso, introducirse variables endógenas retardadas):

$$e_i = \alpha_1 + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki} + \lambda_1 e_{i-1} + \ldots + \lambda_p e_{i-p} + v_i .$$

Se calcula entonces el coeficiente de determinación R^2 de esta regresión auxiliar y con él se construye el estadístico de prueba siguiente:

$$\chi^2_{BG} = n \cdot R^2 \rightarrow \chi^2_p,$$

donde *n* es el número de observaciones que conforman la muestra y *p* es el número de retardos de los residuos que se introducen en la regresión auxiliar.

EViews ofrece la posibilidad de realizar directamente este contraste, solicitando como información el número de retardos *p* a introducir en la regresión auxiliar que se plantea. La forma práctica de operar es ir seleccionando progresivamente los retardos, hasta que se acepte la hipótesis nula de no autocorrelación del orden indicado o el contraste de significación individual del residuo correspondiente al último retardo introducido indique no rechazo de la hipótesis nula. En ese momento, acabará el proceso y el orden

de la autocorrelación será el del último retardo que haya resultado significativo (si ninguno lo es, se aceptará entonces la hipótesis nula de ausencia de autocorrelación).

El contraste de Breusch-Godfrey presenta como inconveniente el hecho de que si bien puede indicar el orden de retardos hasta el que llega la autocorrelación de la perturbación aleatoria en un modelo (caso de estar presente), no permite sin embargo discernir cuál es el esquema exacto de la misma; esto es, si es de tipo AR, o bien de tipo MA.

Para aplicar el contraste de Breusch-Godfrey a nuestro ejemplo, abriremos la *Ventana de Ecuación* AJUSTEMCO y seleccionaremos *VIEW / RESIDUAL TESTS / SERIAL CORRELATION LM TEST* (*Figura 38*). Aquí escribiremos, en principio, 1 retardo.

En la *Figura 39*, vemos que *EViews* nos ofrece el valor (8,473176) del estadístico experimental de Breusch-Godfrey (*Obs*R-squared*), siendo su *p-valor* asociado 0,003604, por lo que incluso para un nivel de confianza del 99% dicho estadístico se sitúa en la región crítica, lo que nos lleva a rechazar la hipótesis nula de no autocorrelación. Además, en la parte inferior de la salida se ofrece la regresión auxiliar de los residuos MCO en función de las variables explicativas del modelo y de los retardos elegidos de dichos residuos (en este caso 1). En relación con dicha salida, debemos fijarnos en que el *p-valor* del estadístico *t* correspondiente al primer retardo de los residuos RESID(-1) es igualmente muy pequeño (0,0015), por lo que para los niveles de confianza más exigentes también resulta significativo.

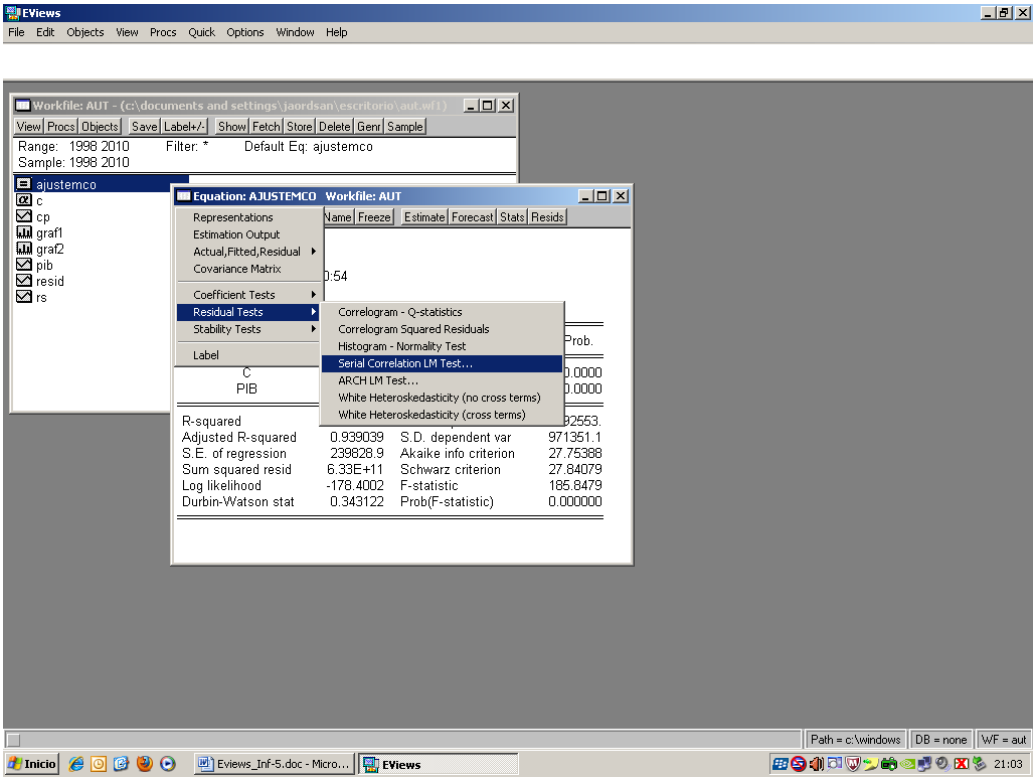


Figura 38

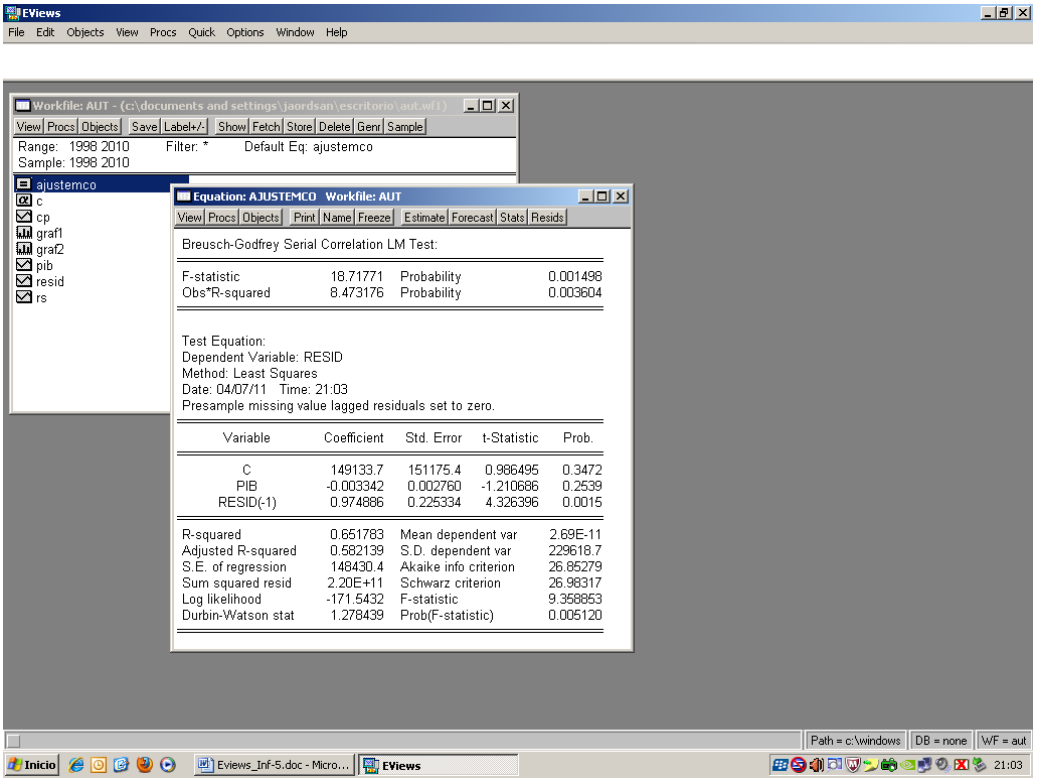


Figura 39

Seguidamente se puede operar de manera análoga con 2 retardos en los residuos de la regresión auxiliar (*Figura 40*).

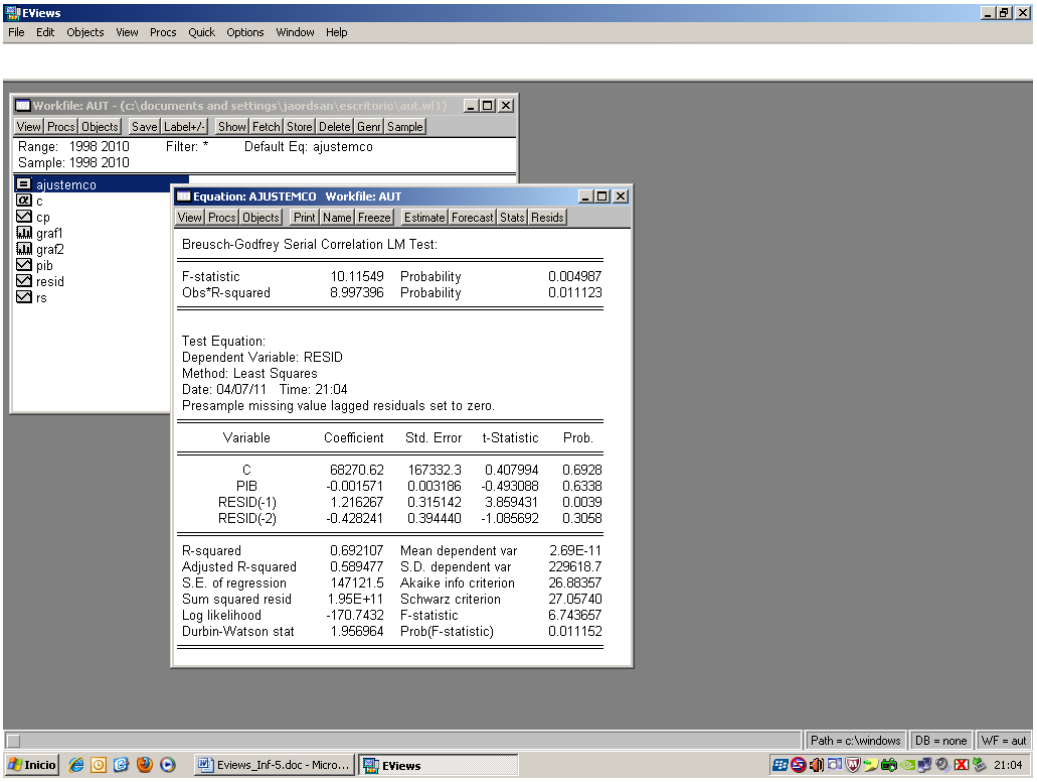


Figura 40

El estadístico experimental del contraste de Breusch-Godfrey, cuyo valor es 8,997396, tiene ahora una probabilidad asociada de 0,011123 y si trabajásemos con un nivel de significación del 5%, el estadístico se situaría, como en el caso anterior, en la región de rechazo de la hipótesis nula de no autocorrelación (en cambio, no sucedería así para un 1%). Si atendiésemos a la regresión auxiliar, el contraste de significación individual del segundo retardo de los residuos RESID(-2) llevaría a no rechazar la hipótesis nula, dado que el *p-valor* de su estadístico t experimental es 0,3058. Así pues, este segundo retardo parece que ya no es significativo en el comportamiento de los residuos MCO. La conclusión del contraste de Breusch-Godfrey es, por tanto, que existe autocorrelación en la perturbación aleatoria y que ésta es de orden 1.

- ***Solución a la autocorrelación: estimación del modelo por el método de MCG***

Como bien sabemos, el método de MCG es el método alternativo a MCO que debe aplicarse para obtener estimadores ELIO cuando el modelo presenta autocorrelación.

El modo de operar con *EViews* es bien sencillo, una vez que se ha identificado la estructura de comportamiento de la perturbación aleatoria del modelo. En el presente ejemplo, hemos visto que parece que se trata de un modelo autorregresivo de orden 1. Bastará entonces con añadir al ajuste de regresión el término AR(1). Para ello, seleccionamos *QUICK / ESTIMATE EQUATION* y escribimos en el cuadro de diálogo:

CP C PIB AR(1)

Después de aceptar (*OK*), el resultado obtenido se puede observar en la *Figura 41*.

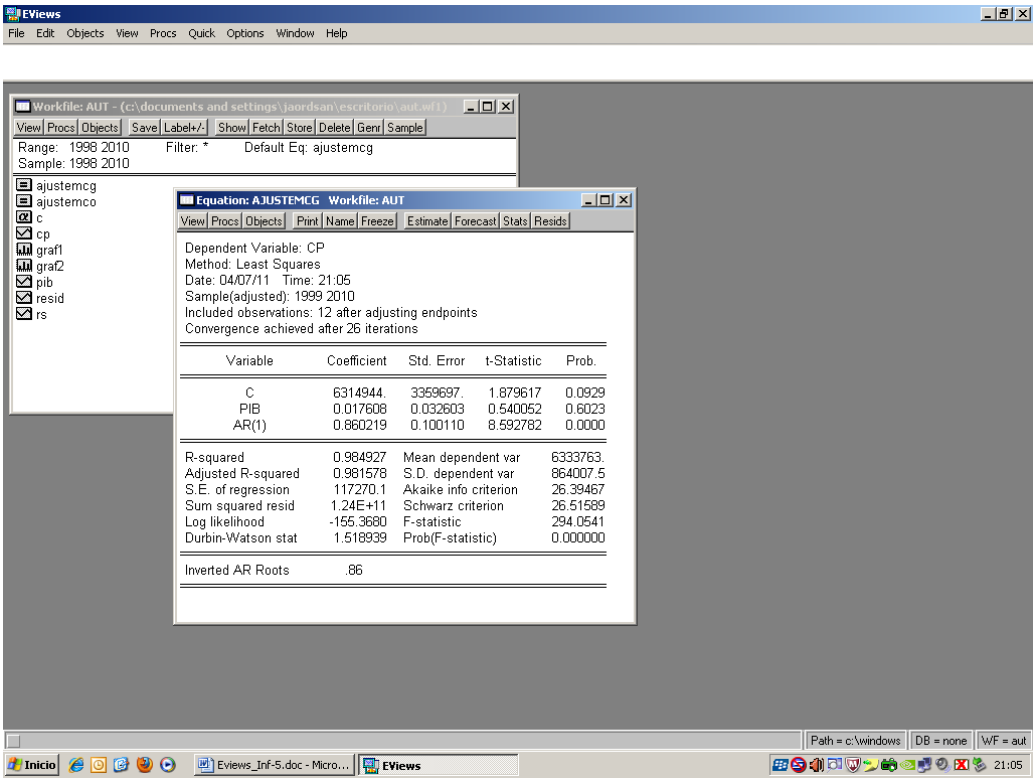


Figura 41

El coeficiente asociado a AR(1) es la estimación del coeficiente de correlación ρ correspondiente al esquema que sigue la perturbación. Puede observarse que dicho coeficiente es estadísticamente significativo a un nivel de confianza máximo de más del 99% y que su valor (0,860219) está bastante próximo a 1, lo que reafirma la existencia de autocorrelación positiva (y además muy elevada) de orden 1.

Los coeficientes asociados a C y PIB son los estimadores MCG del modelo, que resultan ser ELIO.

Podemos guardarlo con el nombre AJUSTEMCG, en *NAME*.

Para concluir, cabe reseñar que si la autocorrelación siguiera un esquema AR(p), se añadirían a la estimación los términos AR(1), AR(2), …, AR(p), siendo sus parámetros asociados las estimaciones de los coeficientes ϕ_i de la regresión: $u_i = \phi_1 u_{i-1} + \cdots + \phi_p u_{i-p} + \varepsilon_i$. En el caso de que se tratase de un modelo de medias móviles de orden *q*, MA(q), se añadirían a la estimación los términos MA(1), MA(2) hasta MA(q); los parámetros asociados a estas variables serán las estimaciones de los coeficientes α_i de la regresión $u_i = \alpha_1 \varepsilon_{i-1} + \alpha_2 \varepsilon_{i-2} + \cdots + \alpha_q \varepsilon_{i-q} + \varepsilon_i$, que establece el modelo MA(q).

Tras esto, el ejercicio estaría concluido y, si lo deseamos, podemos guardar el fichero de trabajo a través de *FILE / SAVE AS*.

 ISBN: 978-84-694-7251-4

TEMA 5

Modelos con variables dependientes discretas

Con la introducción de variables ficticias en el modelo ya vimos que las variables de tipo cualitativo podían estar presentes en el mismo, actuando en tal caso como explicativas.

Pero la presencia de una variable cualitativa en un modelo también puede darse en el papel de variable dependiente. Quizás nuestro objetivo sea estudiar los factores que influyen en la ocurrencia o no de un determinado suceso o fenómeno económico, como la disponibilidad de vivienda, la compra de un determinado bien o el disfrute de un servicio. De esta forma, surgen los *modelos con variables dependientes discretas*.

5.1. Definición de los modelos de elección discreta binaria.-

Los modelos de elección discreta se caracterizan por el hecho de permitir reflejar la elección o toma de decisión por parte de un individuo entre diversas alternativas posibles. Si éstas son solamente dos, hablaremos de modelos de elección binaria. En el caso de que se traten de más de dos, entonces estaremos ante los denominados modelos de elección discreta de respuesta múltiple, donde se pueden encontrar varios tipos, aunque básicamente se puede hacer una primera distinción en relación a si las diferentes alternativas posibles están ordenadas o no.

En nuestro caso, nos vamos a centrar en el estudio de los modelos de elección discreta binaria, esto es: al individuo se le plantea tomar una decisión de entre únicamente dos posibilidades mutuamente excluyentes. La variable endógena Y de estos modelos adopta dos únicos valores numéricos discretos, normalmente 0 y 1; de modo que si el sujeto se decanta por la ocurrencia del suceso objeto de estudio entonces Y toma el valor 1, y 0, si no es así.

Estos modelos facilitan la tarea de identificación de las características o factores que inciden en un comportamiento de los individuos diferente ante los procesos de decisión[1]. Algunas situaciones a las que les son aplicables son, por ejemplo: acudir o no al médico, disponer o no de cobertura aseguradora, adquirir o no una vivienda, etc.

Podemos plantear este tipo de modelos como sigue.

[1] La base económica sobre la que se fundamentan estos modelos es la Teoría de la Utilidad de Von Neumann-Morgenstern, establecida en 1944. De acuerdo con ésta, los sujetos se comportan ante una disyuntiva, de tal modo que tratan de maximizar la utilidad esperada que les reporta cada una de las alternativas posibles sobre las que han de decidirse.

Si suponemos que la variable *Y* depende de un conjunto de variables explicativas *X*, de manera que:

$$Y_i = F(X_i\beta) + u_i,$$

donde $X_i = (X_{1i}, X_{2i}, \ldots, X_{ki})$ hace referencia a las observaciones de todas las variables explicativas del modelo, entonces se tiene que:

$$E[Y_i \mid X_i] = E[F(X_i\beta)] + E[u_i] = F(X_i\beta),$$

manteniendo el supuesto de que $E[u_i] = 0$.

Por otra parte, si se calcula la esperanza condicionada de *Y* en términos probabilísticos, entonces:

$$E[Y_i \mid X_i] = \sum_i Y_i \cdot P(Y_i \mid X_i) = 1 \cdot P(Y_i = 1 \mid X_i) + 0 \cdot P(Y_i = 0 \mid X_i) = P(Y_i = 1 \mid X_i).$$

De donde se deduce, pues, que:

$$E[Y_i \mid X_i] = F(X_i\beta) = P(Y_i = 1 \mid X_i).$$

Teniendo en cuenta que la variable Y_i únicamente puede tomar los valores 1 y 0, el significado del modelo implica que éste asigna cierta probabilidad condicional de que $Y_i = 1$, que denotaremos por P_i, es decir:

$$P(Y_i = 1 \mid X_i) = P_i = F(X_i\beta);$$

y, en consecuencia:

$$P(Y_i = 0 \mid X_i) = 1 - P_i = 1 - F(X_i\beta).$$

El modelo estima, por tanto, la probabilidad para la observación *i* de elegir la opción 1, ya que:

$$\boxed{\hat{Y}_i = F(X_i\hat{\beta}) = \hat{P}_i}.$$

Dependiendo de la forma funcional concreta que adopte $F(X_i\beta)$, se obtienen distintos modelos de elección binaria. Veremos los principales.

5.2. Modelo lineal de probabilidad.-

El *modelo lineal de probabilidad* (MLP) es un modelo de respuesta o elección binaria caracterizado porque $F(X_i\beta)$ adopta la forma de una función lineal, de modo que:

$$F(X_i\beta) = X_i\beta.$$

Así pues:

$$\boxed{Y_i = F(X_i\beta) + u_i = X_i\beta + u_i}.$$

 ISBN: 978-84-694-7251-4

De este modo: $E[Y_i \mid X_i] = F(X_i\beta) = X_i\beta = 1 \cdot P_i + 0 \cdot (1 - P_i) = P_i$, por lo que: $E[Y_i \mid X_i] = P_i = X_i\beta$.

En definitiva, la probabilidad de que ocurra el suceso objeto de estudio, se puede estimar a partir de un ajuste de regresión lineal entre la variable dependiente *Y* y las independientes *X*:

$$\boxed{\hat{Y}_i = \hat{P}_i = X_i\hat{\beta}};$$

es decir, los valores estimados del modelo $\hat{Y}_i$, representan las estimaciones de la probabilidad de que $Y_i = 1$.

En cuanto a la interpretación de los valores estimados de los parámetros o coeficientes de regresión del MLP, éstos miden el efecto, en términos de probabilidad de elección de la alternativa estudiada, de un cambio unitario en cada una de las variables explicativas X_j (j = 1, 2,…, k):

$$\boxed{\frac{\partial P_i}{\partial X_{ji}} = \frac{\partial F(X_i\beta)}{\partial X_{ji}} = \beta_j}.$$

En el caso de que X_j sea una variable ficticia o *dummy*, entonces el efecto de una variación de dicha variable sobre la probabilidad de que Y tome el valor 1 se calcula a través de la diferencia entre los valores obtenidos por $E[Y_i \mid X_{ji} = 1]$ y $E[Y_i \mid X_{ji} = 0]$.

Pese a la facilidad de planteamiento de este modelo, presenta importantes limitaciones. Éstas son:

- No normalidad de la perturbación aleatoria; en efecto: para $Y_i = 1$ tenemos que $u_i = 1 - X_i\beta = 1 - P_i$, y para $Y_i = 0$, $u_i = 0 - X_i\beta = -P_i$, por lo que $-1 \le u_i \le 1$, cuando el rango de variación de una variable aleatoria normal es $(-\infty, +\infty)$. En concreto, u_i sigue una distribución binomial.
- Escasa fiabilidad del coeficiente de determinación R^2: sus valores suelen ser muy bajos debido a lo elevadas que son las sumas cuadráticas de los residuos; téngase en cuenta que lo que se trata es de ajustar una función lineal a observaciones que adoptan dos únicos valores extremos: 1 y 0 (*Figura 1*).
- Estimaciones de la probabilidad no acotadas: el MLP no garantiza matemáticamente que los valores que se estimen, $\hat{Y}_i = \hat{P}_i$, estén comprendidos entre 0 y 1, lo que contradice gravemente el concepto teórico de probabilidad. Éste es uno de los principales problemas de este modelo.
- Planteamiento poco realista sobre el comportamiento de la probabilidad; ésta es otra grave limitación. El hecho de suponer que la probabilidad aumenta

linealmente con los valores de *X*, es decir, que el efecto marginal de *X* se mantiene constante a lo largo de todas las observaciones, resulta difícil de admitir empíricamente[2].

- Heteroscedasticidad de la perturbación aleatoria; aún cuando se puedan mantener en este modelo las hipótesis de que la perturbación aleatoria tiene media 0 y que es serialmente independiente, si acudimos a la definición de su varianza, veremos que ésta tiene un valor distinto para cada observación *i-ésima*.

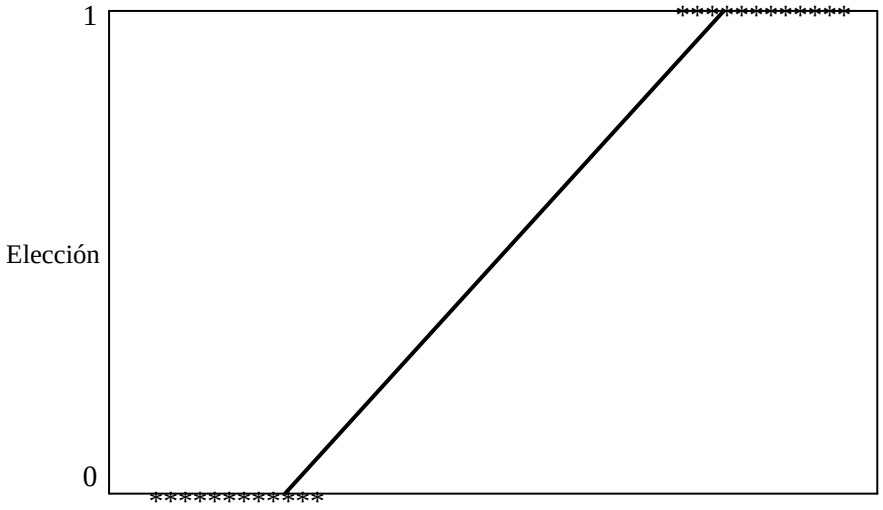


Figura 1

Debido al problema de heteroscedasticidad de este modelo, la aplicación de MCO daría lugar a estimadores que no serían ELIO, porque si bien serían lineales e insesgados, no tendrían mínima varianza. El método de estimación más adecuado sería entonces el de los mínimos cuadrados generalizados (MCG)[3].

Pero aunque la estimación por el método de MCG permite obtener estimadores eficientes del modelo, los problemas persisten:

- Los estimadores son eficientes, pero si se eliminan aquellas estimaciones incoherentes, entonces ya no son robustos.

[2] Si se piensa por ejemplo en la adquisición de una póliza privada de enfermedad por parte de los hogares en función de su renta, parece evidente que a niveles bajos, la probabilidad irá creciendo lentamente, ya que se carece de posibilidades; a partir de un cierto momento comenzará a subir más rápidamente; y, de nuevo, en niveles de probabilidad cercanos a 1, correspondientes a altos niveles de renta, el efecto marginal será de nuevo menor, pues muchos hogares dispondrán ya de dicha cobertura.

[3] Como ya sabemos, si bien el modo más correcto de actuar sería estimar el modelo por el método de MCG, una opción "intermedia", fácil y más eficiente que MCO, sería aplicar la *estimación consistente de White*. Recuérdese que, mediante este método, la estimación de los coeficientes de regresión del modelo es la misma que por MCO, pero sin embargo, la matriz de varianzas-covarianzas de éstos se estima correctamente, lo que repercute en una mayor fiabilidad de los contrastes de hipótesis que se planteen.

 ISBN: 978-84-694-7251-4

- Con la transformación realizada, el modelo pierde su término independiente, lo que puede originar problemas sobre R^2, ya de por sí subestimado, pues puede llegar a tener valor negativo.
- La perturbación aleatoria continúa sin seguir una distribución normal, lo que invalida todos los desarrollos inferenciales establecidos por la Teoría Econométrica para los modelos clásicos de regresión lineal. No obstante, este problema se puede solventar empleando muestras con un tamaño "suficientemente" grande. De acuerdo con el Teorema Central del Límite, la distribución binomial en el límite se comporta como una distribución normal.
- Pero los principales problemas del MLP permanecen: la adopción del supuesto lineal y que no hay nada que asegure que las estimaciones de la probabilidad estén entre 0 y 1.

Está claro, pues, que la solución pasa por encontrar algún tipo de función $F(X)$ que en lugar de ser lineal, tenga forma de "*S*" (véase la *Figura 2*). Es aquí cuando surgen entonces modelos como el *logit* o el *probit*.

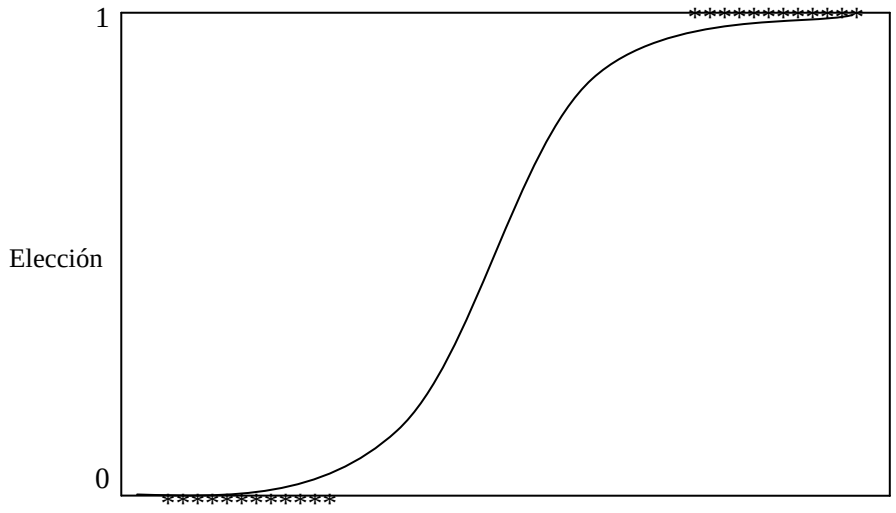


Figura 2

5.3. Modelo logit y modelo probit.-

Los modelos *logit* y *probit* son modelos de elección binaria no lineales muy semejantes. Presentan una serie de propiedades que justifican su utilización:

- Se definen para todo valor de las variables explicativas: $(-\infty,+\infty)$.
- Sus funciones de distribución son continuas y toman siempre valores comprendidos entre 0 y 1.

- Cuando $X_i\beta \to -\infty \Rightarrow P_i \to 0$.
- Cuando $X_i\beta \to +\infty \Rightarrow P_i \to 1$.
- Incrementan monótonamente respecto a $X_i\beta$.

La diferencia básica entre ellos radica en la función de distribución de probabilidad sobre la que se basan, $F(.)$. En concreto, el *logit* emplea la función logística de distribución: $\Lambda(.)$ y el *probit*, la correspondiente a la distribución normal: $\Phi(.)$. Estas funciones de distribución se caracterizan por tener forma de "*S*"; su punto de inflexión depende de cada una de ellas.

Definición del modelo logit

El modelo *logit* es un modelo de respuesta o elección binaria, caracterizado porque la función $F(X_i\beta)$ se corresponde con la de la distribución logística, de modo que si denotamos ésta por $\Lambda(.)$, tenemos que:

$$\boxed{Y_i = \Lambda(X_i\beta) + u_i}.$$

De esta forma:

$$E\left[Y_i \mid X_i\right] = P\left(Y_i = 1 \mid X_i\right) = P_i = \Lambda(X_i\beta) = \frac{1}{1+e^{-X_i\beta}}$$

y

$$P\left(Y_i = 0 \mid X_i\right) = 1 - P_i = 1 - \frac{1}{1+e^{-X_i\beta}} = \frac{1}{1+e^{X_i\beta}}.$$

Así pues, la estimación del modelo proporciona la cuantificación de la probabilidad de elegir la opción 1; es decir:

$$\boxed{\hat{Y}_i = \hat{P}_i = \Lambda(X_i\hat{\beta}) = \frac{1}{1+e^{-X_i\hat{\beta}}}}.$$

La interpretación de los parámetros β_j en un modelo *logit* no es inmediata, ya que no nos proporciona, como en el caso de los modelos lineales, el efecto marginal de un cambio unitario en X_j sobre la probabilidad de elegir la alternativa 1 (la variable endógena *Y*).

En efecto, dada la no linealidad del modelo, se tiene que:

$$\frac{P_i}{1-P_i} = \frac{\dfrac{1}{1+e^{-X_i\beta}}}{\dfrac{1}{1+e^{X_i\beta}}} = e^{X_i\beta},$$

de donde:

 ISBN: 978-84-694-7251-4

$$X_i\beta = \ln\left(\frac{P_i}{1-P_i}\right) = L_i .$$

L_i recibe el nombre de *logit*, y es lo que da nombre al modelo. El *logit* representa el logaritmo neperiano de la razón de la probabilidad a favor de la alternativa 1. Por tanto:

$$\boxed{\beta_j = \frac{\partial L_i}{\partial X_{ji}}}.$$

Es decir, los parámetros miden el cambio en el *logit* ocasionado por un cambio unitario en la variable X_j; esto es, cuánto varía el logaritmo de la razón de probabilidades a favor de la ocurrencia de la opción 1, ante incrementos unitarios de X_j , no el efecto marginal de un cambio unitario en X_j sobre la probabilidad de ocurrencia de la opción 1, P_i.

Éste último viene dado por la expresión:

$$\boxed{\frac{\partial P_i}{\partial X_{ji}} = \frac{\partial \Lambda(X_i\beta)}{\partial X_{ji}} = \lambda(X_i\beta)\cdot\beta_j},$$

donde $\lambda(X_i\beta)$ es la función de densidad de la distribución logística.

Esto es, la variación en la probabilidad de la ocurrencia de la opción estudiada ante variaciones unitarias de X_j viene dada por el producto de β_j por el valor que toma la función de densidad de la distribución logística en la observación *i-ésima*.

Este último detalle es importante. La magnitud de la variación de la probabilidad, dado un incremento unitario de la correspondiente variable explicativa, depende de su nivel de partida y, por consiguiente, de los valores de todos y cada uno de los regresores y coeficientes en la observación donde se estudie. Esto supone, como puede observarse, la superación de la limitación que ofrecía el MLP respecto a la consideración de un efecto marginal de las variables explicativas sobre la probabilidad constante para todas las observaciones; supuesto que considerábamos poco ajustado a la realidad.

En el caso de que X_j sea una variable *dummy*, el análisis del efecto marginal de ésta sobre la probabilidad se calcula a través de la diferencia de los valores proporcionados por $E\left[Y_i \mid X_{ji} = 1\right]$ y $E\left[Y_i \mid X_{ji} = 0\right]$.

Como el efecto marginal de un cambio unitario de X_j sobre la probabilidad varía dependiendo de la observación donde se realice, para obtener un valor representativo éste se suele medir en los valores medios de los regresores.

Es importante reseñar que el signo de β_j sí indica el sentido del cambio en la probabilidad, aunque no su cuantía.

La estimación de este modelo se realiza habitualmente a través del método de máxima verosimilitud (MV). Los estimadores MV resultan ser consistentes y asintóticamente eficientes, por lo que este método es recomendable que se emplee para muestras "suficientemente" grandes.

Definición del modelo probit

Otro de los modelos de elección binaria es el modelo *probit*. Este modelo se caracteriza porque la función de distribución que utiliza es la correspondiente a la normal: $\Phi(.)$. De este modo, se tiene que:

$$\boxed{Y_i = \Phi(X_i\beta) + \varepsilon_i}.$$

Y consiguientemente:

$$E[Y_i \mid X_i] = P(Y_i = 1 \mid X_i) = P_i = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \phi(s)\, ds\,,$$

donde $\phi(s) = \dfrac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}}$ es la función de densidad de la distribución normal y s es una variable "muda" de integración con media cero y varianza 1.

Así que, la estimación del modelo ofrece la cuantificación de la probabilidad de elegir la alternativa 1; esto es:

$$\boxed{\hat{Y}_i = \hat{P}_i = \Phi(X_i\hat{\beta}) = \int_{-\infty}^{X_i\hat{\beta}} \frac{1}{(2\pi)^{1/2}} e^{\frac{-s^2}{2}}\, ds}.$$

En cuanto a la interpretación de los parámetros, en el modelo *probit* sucede exactamente igual que en el modelo *logit*. Los parámetros estimados no determinan el efecto marginal de variaciones de las variables explicativas X_j sobre la probabilidad, aunque su signo sí determina el sentido del cambio. El efecto marginal resulta del producto del valor de la función de densidad del modelo (en este caso la normal) en un punto determinado y el parámetro correspondiente:

$$\boxed{\frac{\partial P_i}{\partial X_{ji}} = \frac{\partial \Phi(X_i\beta)}{\partial X_{ji}} = \phi(X_i\beta)\cdot\beta_j}.$$

Como se puede observar, la magnitud de las variaciones de la probabilidad depende del nivel donde se observe, por lo que (de forma análoga a como sucede en el modelo *logit*) es función de los valores de todas y cada una de las variables explicativas y de sus coeficientes en aquella observación donde se estudie.

 ISBN: 978-84-694-7251-4

Si X_j fuese una variable ficticia o *dummy*, el análisis de su efecto marginal sobre la probabilidad, se realiza a través de la diferencia de los valores proporcionados por $E\left[Y_i \mid X_{ji}=1\right]$ y $E\left[Y_i \mid X_{ji}=0\right]$.

De nuevo, para obtener un valor representativo de los efectos marginales éstos se suelen medir en los valores medios de los regresores.

Un aspecto más que se puede considerar, a la hora de interpretar el significado de estos modelos, es el relativo a los denominados *odds* y *ratio odds*.

El estadístico *odds* mide el cociente de probabilidades para una observación *i* de elegir la opción 1 frente a la opción 0; es decir:[4]

$$Odds = \frac{P_i}{1-P_i}.$$

Si lo que se quiere es comparar la utilidad que la opción elegida proporciona al individuo (observación) *i*, con respecto a la utilidad percibida por el individuo (observación) *m*, entonces se define el *cociente* o *ratio odds*:

$$Ratio\,odds = \frac{\dfrac{P_i}{1-P_i}}{\dfrac{P_m}{1-P_m}}.$$

De este modo, si la *ratio odds* es:

- mayor que 1: la utilidad para el individuo *i* es mayor que para el individuo *m*;
- menor que 1: la utilidad para el individuo *i* es menor que para el individuo *m*;
- igual a 1: la utilidad para ambos individuos, *i* y *m*, es la misma.

Al igual que el modelo logit, el modelo probit suele estimarse por el método de MV.

- ***Contraste de significatividad individual***

La estimación de los parámetros de estos modelos por MV resulta asintóticamente normal, por lo que los contrastes resultan asintóticos. Éstos se realizan de forma similar a como se hace en el modelo de regresión lineal, con la diferencia de que en lugar de emplear estadísticos que siguen la distribución *t-Student*, en este caso se trabaja con la distribución *normal tipificada*.

De este modo, para contrastar la hipótesis nula $H_0 : \beta_j = 0$, se tendrá que:

[4] Obsérvese cómo lo que se definió como *logit*, no es más que el logaritmo del estadístico *odds*.

$$P\left(-N_{\alpha/2} < \frac{\hat{\beta}_j}{\widehat{ES(\hat{\beta}_j)}} < N_{\alpha/2}\right) = 1-\alpha,$$

donde para un nivel de confianza $1-\alpha$, $N_{\alpha/2}$ y $-N_{\alpha/2}$ son los valores críticos (simétricos) de la distribución normal tipificada, $\hat{\beta}_j$ se refiere a cada uno de los parámetros estimados, y $\widehat{ES(\hat{\beta}_j)}$ es la estimación del error estándar del parámetro estimado correspondiente.

Como bien sabemos, si la desigualdad se verifica, ello supondrá aceptar la hipótesis nula, con lo que el parámetro β_j, y con ello la variable explicativa X_j, no serán significativos. Si por el contrario, no se verificase, querrá decir que la variable en cuestión es relevante en la explicación de la variable dependiente del modelo.

- ***Medidas de bondad del ajuste y de contraste de significatividad global de los modelos de elección discreta binaria***

En estos modelos, el habitual coeficiente de determinación R^2 no resulta válido como medida de bondad del ajuste. Por ello, se han desarrollado otras medidas alternativas.

- *R^2 de McFadden*:

Su expresión es:

$$R^2 \text{ de McFadden} = 1 - \frac{\ln(L)}{\ln(L_R)},$$

donde $\ln(L_R)$ es el logaritmo neperiano de la función de verosimilitud del modelo restringido, bajo la hipótesis nula $H_0 : \beta_2 = \ldots = \beta_k = 0$, mientras que $\ln(L)$ es el logaritmo de la función de verosimilitud del modelo original (sin restricciones).

El valor de este estadístico está comprendido entre 0, que implica nula significatividad del modelo, y 1, que supone un ajuste perfecto, pero para los valores intermedios su significado no es tan claro ni directo como en el caso del R^2 lineal.

- *Estadístico LR* (*LR-statistic*) o *test de la razón de verosimilitud*:

El estadístico *LR* se define de la forma:

$$LR = -2\frac{\ln(L_R)}{\ln(L)},$$

donde, al igual que en el caso del R^2 *de McFadden*, $\ln(L_R)$ es el logaritmo neperiano de la función de verosimilitud del modelo restringido, bajo la hipótesis nula $H_0 : \beta_2 = \ldots = \beta_k = 0$ y $\ln(L)$ es el logaritmo de la función de verosimilitud del modelo original (sin restricciones).

 ISBN: 978-84-694-7251-4

Este estadístico sigue una distribución χ^2 con $k-1$ grados de libertad.

El rechazo de H_0 implica que el modelo es, en su conjunto, significativo. Por el contrario, su aceptación implica que no lo es.

- *Pseudo R^2 de predicción*:

Esta medida indica la proporción de predicciones correctas que realiza el modelo. Concretamente, se define como:

$$Pseudo\, R^2\, de\, predicción = \frac{\text{Predicciones correctas}}{n},$$

donde *n* es el número de observaciones muestrales.

Normalmente, el valor umbral que se suele adoptar para asignarle un valor a una predicción es 0,5; de tal manera que:

- si $\hat{Y}_i \geq 0{,}5$, entonces se asigna a la predicción el valor 1;
- y si $\hat{Y}_i < 0{,}5$, se asigna a la predicción el valor 0.

A partir de aquí, se cuentan entonces los valores 1 y 0 asignados, comparándolos con los de las observaciones reales, para calcular así el Pseudo R^2 de predicción.

- ***Comparación y elección entre modelos de elección discreta binaria***

Los siguientes estadísticos que vamos a tratar se refieren a "pérdida de información". De algún modo sirven también para evaluar la bondad de un modelo, por cuanto se utilizan, en base a sus valores, para comparar las estimaciones realizadas por distintas modelizaciones y seleccionar de este modo la "mejor".

- *Akaike Information Criterion* (*AIC*): $AIC = \dfrac{2k - 2\ln(L)}{n}$

- *Schwarz Criterion* (*SC*): $SC = \dfrac{k \cdot \ln(n) - 2\ln(L)}{n}$

- *Hannan-Quinn Information Criterion* (*HQ*): $HQ = \dfrac{2k \cdot \ln\left(\ln(n)\right) - 2\ln(L)}{n}$

donde, para los tres estadísticos:
- *k*: número de regresores (incluido el término independiente);
- *n*: tamaño de la muestra;
- *L*: valor de la función de verosimilitud.

A la hora de establecer comparaciones entre distintos modelos, se considera con mejor ajuste aquél que presente unos valores más bajos en estos estadísticos. Además de estos

 ISBN: 978-84-694-7251-4

estadísticos, también se podría emplear el *estadístico de la razón de verosimilitud* (*LR-statistic*), ya visto anteriormente.

5.4. Estimación de modelos de elección discreta binaria con *EViews*.-

En este punto vamos a mostrar un ejemplo de especificación y estimación de un modelo de elección discreta binaria. En concreto, vamos a estudiar la demanda de compra de un seguro privado de enfermedad por parte de los hogares, en función de diversas características de la persona principal del hogar: edad y nivel máximo de estudios alcanzado, así como también de los ingresos netos familiares. La información procede de una muestra de 3.000 hogares españoles de 1998. Las variables del modelo son:

$$\text{SEGPRIV} = \begin{cases} 1, & \text{si el hogar posee seguro privado de enfermedad} \\ 0, & \text{en caso contrario} \end{cases}$$

EDAD = Edad de la persona principal del hogar (en años)

$$\text{SECUNDAR} = \begin{cases} 1, & \text{si los máximos estudios de la persona ppal. del hogar son secundarios} \\ 0, & \text{en caso contrario} \end{cases}$$

$$\text{SUPERIOR} = \begin{cases} 1, & \text{si los máximos estudios de la persona principal del hogar son superiores} \\ 0, & \text{en caso contrario} \end{cases}$$

LNING = Logaritmo de los ingresos netos familiares

Obsérvese que, en el nivel de estudios, se considera el PRIMARIO como su categoría base.

La variable a explicar de nuestro modelo es una variable dicotómica, que únicamente toma dos posibles valores, 1 ó 0, según si el hogar dispone o no de seguro privado de enfermedad. Por tanto, el objeto de estudio de nuestro modelo es estimar la probabilidad de que el hogar *i* elija la opción 1: $\hat{Y}_i = F(X_i\hat{\beta}) = \hat{P}_i$; es decir, estimar la probabilidad de que un hogar disponga de seguro privado de enfermedad. Respecto a las variables explicativas, entre ellas hay tanto cuantitativas (edad y logaritmo de los ingresos), como cualitativas (nivel de estudios), y lo que haremos es ver si realmente son significativas o no en la decisión de compra del mencionado seguro, así como determinar cuál es su aportación en términos cuantitativos.

A partir de aquí, podemos optar por diversas formas funcionales $F(X_i\beta)$ para especificar nuestro modelo. En particular, se han visto de forma teórica tres: el modelo lineal de probabilidad (MLP), el modelo logit y el modelo probit:

 ISBN: 978-84-694-7251-4

$$\text{MLP} \;-\; Y_i = X_i\beta + u_i \longrightarrow \hat{Y}_i = \hat{P}_i = X_i\hat{\beta}$$

$$\text{Logit} \;-\; Y_i = \Lambda(X_i\beta) + u_i \longrightarrow \hat{Y}_i = \hat{P}_i = \Lambda(X_i\hat{\beta}) = \frac{1}{1+e^{-X_i\hat{\beta}}}$$

$$\text{Probit} \;-\; Y_i = \Phi(X_i\beta) + u_i \longrightarrow \hat{Y}_i = \hat{P}_i = \Phi(X_i\hat{\beta}) = \int_{-\infty}^{X_i\hat{\beta}} \frac{1}{(2\pi)^{1/2}} e^{\frac{-s^2}{2}}\, ds$$

La *Figura 3* muestra el cuadro de diálogo de *EViews* para llevar a cabo la estimación del MLP. Obsérvese que en "*Options*" se elige la estimación consistente de White (*Heteroskedasticity Consistent Covariance*).

Las *Figuras 4* y *5*, por su parte, nos ofrecen los pasos necesarios que deben seguirse en *EViews* para estimar los modelos logit y probit.

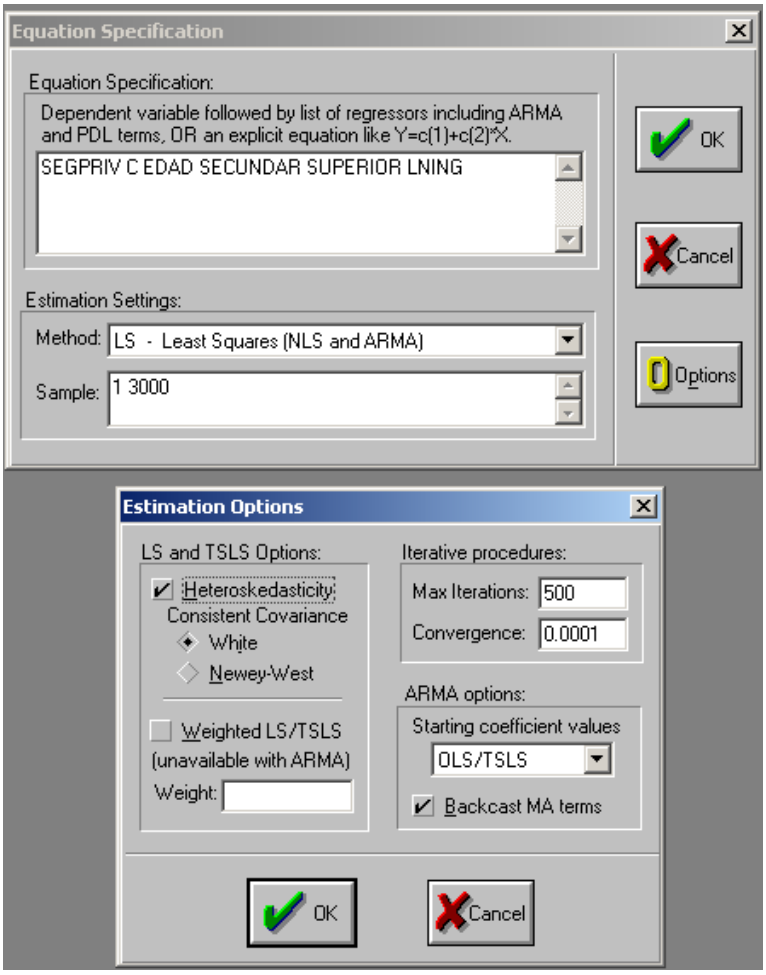


Figura 3

 ISBN: 978-84-694-7251-4

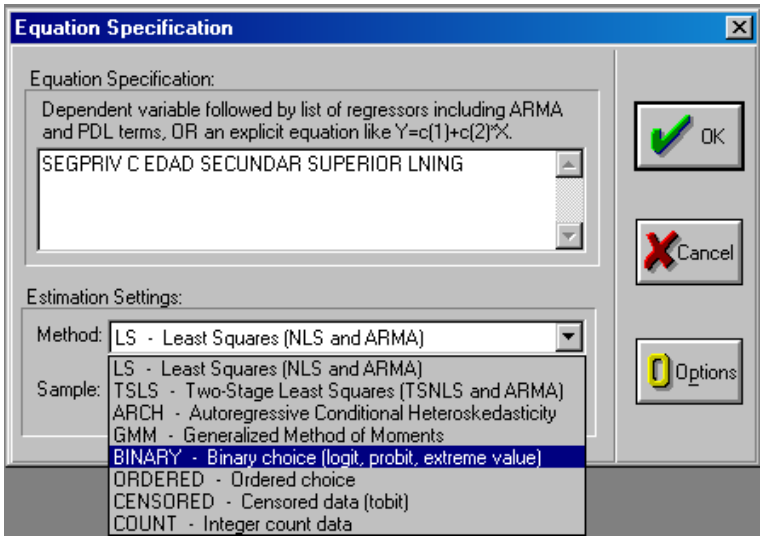


Figura 4

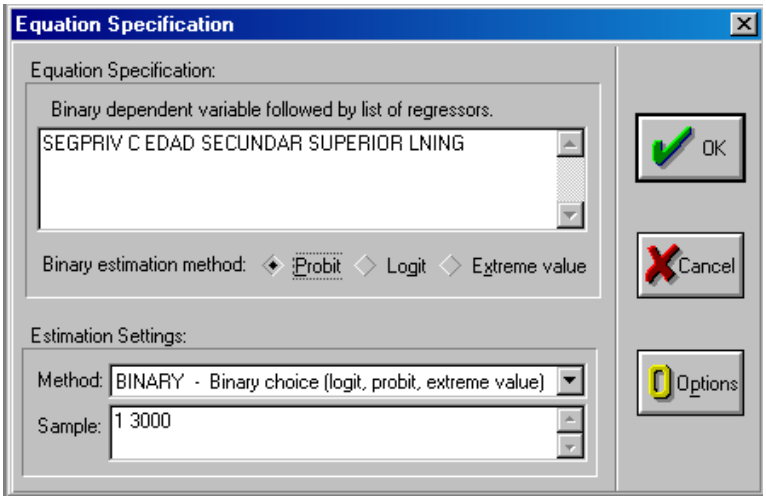
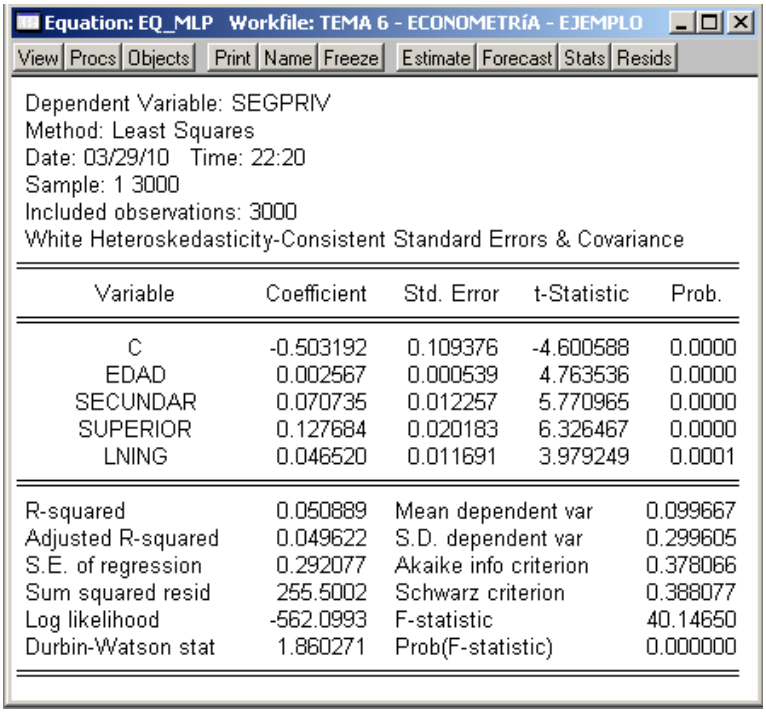


Figura 5

En las *Figuras 6*, *7* y *8* se muestran, respectivamente, las regresiones estimadas de cada una de las tres posibilidades señaladas.

Lo primero que puede afirmarse es que para las tres estimaciones realizadas, todas las variables explicativas resultan significativas, como indican los *p-values* asociados a los estadísticos *t*-Student (en el caso del MLP) o *z*-normales (para el logit y el probit) de cada uno de los respectivos coeficientes de regresión.

 ISBN: 978-84-694-7251-4



Equation: EQ_MLP Workfile: TEMA 6 - ECONOMETRÍA - EJEMPLO

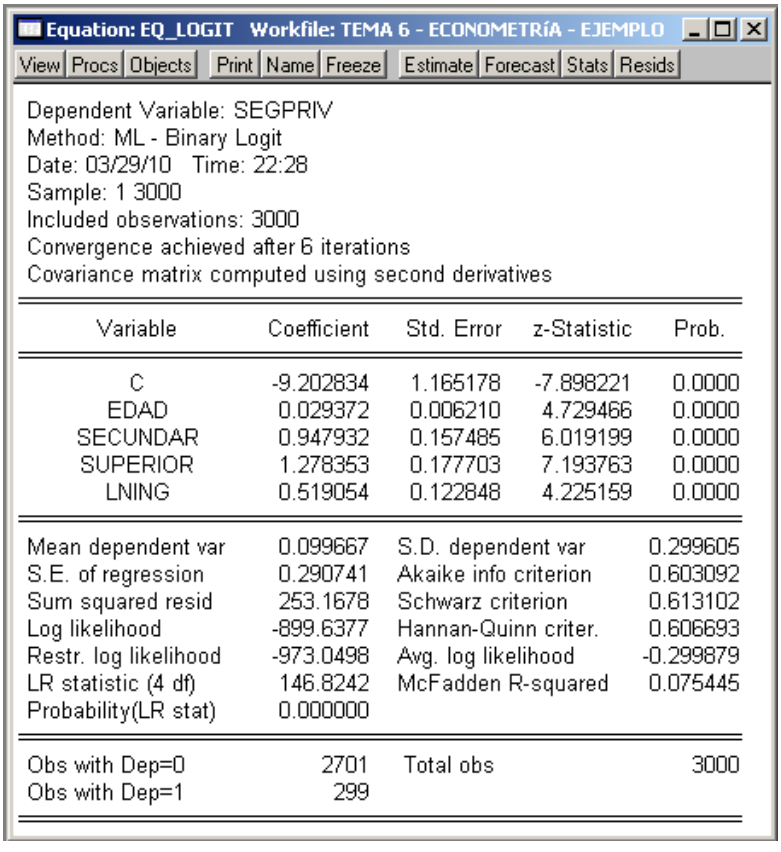
View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SEGPRIV
Method: Least Squares
Date: 03/29/10 Time: 22:20
Sample: 1 3000
Included observations: 3000
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.503192	0.109376	-4.600588	0.0000
EDAD	0.002567	0.000539	4.763536	0.0000
SECUNDAR	0.070735	0.012257	5.770965	0.0000
SUPERIOR	0.127684	0.020183	6.326467	0.0000
LNING	0.046520	0.011691	3.979249	0.0001

R-squared	0.050889	Mean dependent var	0.099667
Adjusted R-squared	0.049622	S.D. dependent var	0.299605
S.E. of regression	0.292077	Akaike info criterion	0.378066
Sum squared resid	255.5002	Schwarz criterion	0.388077
Log likelihood	-562.0993	F-statistic	40.14650
Durbin-Watson stat	1.860271	Prob(F-statistic)	0.000000

Figura 6



Equation: EQ_LOGIT Workfile: TEMA 6 - ECONOMETRÍA - EJEMPLO

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SEGPRIV
Method: ML - Binary Logit
Date: 03/29/10 Time: 22:28
Sample: 1 3000
Included observations: 3000
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

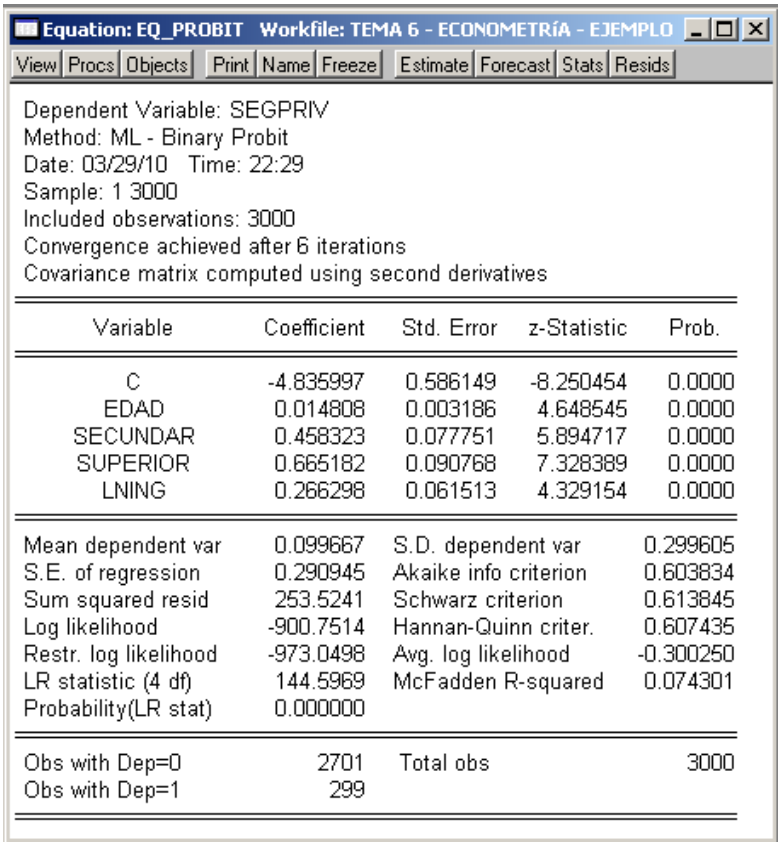
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-9.202834	1.165178	-7.898221	0.0000
EDAD	0.029372	0.006210	4.729466	0.0000
SECUNDAR	0.947932	0.157485	6.019199	0.0000
SUPERIOR	1.278353	0.177703	7.193763	0.0000
LNING	0.519054	0.122848	4.225159	0.0000

Mean dependent var	0.099667	S.D. dependent var	0.299605
S.E. of regression	0.290741	Akaike info criterion	0.603092
Sum squared resid	253.1678	Schwarz criterion	0.613102
Log likelihood	-899.6377	Hannan-Quinn criter.	0.606693
Restr. log likelihood	-973.0498	Avg. log likelihood	-0.299879
LR statistic (4 df)	146.8242	McFadden R-squared	0.075445
Probability(LR stat)	0.000000		

Obs with Dep=0	2701	Total obs	3000
Obs with Dep=1	299		

Figura 7

 ISBN: 978-84-694-7251-4



Equation: EQ_PROBIT Workfile: TEMA 6 - ECONOMETRíA - EJEMPLO

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SEGPRIV
Method: ML - Binary Probit
Date: 03/29/10 Time: 22:29
Sample: 1 3000
Included observations: 3000
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-4.835997	0.586149	-8.250454	0.0000
EDAD	0.014808	0.003186	4.648545	0.0000
SECUNDAR	0.458323	0.077751	5.894717	0.0000
SUPERIOR	0.665182	0.090768	7.328389	0.0000
LNING	0.266298	0.061513	4.329154	0.0000

Mean dependent var	0.099667	S.D. dependent var	0.299605
S.E. of regression	0.290945	Akaike info criterion	0.603834
Sum squared resid	253.5241	Schwarz criterion	0.613845
Log likelihood	-900.7514	Hannan-Quinn criter.	0.607435
Restr. log likelihood	-973.0498	Avg. log likelihood	-0.300250
LR statistic (4 df)	144.5969	McFadden R-squared	0.074301
Probability(LR stat)	0.000000		

Obs with Dep=0	2701	Total obs	3000
Obs with Dep=1	299		

Figura 8

Si nos centramos en la salida del MLP, puede comprobarse el bajo valor reflejado por el coeficiente de determinación lineal R^2, ya comentado a nivel teórico.

Como ya se ha indicado, conceptualmente los modelos logit y probit resultan más apropiados que el MLP. A la hora de elegir entre uno u otro, podemos atender a los resultados arrojados por medidas como el R^2 de McFadden (*McFadden R-squared*), el estadístico *LR* o razón de verosimilitud (*LR-statistic*), o los estadísticos de Akaike, Schwarz y Hannan-Quinn de pérdida de información. Al observar los valores de todos estos indicadores en nuestras estimaciones de los modelos logit y probit, puede comprobarse que, de acuerdo con todos ellos, el modelo logit parece (aunque por muy poco) más adecuado.

En el siguiente Cuadro se ofrece, para los tres modelos estimados, los efectos marginales de cada variable explicativa sobre la probabilidad de tenencia de seguro privado de enfermedad (junto con los valores medios de cada variable, necesarios para el cálculo de dichos efectos en los modelos logit y probit[5]).

[5] Debe indicarse que los efectos marginales de los modelos logit y probit no son proporcionados por *EViews*, sino que se han calculado aparte, tomando no obstante la información básica precisa para ello de los resultados de *EViews*.

 ISBN: 978-84-694-7251-4

Variables	Valores medios	Efectos marginales MLP	Efectos marginales Logit	Efectos marginales Probit
C	1	-0,5032	-0,6916	-0,7512
EDAD	42,17733	0,0026	0,0022	0,0023
SECUNDAR	0,331667	0,0707	0,0694	0,0684
SUPERIOR	0,168667	0,1277	0,1087	0,1146
LNING	9,664755	0,0465	0,0390	0,0414

Para explicar el significado de los efectos marginales derivados de los coeficientes de regresión de estos modelos, vamos a basarnos, a modo de ejemplo, en las estimaciones obtenidas a partir del modelo logit, ya que éste parece la elección más adecuada. En los otros, el significado es análogo.

- C - Como siempre, aquí se hace referencia a la ordenada en el origen y, a veces, no tiene sentido económico. En este caso, así ocurre.
- EDAD – El efecto marginal de esta variable indica que por cada año adicional que tenga la persona principal del hogar, la probabilidad de que se éste disponga de un seguro privado de enfermedad se incrementa en un 0,22%.
- SECUNDAR - El efecto marginal asociado al nivel de estudios secundarios de la persona principal del hogar muestra una relación de este nivel de estudios, con la probabilidad de que el hogar posea un seguro privado de enfermedad, superior a la asociada a la categoría base. En concreto, un 6,94% mayor que en el caso de los hogares cuya persona principal tiene sólo estudios primarios.
- SUPERIOR - El significado de este efecto es similar al anterior. Los hogares donde la persona principal posee estudios superiores evidencian un 10,87% más de probabilidad de tener un seguro privado de enfermedad que aquéllos donde la persona principal tiene únicamente estudios primarios.
- LNING - El efecto marginal correspondiente al logaritmo de los ingresos netos familiares indica que la relación entre esta variable y la probabilidad de que el hogar tenga seguro privado de enfermedad es positiva. Dado que la variable ingresos viene medida en logaritmos, este coeficiente de regresión coincide con la elasticidad renta de este bien, de forma que un incremento de un 1% en los ingresos, conlleva un aumento de la referida probabilidad en un 0,0390%.