CONCEPTOS BÁSICOS SOBRE LA HETEROCEDASTICIDAD EN EL MODELO BÁSICO DE REGRESIÓN LINEAL

TRATAMIENTO CON E-VIEWS

Prof. Rafael de Arce
Dpto. de Economía Aplicada
Universidad Autónoma de Madrid
rafael.dearce@uam.es

Abril de 2001

ÍNDICE DE CONTENIDOS

1.- Qué es

La heterocedasticidad es la existencia de una varianza no constante en las perturbaciones aleatorias de un modelo econométrico. En ese caso, la matriz de varianzas-covarianzas de las perturbaciones se representaría del siguiente modo:

$$E(UU') = \begin{bmatrix} E(u_1)^2 & & & . \\ E(u_1u_2) & E(u_2)^2 & & \\ & & \dots & \\ E(u_1u_n) & E(u_2u_n) & & E(u_n)^2 \end{bmatrix} = \begin{bmatrix} E(u_1)^2 & & & . \\ 0 & E(u_2)^2 & & \\ 0 & 0 & \dots & \\ 0 & 0 & 0 & E(u_n)^2 \end{bmatrix} = s_i^2 I_n = s^2 \Sigma$$

Como caso concreto de la presencia de una matriz de varianzas-covarianzas no escalar de las perturbaciones aleatorias, la estimación correcta de los parámetros del modelo debiera hacerse a partir del método de los Mínimos Cuadrados Generalizados:

$$b^{MCG} = \left[X'\Sigma^{-1}X\right]^{-1} X'\Sigma^{-1}Y$$

donde, para la aplicación de esta fórmula en un modelo de "n" observaciones y "k" variables explicativas, sería necesario estimar "k" parámetros sobre las variables y "n" varianzas distintas de las perturbaciones aleatorias.

Por supuesto, como solo contamos con "n" observaciones muestrales, es imposible estimar simultáneamente n+k valores: hay más incógnitas que ecuaciones independientes se puedan construir. Por ello, habrá que hacer algún supuesto simplificador sobre la causa de la heterocedasticidad una vez esta sea detectada. Evidentemente, encontrar una simplificación correcta dotará de plena utilidad (eficiencia) a la estimación con MCG y, a sensu contrario, un mal diseño de la causa de la heterocedasticidad (de la matriz S) producirá un valor ineficiente de dichos parámetros.

2.- Causas frecuentes de heterocedasticidad

Aunque las que se citan a continuación no son las únicas posibilidades que dan lugar a un modelo heterocedástico, sí son las más frecuentes.

A.- Variables explicativas cuyo recorrido tenga una gran dispersión respecto a su propia media.

En esta situación, los modelos de corte transversal son especialmente susceptibles a registrar heterocedasticidad. La disposición arbitraria de las observaciones en este caso (puede responder, por ejemplo al orden alfabético de las observaciones de la endógena o al modo en que se han obtenido los datos o a cualquier otra razón) pueden agrupar, casualmente, observaciones que presenten valores grandes en una determinada variable explicativa y lo mismo con valores pequeños de esta misma variable. Si esta variable es la que está produciendo la distorsión en el modelo de heterocedasticidad, dicha distorsión será probablemente mayor en aquellas observaciones que contengan una mayor carga de ésta y menor en las que su peso sea más pequeño. Por ello, la varianza de las perturbaciones aleatorias estimada por subperíodos distintos de la muestra sería diferente; es decir, habría heterocedasticidad.

La misma situación se puede dar en modelos de corte temporal en los que la evolución histórica haya marcado diferentes períodos en cuanto a los valores de una variable en relación a su media, agrupando en algún subperíodo valores altos y en otros valores pequeños.

B.- Omisión de variables relevantes en el modelo especificado.

Evidentemente, cuando se ha omitido una variable en la especificación, dicha variable quedará parcialmente recogida en el comportamiento de las perturbaciones aleatorias, pudiendo introducir en éstas su propia variación, no necesariamente fija.

Recuérdese que la hipótesis inicial del MBRL de homocedasticidad hacía referencia a la varianza constante de las perturbaciones aleatorias, pero no obligaba a que las variables explicativas tuvieran también varianza constante, hecho que, además, sería una restricción muy poco plausible.

C.- Cambio de estructura

El hecho de que se produzca un cambio de estructura determina un mal ajuste de los parámetros al conjunto de los datos muestrales. Este no tiene porque influir del mismo modo en todo el recorrido de la muestra[1], pudiendo producir cuantías de desajuste del modelo diferentes y, por tanto, varianza no constante por subperíodos.

Al fin y al cabo, el fenómeno del cambio de estructura es equiparable a una especificación incorrecta por omisión de variables relevantes: precisamente faltaría la variable ficticia que distingue entre las dos situaciones o estructuras distintas que conviven en el período muestral elegido en el modelo.

D.- Empleo de variables no relativizadas

De un modo similar al comentado en el caso A), aquellas observaciones que contengan un valor mayor de una variable explicativa concreta (sospechosa de ser la que produce la heterocedasticidad) pueden originar valores del error diferentes.

Observadas las causas frecuentes de heterocedasticidad, es fácil deducir que la varianza no constante de las perturbaciones aleatorias viene casi siempre inducida por alguna variable, presente o no en el modelo, por lo que se podrían distinguir dos componentes en la varianza heterocedástica resultante del modelo: una cambiante, proveniente de esa variable que induce el problema, y una constante, que sería la que se daría si el modelo hubiera sido bien planteado. Matemáticamente podríamos escribir esto del siguiente modo:

$$s_i^2 = f(s^2 Z_i)$$

donde s^2 sería el parámetro fijo o parte fija de la varianza, y Z_i sería la matriz de variable o variables que está produciendo ese comportamiento no constante de la varianza de las perturbaciones aleatorias. Esta función podría ser empleada precisamente como el "supuesto simplificador" al que anteriormente se hacía referencia para posibilitar la estimación mediante MCG sin encontrarnos con más incógnitas que observaciones.

[1] De hecho, los parámetros estimados "recogerán mejor" el comportamiento de la serie en aquella de las dos estructuras distintas que se produzca durante mayor número de observaciones, ya que los parámetros estimados en presencia de un cambio de estructura serán una media ponderada de los que resultarían de una estimación particular para cada una de las dos submuestras

3.- Efectos de la heterocedasticidad sobre el MBRL

A. Incorrecta estimación de los parámetros

Dado que la matriz de varianzas-covarianzas es no escalar, el procedimiento correcto de estimación debe incluir la determinación de la matriz S; es decir, lo apropiado sería emplear los estimadores MCG o de Aitken cuya expresión es:

$$b^{MCG} = \left[X'\Sigma^{-1}X\right]^{-1}X'\Sigma^{-1}Y$$

Por supuesto, se ha demostrado que estos estimadores son lineales, insesgados, óptimos y consistentes para la estimación de una estructura no escalar de la matriz de varianzas-covarianzas siempre y cuando la estimación de la matriz S sea correcta. Esto, que pudiera parecer una perogrullada, debe llevarnos a una reflexión importante si miramos por un momento el carácter más aplicado de la cuestión.

Como ya se ha comentado, el elevado número de incógnitas a estimar respecto al número de observaciones (datos) nos obliga a hacer un supuesto simplificador sobre el comportamiento de la varianza heterocedástica. Evidentemente, es muy probable que, como con todo supuesto simplificador, al realizar la estimación de la matriz S bajo éste estemos sufriendo un cierto error o sesgo, con lo que la eficiencia absoluta teórica del estimador de Aitken frente al MCO en presencia de heterocedasticidad quedaría en entredicho.

B. Cálculo incorrecto de las varianzas y parámetros ineficientes

En el caso de obviar la heterocedasticidad para la estimación de los parámetros; es decir, seguir empleando la expresión MCO, caben dos opciones:

- Estimar también las varianza como si hubiera homocedasticidad en el modelo.
- Estimar los parámetros con MCO, pero calcular la verdadera varianza que les correspondería a estos cuando la matriz de varianzas-covarianzas de la perturbación aleatoria es no escalar.

Sobre esta reflexión es interesante recordar el experimento realizado por Goldfeldt y Quandt (1972)[2] en el que pretendían juzgar la ganancia en eficiencia (menor varianza) en los siguientes casos:

- Estimación de los parámetros con la expresión de MCG y cálculo correcto de sus varianzas correspodientes: $\text{cov}-\text{var}(\hat{b}) = s^2\left[X'\Sigma^{-1}X\right]^{-1}X'\Sigma^{-1}Y$

- Estimación de los parámetros con la expresión MCO y cálculo de las varianzas con la expresión que correspondería a un supuesto de homocedasticidad: $s^2\left[X'X\right]^{-1}$

[2] GOLFEDLD,SM Y QUANDT (1972): *Non Linnear Methods in Econometrics.* North Holland, pag. 280.

- Estimación de los parámetros con la expresión MCO y cálculo de las varianzas con la expresión que correspondería a un supuesto de heterocedasticidad: $\text{cov}-\text{var}(\hat{b}) = s^2[X'X]^{-1}X'\Sigma X[X'X]^{-1}$

Sobre un experimento controlado de generación de la varianza heterocedástica se llegaba a las siguientes conclusiones[3]:

1. La mayor varianza por empleo de MCO en vez de MCG en presencia de heterocedasticidad puede producir un incremento de más de 10 veces en la varianza estimada del parámetro constante y valores hasta 4 veces mayores en las varianzas de los parámetros que acompañan a variables explicativas.

2. Calcular la varianza de los estimadores ignorando la heterocedasticidad según la expresión que correspondería a una matriz de varianzas-covarianzas escalar, produce un sesgo por infravaloración de la real del orden del doble.

C. Invalidación de los contrastes de significatividad

Los contrastes que emplean para su cálculo estimaciones de la varianza o de su raíz cuadrada (desviación típica), sufrirán un claro sesgo deducible de lo dicho anteriormente:

- Si se elude el problema de la heterocedasticidad y se siguen empleando MCO, calculando erróneamente la varianza que correspondería a estos en el caso de que hubiera homocedasticidad $s^2[X'X]^{-1}$, ya se ha comentado que se estaría infravalorando la varianza real, por lo que contrastes de significatividad de los parámetros como la t-estadística o la F rechazarían la hipótesis nula con mayor frecuencia de la debida; es decir, aceptarían la validez de determinadas variables para explicar la endógena en casos en los que esto es falso.

- Si se emplearan MCO en la estimación, calculando correctamente su varianza en caso de heterocedasticidad ($s^2[X'X]^{-1}X'\Sigma X[X'X]^{-1}$), ya se ha comentado que estos parámetros arrojarían una importante ineficiencia respecto al empleo de MCG, por lo que, al contrario que en el caso anterior, se aceptaría la hipótesis nula de los contrastes de significatividad más veces de las reales.

4.- Cómo se contrasta

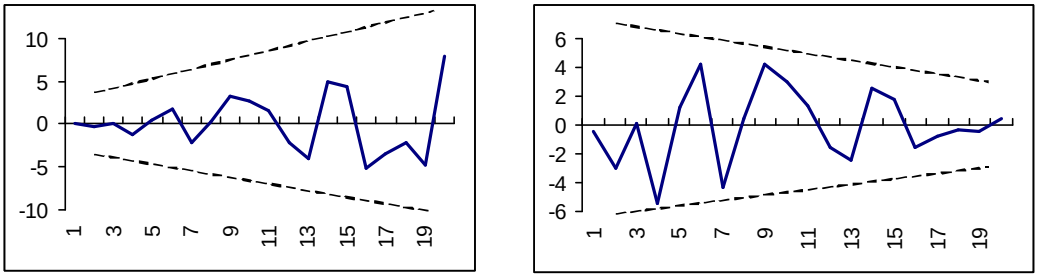
A. Contrastes Gráficos más habituales

A.1) Gráfica del error a través de las distintas observaciones del modelo

Dado que las series económicas presentan casi siempre una tendencia definida (positiva o negativa), la simple gráfica de error puede servir para conocer intuitivamente si el mero transcurso del tiempo da lugar a un incremento/decremento continuado del error, lo que sería significativo de una relación entre la evolución de las variables del modelo y los valores cada vez mayores o cada vez menores de éste.

[3] Sobre este experimento, se puede encontrar un detalle más profuso en la página 272 de Pulido, A. (1989).

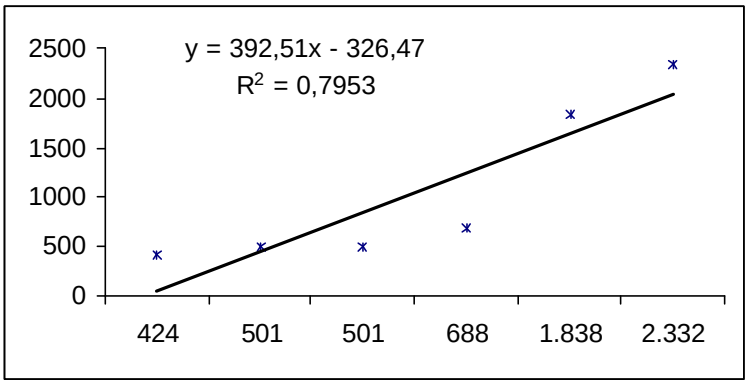
Gráficos del error sintomáticos de presencia de heterocedasticidad



En ambos, la mera evolución del tiempo está correlacionada con valores cada vez mayores (izquierda) del error o cada vez menores (derecha), con lo que el cálculo de la varianza por subperíodos arrojaría valores significativamente diferentes; es decir la serie del error sería heterocedástica.

A.2) Gráfica del valor absoluto del error en función de una explicativa sospechosa de producir la heterocedasticidad en el modelo

Si se ordena de menor a mayor la variable sobre la que se quiere investigar si produce o no heterocedasticidad y, con ella, los valores absolutos del error estimado, el hecho de que la nube de puntos obtenida en su gráfica conjunta (la variable en el eje de abcisas y el error absoluto en el eje de ordenadas) se pudiera aproximar correctamente con una regresión lineal significaría que el incremento de la variable explicativa da lugar a un incremento de las perturbaciones aleatorias, luego sería causa de heterocedasticidad en éstas.



El coeficiente de correlación por rangos de Spearman, explicado posteriormente, podría resultar un sistema numérico de contrastar la misma idea que subyace en la construcción de este gráfico.

B. Contrastes paramétricos

Varios de los contrastes que se desarrollan en este apartado tendrán un método para dirimir la significatividad de los valores obtenidos a partir de las tablas estadísticas de las funciones de densidad conocidas según la cual se distribuyen en cada caso los ratios propuestos. Es por esta razón por la que se llaman "paramétricos".

En particular, los contrastes que se presentan parten de una estructura acorde a la del Multiplicador de Lagrange. De forma muy intuitiva, sin querer hacer una argumentación estrictamente académica[4], diremos que en este tipo de contrastes se propone siempre dos

[4] Con mayor rigor técnico, el contraste del multiplicador de Lagrange evalúa el máximo obtenido en la función de verosimilitud cuando se estima un modelo con una serie de parámetros A o con estos más otros no incluidos

modelos, uno inicial y otro en el que se incorpora algún añadido en la especificación. A partir de un ratio sobre los errores de cada uno de estos modelos (o alguna transformada de estos), se compara si el modelo más completo aporta suficiente explicación adicional de la endógena como para compensar el coste de incorporar más variables.

B.1.) Contraste de Breusch-Pagan

La idea del contraste es comprobar si se puede encontrar un conjunto de variables Z que sirvan para explicar la evolución de la varianza de las perturbaciones aleatorias, estimada ésta a partir del cuadrado de los errores del modelo inicial sobre el que se pretende comprobar si existe o no heterocedasticidad.

El proceso a seguir para llevar a cabo este contraste es el siguiente:

1. Estimar el modelo inicial, sobre el que se pretende saber si hay o no heterocedasticidad, empleando MCO y determinando los errores.

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + u_i$$
$$\hat{b} = [X'X]^{-1} X'Y$$
$$e_i = y_i - \hat{y}_i$$

2. Calcular una serie con los errores del modelo anterior al cuadrado estandarizados:

$$\tilde{e}_i^2 = \frac{e_i^2}{\hat{s}^2}$$
$$\hat{s}^2 = \frac{e'e}{n}$$

   - Este valor al cuadrado nos elimina problemas de interpretación sobre la evolución media del error en el tiempo debidos a la compensación de signos que se produciría en cualquier cálculo agregado.
   - Por otra parte, la estandarización elimina distorsiones debidas a las posibles distintas dimensiones de los errores originales.

3. Se estima una regresión del error calculado en el paso (2) explicado por una constante y el conjunto de las variables Z que se pretende saber si producen o no heterocedasticidad en el modelo, obteniéndose la R^2 de este modelo y la varianza de la estimada:

$$\tilde{e}_i^2 = a_0 + a_1 z_{1i} + a_2 z_{2i} + ... + a_p z_{pi} + e_i$$
$$R^2_{\tilde{e}}$$

4. En principio, dado que el modelo tiene término constante, se cumple la regla general de las regresiones según la cual la varianza de la endógena real es igual a la suma de la varianza de la endógena estimada más la varianza del error obtenido en el modelo ($S^2_{\tilde{e}^2} = S^2_{\hat{e}^2} + S^2_{\hat{e}}$). Por ello, si el modelo es "malo" la varianza de la endógena estimada será pequeña (es lo mismo que decir que la varianza del error estimado es grande o que el "modelo tiene mucho error"). En definitiva, y siguiendo el interés que aquí buscamos,

previamente. Se emplea para dirimir la cuestión de qué modelo alcanza un máximo el cuadrado de la pendiente de la función de verosimilitud evaluada con cada uno de los conjuntos de parámetros.

si la varianza de la endógena estimada en este segundo modelo es muy pequeña, estaremos afirmando que el poder explicativo del conjunto de variables Z sobre la representación de la varianza de las perturbaciones aleatorias es escaso. A partir de esta afirmación, podríamos generar un contraste calculado con esta varianza, a sabiendas de que cuanto más cerca de cero se encuentre, más probabilidades de homocedasticidad habrá en el modelo. El contraste propuesto es:

$$\frac{S^2_{\hat{e}^2}}{2}$$

los autores demuestran que, en el caso de un modelo homocedástico, se distribuye como una c^2_p, con lo que, si el valor del ratio supera al valor de tablas, se rechaza la hipótesis nula; es decir, se acepta que el conjunto de variables Z no está produciendo heterocedasticidad en el modelo original.

El contraste de Breusch Pagan efectivamente nos servirá para aceptar o descartar la presencia de heterocedasticidad debida a ese conjunto de variables Z citado, pero su operatividad es limitada. Si el conjunto de las variables Z contiene variables no incluidas en el modelo original, parece difícil no haberlas tenido en cuenta antes para realizar una buena especificación y sí tenerlas en cuenta ahora para la contrastación. Por otro lado, la lista de variables Z debe ser necesariamente pequeña para poder realizarse el contraste.

B.2.) Contraste de Glesjer

De forma similar al caso anterior, Glesjer propone descartar la variación del error en función de una variable z, que ahora pueden estar elevadas a una potencia "h" que estaría comprendida entre -1 y 1. El modelo que se propone es:

1. Estimar el modelo inicial, sobre el que se pretende saber si hay o no heterocedasticidad, empleando MCO y determinando los errores.

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + u_i$$
$$\hat{b} = [X'X]^{-1} X'Y$$
$$e_i = y_i - \hat{y}_i$$

2. Estimar cuatro regresiones para los valores absolutos del error del modelo anterior en función de una variable elevada consecutivamente a "h", que para cada modelo tomaría los valores -1, -0,5, 0,5 y 1.

$$|e_i| = a_0 + a_1 z^h e_i \qquad h\{-1,-0.5,0.5,1\}$$

Se escogerá la regresión de las cuatro con parámetros significativos y con mayor R^2.

3. Se entiende que, si el valor de esta R^2 es suficientemente grande, se estará confirmando que existe heterocedasticidad producida por la variable z, ya que esta es capaz de explicar la evolución de la evolución del error como estimada de la evolución de las perturbaciones aleatorias.

B.3.) Contraste de White

En este contraste la idea subyacente es determinar si las variables explicativas del modelo, sus cuadrados y todos sus cruces posibles no repetidos sirven para determinar la evolución del error al cuadrado. Es decir; si la evolución de las variables explicativas y de sus varianzas y covarianzas son significativas para determinar el valor de la varianza muestral de los errores, entendida ésta como una estimación de las varianzas de las perturbaciones aleatorias.

El proceso a seguir para realizar este contraste sería el siguiente:

1. Estimar el modelo original por MCO, determinando la serie de los errores. Escrito esto en forma matricial para un modelo con "n" observaciones y "k" variables explicativas:

$$Y = Xb + U$$
$$\hat{b} = \left[X'X\right]^{-1} X'Y$$
$$\hat{Y} = X\hat{b}$$
$$e = Y - \hat{Y}$$

2. Estimar un modelo en el que la endógena sería los valores al cuadrado de los errores obtenidos previamente (paso 1) con todas las variables explicativas del modelo inicial, sus cuadrados y sus combinaciones no repetidas.

$$e_i^2 = a_0 + a_1 x_{1i} + ... + a_k x_{ki} + a_{k+1} x_{1i}^2 + ... + a_{k+k} x_{ki}^2 + a_{k+k+1} x_{1i} x_{2i} +$$
$$a_{k+k+2} x_{1i} x_{3i} + ... + a_{3k+1} x_{2i} x_{3i} + ... + e_i$$

3. El valor de la R_e^2 de este segundo modelo (paso 2) nos dirá si las variables elegidas sirven o no para estimar la evolución variante del error al cuadrado, representativo de la varianza estimada de las perturbaciones aleatorias. Evidentemente, si la varianza de éstas fuera constante (homocedasticidad), el carácter no constante de las variables explicativas implicadas en el modelo no serviría para explicar la endógena, luego la R_e^2 debiera ser muy pequeña.

En principio, la R_e^2, como proporción de la varianza de la endógena real[5] que queda explicada por la estimada, debiera ser muy pequeña si la capacidad explicativa de los regresores considerados también es muy pequeña, siendo estos regresores, por su construcción, representativos de varianzas y covarianzas de todas las explicativas del modelo original. Dicho esto, evidentemente un valor de la R^2 suficientemente pequeño servirá para concluir que no existe heterocedasticidad en el modelo producida por los valores de las explicativas consideradas en el modelo inicial. Para encontrar el valor crítico en esa consideración de "suficientemente pequeño" se emplea la expresión deducida por Breusch y Pagan como producto del coeficiente R^2 por el número de datos del modelo, que se distribuiría del siguiente modo:

$$n \cdot R_e^2 \rightarrow c_{p-1}$$

[5] En este caso, la endógena real será el valor del error muestral al cuadrado de la primera regresión practicada. En el caso de homocedasticidad, este debe ser casi constante, por lo que difícilmente la evolución de otras variables podría explicar un valor fijo. Por ello es intuitivo pensar que cuanto mayor sea la R_2 de este modelo, más probable será la heterocedasticidad.

En definitiva, si obtenemos un valor del producto $n \cdot R_e^2$ mayor que el reflejado por las tablas de c_{p-1}, afirmaremos que existe heterocedasticidad, y viceversa, si este valor es más pequeño diremos que no se mantiene la homocedasticidad.

Otro modo de contrastar la existencia de heterocedasticidad en el modelo a partir de la validez o no de los parámetros incluidos en la regresión propuesta por White vendría dado por el valor del contraste de significación conjunta F. Si dicho contraste afirmara que, en conjunto, las variables explicitadas tienen capacidad explicativa sobre la endógena, estaríamos afirmando la presencia de heterocedasticidad en el modelo.

B.4.) Contraste a partir del coeficiente de correlación por rangos de Spearman

La filosofía de este contraste reside en que la variable sospechosa de producir heterocedasticidad debería provocar un crecimiento del residuo estimado al mismo ritmo que ella va creciendo. Por ello, si ordenáramos de menor a mayor tanto la variable "sospechosa", por ejemplo x_{ji}, como el valor absoluto del residuo, $|e_i|$, el cambio de puesto en ambas, y para cada una de las observaciones, debiera ser del mismo número de puestos respecto al orden original de las series. En la medida en la que este cambio de puesto respecto al original no sea el mismo para las dos (una vez ordenadas) se podría hablar de movimientos no correlacionados. Dado que la correlación se mide entre uno y menos uno, Spearman propone determinar un grado de correlación en ese "cambio de puesto respecto al inicial" de cada una de las variables a partir de la diferencia entre el nuevo puesto y el inicial:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

En esta expresión, una coincidencia máxima (todas las distancias son igual a cero), daría lugar a una correlación de Spearman igual a uno; mientras que una distancia máxima, provocaría un valor cero de dicho coeficiente de correlación[6].

En la siguiente tabla se hace un pequeño ejemplo numérico de cálculo del coeficiente de Spearman para clarificar lo dicho hasta ahora.

Series originales			Series ordenadas				d	d^2
Puesto	x_{ji}	$\|e_i\|$	x_{ji}	Puesto original	$\|e_i\|$	Puesto original		
1	1.838	1,6	424	2	1,2	3	2-3=-1	1
2	424	1,4	501	3	1,3	4	3-4=-1	1
3	501	1,2	688	5	1,4	2	5-2=3	9
4	2.332	1,3	1.838	1	1,5	5	1-5=-4	16
5	688	1,5	2.332	4	1,6	1	4-1=3	9

[6] Realmente, el coeficiente de correlación por rangos de Spearman es equivalente a emplear el coeficiente de correlación lineal $r(x, y) = \frac{\text{cov}(x, y)}{S_x S_y}$ a las variables de puntuación de orden de ambas colocadas según la progresión de una de ellas. Para ver el detalle del denominador, se puede acudir a Martín-Guzmán y Martín Pliego (1985), páginas 312-314.

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = 1 - \frac{6 * 30}{5(25 - 1)} = 1 - 1{,}8 = -0{,}8$$

En este caso, el grado de correlación negativa de ambas series sería bastante elevado, dado que los extremos de correlación serían +/-1.

Para valorar la significatividad o no de esta correlación, se conoce la función de distribución del siguiente ratio bajo la hipótesis nula de no significatividad, demostrado por el autor:

$$\frac{r_s\sqrt{N-2}}{\sqrt{1-r_S^2}} \rightarrow t_{n-2}$$

Con ello, si el resultado del ratio es superior al valor de tablas podremos afirmar que la correlación es significativa o, de cara a nuestro interés en este caso, que hay indicios de heterocedasticidad en el modelo provocada por la variable x_{ji}.

B.5) Otros contrastes

Aunque no se comentarán aquí, si es conveniente citar otros contrastes habituales para la determinación de la heterocedasticidad, como:

- Contraste de Harvey
- Contraste RESET de Ramsey
- Golfeld-Quandt
- Contraste de picos
- LM Arch

5.- Cómo se corrige

Como hemos venido viendo repetidas veces a lo largo del tema, la heterocedasticidad viene producida por la dependencia de la varianza de las perturbaciones aleatorias de una o más variables que, a su vez, pueden estar presentes en el modelo o no. Los distintos métodos de detectar este problema servían para probar, en el caso en el que ésta realmente se diese, la dependencia de la varianza de la perturbación aleatoria de un conjunto de variables, a partir de lo que hemos llamado un supuesto simplificador:

$$s_i^2 = f(s^2 Z_i)$$

Por lógica, el modo de subsanar el problema detectado será operar convenientemente la variables del modelo precisamente eliminando la fuente de heterocedasticidad que habremos podido definir cuando detectamos la misma. Como veremos a continuación, si el conjunto total de las variables del modelo (endógena incluida) es dividido por la forma estimada de esta función de la raíz de la varianza heterocedástica (una vez algún método de detección nos haya confirmado que efectivamente el comportamiento de esta varianza se puede seguir convenientemente con dicha función) estaremos corrigiendo el modelo.

Para comprobar esto, podemos volver a la forma matricial de varianzas covarianzas no escalar:

$$E(UU') = \begin{bmatrix} E(u_1)^2 & & & . \\ E(u_1u_2) & E(u_2)^2 & & \\ & & ... & \\ E(u_1u_n) & E(u_2u_n) & & E(u_n)^2 \end{bmatrix} = \begin{bmatrix} E(u_1)^2 & & & . \\ 0 & E(u_2)^2 & & \\ 0 & 0 & ... & \\ 0 & 0 & 0 & E(u_n)^2 \end{bmatrix} = s_i^2 I_n = s^2\Sigma$$

En esta matriz, si dividimos por la raíz de $s_i^2 = f(s^2 Z_i)$, obtendremos una diagonal principal de unos; es decir, volveríamos al caso de una matriz de varianzas covarianzas escalar tal y como la que se supone en el modelo básico de regresión lineal.

Formalmente, para probar esto seguimos los siguientes pasos. Dado que la matriz S es una matriz semidefinida positiva (todos los elementos de su diagonal principal son necesariamente positivos), siempre podremos descomponerla en dos matrices de la forma:

$$\Sigma = PP' \Leftrightarrow \Sigma^{-1} = P^{-1}P^{-1}'$$

Volviendo a la matriz de varianzas covarianzas no escalar y uniendo esto a la función que hemos comprobado sirve para definir esta varianza no constante $s_i^2 = f(s^2 Z_i)$, es fácil llegar a que la descomposición $\Sigma = PP' \Leftrightarrow \Sigma^{-1} = P^{-1}P^{-1}'$ es:

$$\begin{bmatrix} E(u_1)^2 & & & . \\ 0 & E(u_2)^2 & & \\ 0 & 0 & ... & \\ 0 & 0 & 0 & E(u_n)^2 \end{bmatrix} = \begin{bmatrix} s_1^2 & & & . \\ 0 & s_2^2 & & \\ 0 & 0 & ... & \\ 0 & 0 & 0 & s_n^2 \end{bmatrix} = s^2\Sigma =$$

$$\begin{bmatrix} s_1 & & & . \\ 0 & s_2 & & \\ 0 & 0 & ... & \\ 0 & 0 & 0 & s_n \end{bmatrix}' \begin{bmatrix} s_1 & & & . \\ 0 & s_2 & & \\ 0 & 0 & ... & \\ 0 & 0 & 0 & s_n \end{bmatrix} = s^2 PP'$$

Si multiplicamos cada variable del modelo por esta matriz P, tal y como se ha sugerido, obtenemos unas nuevas variables del siguiente tipo:

$$P^{-1}Y = P^{-1}xb + P^{-1}U \Leftrightarrow Y^* = X^*b + U^*$$

donde:

$$E(U^*U^{*}') = E(P^{-1}UU'P^{-1}') = P^{-1}P^{-1}'E(UU') = \begin{Bmatrix} E(UU') = s^2\Sigma \\ P^{-1}P^{-1}' = \Sigma^{-1} \end{Bmatrix} = \Sigma^{-1}s^2\Sigma = s^2 I_n$$

Por lo que podemos afirmar que el modelo transformado (aquel por el que se han dividido todas las variables por la desviación típica estimada de las perturbaciones aleatorias) soporta una matriz de varianzas covarianzas de las perturbaciones aleatorias escalar, con lo que se puede estimar con toda garantía por MCO.

En definitiva, y a modo de breve "receta", los pasos para la corrección de la heterocedasticidad serían los siguientes:

a) Se estiman los parámetros del modelo por MCO, ignorando por el momento el problema de la heterocedasticidad de las perturbaciones aleatorias
b) Se establece un supuesto acerca de la formación de s_i^2 y se emplean los residuos de la regresión por MCO para estimar la forma funcional supuesta.
c) Se divide cada observación por $\sqrt{\hat{s}_i^2}$ según el paso anterior (según el valor de esa heterocedasticidad supuesta estimada, siempre y cuando un contraste nos haya confirmado que el "modelo simplificador" es bueno).
d) Se estima el modelo original ahora con todas las variables transformadas según el paso c).

TRATAMIENTO DE LA HETEROCEDASTICIDAD EN E-VIEWS

Se propone un modelo para cuantificar las ventas de Burger King (VTASBK) en una serie de 20 países, proponiéndose como explicativas las siguientes variables:

PRECIOSBK:	Precios Hamburguesa Whoper
PRECIOSMAC:	Precios Hamburguesa Big Mac
RENTAPC:	Renta per capita del país

Realizada una primera regresión, los resultados obtenidos son los siguientes:

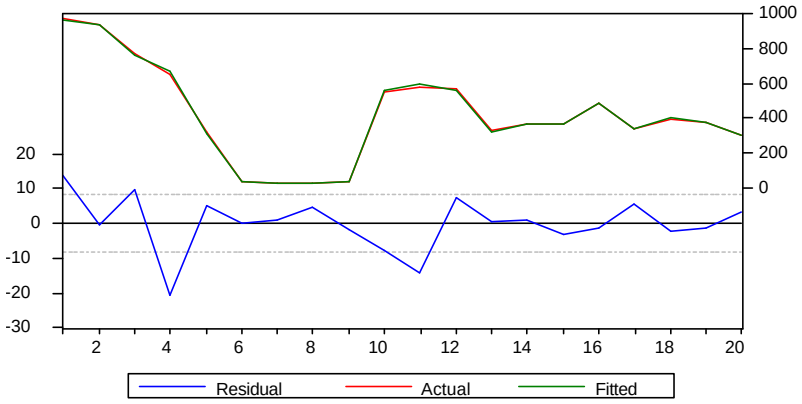
Equation: UNTITLED Workfile: PARA DOC HETEROCEDASTIC

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: VTASBK
Method: Least Squares
Date: 04/04/01 Time: 17:36
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23.78791	16.26076	1.462903	0.1629
PRECIOSBK	-2.356251	12.96562	-0.181731	0.8581
PRECIOSMAC	-16.74075	19.25262	-0.869531	0.3974
RENTAPC	0.025278	0.000189	133.7319	0.0000

R-squared	0.999224	Mean dependent var	421.8982
Adjusted R-squared	0.999078	S.D. dependent var	278.2593
S.E. of regression	8.447007	Akaike info criterion	7.282358
Sum squared resid	1141.631	Schwarz criterion	7.481504
Log likelihood	-68.82358	F-statistic	6867.346
Durbin-Watson stat	2.376763	Prob(F-statistic)	0.000000



Matriz de correlaciones de las variables

	VTASBK	PRECIOSBK	PRECIOSMAC	RENTAPC
VTASBK	1.000000	0.360900	0.226085	0.999566
PRECIOSBK	0.360900	1.000000	0.704328	0.367945
PRECIOSMAC	0.226085	0.704328	1.000000	0.235402
RENTA PC	0.999566	0.367945	0.235402	1.000000

No se da ninguna correlación entre variables explicativas superior al R^2 obtenido en el modelo, por lo que no parece haber indicios de multicolinealidad. Tan sólo existe una fuerte correlación entre PRECIOSBK y PRECIOSMAC (0,7043), en cualquier caso más pequeño que el 0,99.

A la luz del gráfico de residuos, podría pensarse que que los cinco primeros países presentarían una varianza mayor que los siguientes, aunque, como suele ocurrir con los gráficos, no se puede apreciar nada claramente.

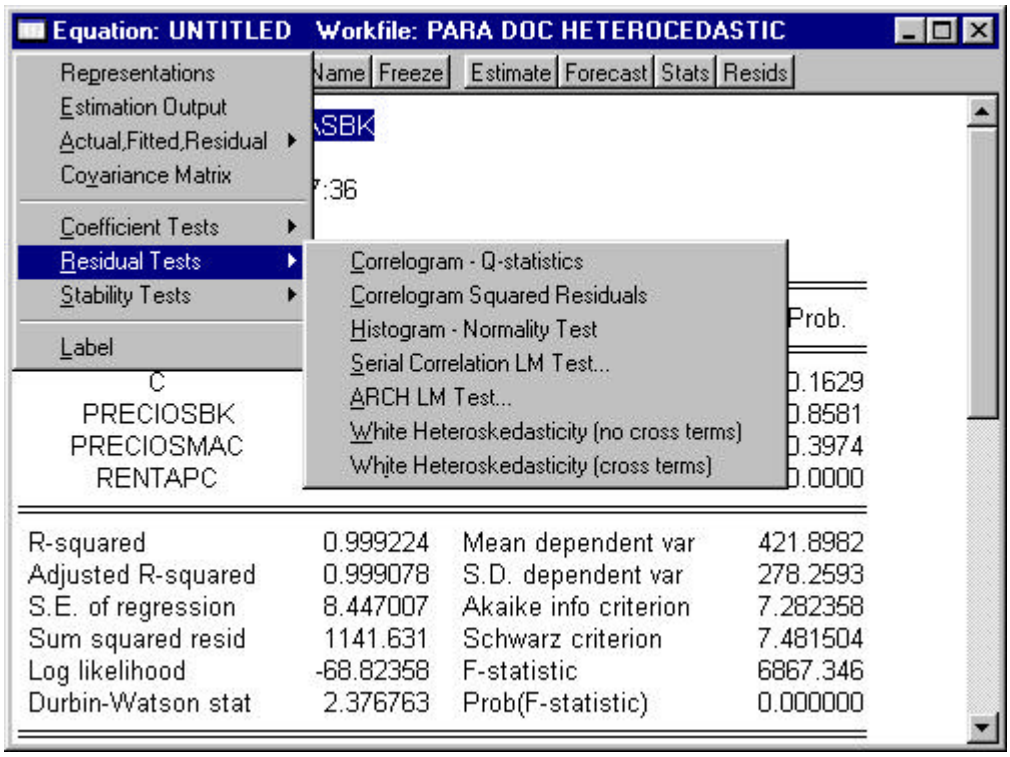
El siguiente elemento a contrastar sería la presencia de heterocedasticidad en el modelo. El programa E-Views suministra, con este fin, la posibilidad de detectar la heterocedasticidad, entre otros, a partir del Test de Residuos de White, ofreciendo dos posibilidades:

- ***No Cross Terms***: Realizar la regresión de los errores al cuadrado de la regresión inicial del modelo escribiendo como explicativas todas las exógenas de la inicial y sus valores al cuadrado.
- ***Cross Terms:*** igual que la anterior, pero incluyendo además, como explictivas del error al cuadrado, los productos no repetidos de todas las variables explicativas del modelo inicial entre sí.

En principio, el contraste expresado por White sería la segunda opción, pero, en modelos con escasas observaciones, a lo mejor no es posible realizar la estimación con tantos regresores y es más recomendable la primera opción (por no eliminar completamente los grados de libertad).

En nuestro caso, el número de observaciones es 20 (países) y el número de explicativas tres más la constante, luego el contraste de White con términos cruzados equivaldría a incluir 10 variables explicativas sobre el cuadrado de los errores de la regresión inicial (la constante, las tres explicativas, sus tres cuadrados y los tres cruces posibles no repetidos entre ellas).

Para aplicar este contraste en E-views, desde la misma ventana donde se ha realizado la regresión, se sigue el siguiente trayecto:



Los resultado de este Test de residuos White heteroskedasticity (cross terms) son:

White Heteroskedasticity Test:			
F-statistic	7.458779	Probability	0.002102
Obs*R-squared	17.40694	Probability	0.042712

Como resultado, se nos ofrecen dos formas de contrastar la validez de las variables elegidas para explicar un comportamiento no homogéneo del error al cuadrado (estimador de la varianza de la perturbación aleatoria en este caso):

- F-stastitic (como siempre con k-1; n-k grados de libertad), nos vendría a dar una medida de la bondad del modelo (probabilidad de heterocedasticidad si se confirma la validez conjunta de las variables elegidas para determinar la variación del error al cuadrado - la endógena-).

- Obs*R-squared ($n \cdot R_e^2 \rightarrow c_{p-1}$): supuesta la hipótesis nula de homocedasticidad, el cálculo propuesto debería comportarse como una c_{p-1} con p-1 grados de libertad. En nuestro caso p=10 (las explicativas de la regresión practicada). El valor de tablas de c_{10-1}^2, para el 95% de confianza, es 16,9.

A la luz de lo dicho, ambos estadísticos propuestos afirman, con un 97,9% de probabilidades el primero y con un 96,73% de probabilidades el segundo, la existencia de heterocedasticidad.

La misma salida nos muestra la regresión utilizada para realizar estos cálculos, que sería la siguiente:

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/04/01 Time: 18:13
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1244.796	761.4100	-1.634856	0.1331
PRECIOSBK	-3872.145	4225.294	-0.916420	0.3810
PRECIOSBK^2	1071.919	452.8574	2.367012	0.0395
PRECIOSBK*PRECIOSMAC	-423.3863	3433.568	-0.123308	0.9043
PRECIOSBK*RENTAPC	0.065588	0.019299	3.398529	0.0068
PRECIOSMAC	6562.125	4306.578	1.523745	0.1586
PRECIOSMAC^2	-2332.049	3209.945	-0.726507	0.4842
PRECIOSMAC*RENTAPC	0.048495	0.039574	1.225423	0.2485
RENTAPC	-0.090230	0.034504	-2.615066	0.0258
RENTAPC^2	-7.94E-07	2.13E-07	-3.719294	0.0040
R-squared	0.870347	Mean dependent var		57.08155
Adjusted R-squared	0.753659	S.D. dependent var		104.1942
S.E. of regression	51.71438	Akaike info criterion		11.03620
Sum squared resid	26743.77	Schwarz criterion		11.53407
Log likelihood	-100.3620	F-statistic		**7.458779**
Durbin-Watson stat	1.810789	Prob(F-statistic)		0.002102

A la luz de esta regresión, es fácil comprobar la significatividad de la variable rentapc y rentapc^2 para explicar la varianza del error. También los es preciosbk^2 y preciosbk*rentapc.

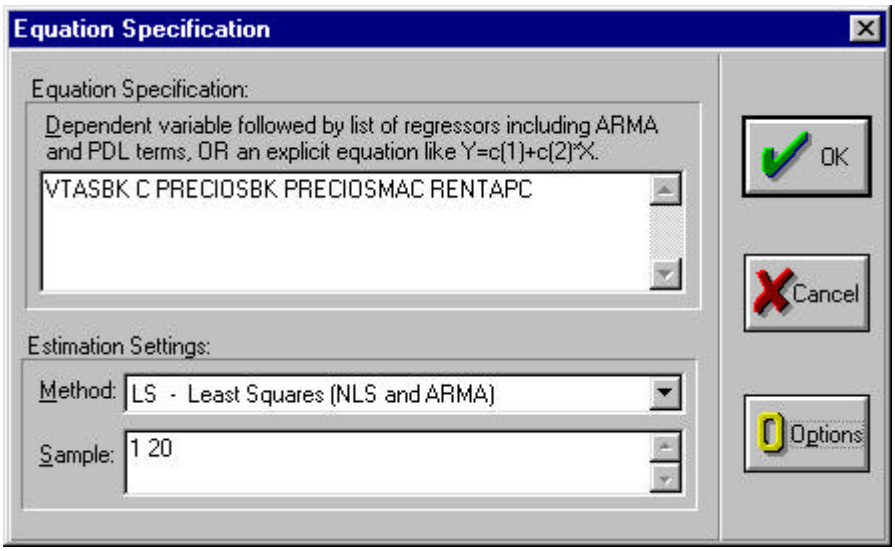
Para corregir el problema de la heterocedasticidad, habría que emplear Mínimos Cuadrados Generalizados, o bien transformar todas las variables del modelo predividiendo todas sus observaciones por la raíz cuadrada del valor estimado del error al cuadrado en el modelo que se ha utilizado para contrastar la presencia de heterocedasticidad y que nos ha informado sobre la presencia de la misma y la buena explicación del comportamiento no constante de la varianza.

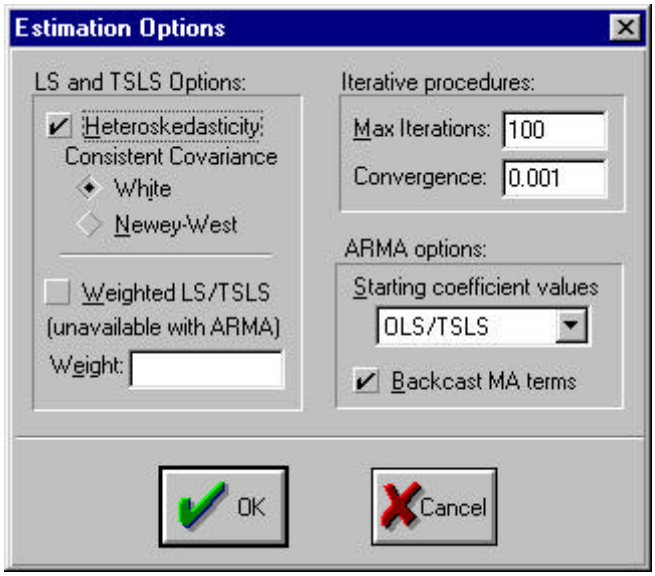
El programa E-views permite realizar la estimación por MCG usando como valor de S el obtenible a partir de la propuesta de White (1980).

El estimador consistente de la matriz de covarianzas para lograr una estimación correcta de los parámetros en presencia de heterocedasticidad es el siguiente:

$$\hat{\Sigma} = \frac{n}{n-k}[X'X]^{-1}\left(\sum_{i=1}^{n} e_{it}^2 x'_t \; x_t\right)[X'X]^{-1}$$

Para lograr una estimación empleando esta corrección en E-views, es necesario, una vez se ejecuta una estimación lineal normal, pulsar el botón de "estimate". Aparecerá entonces, a la derecha, un botón de opciones que, pulsado, permite señalar "Heteroskedasticity: consistent covariance White".





Estimando según esta propuesta, ya que hemos confirmado la presencia de heterocedasticidad, los resultados serían los siguientes:

Dependent Variable: VTASBK
Method: Least Squares
Date: 04/20/01 Time: 13:37
Sample: 1 20
Included observations: 20

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23.78791	8.785312	2.707691	0.0155
PRECIOSBK	-2.356251	7.695290	-0.306194	0.7634
PRECIOSMAC	-16.74075	13.70312	-1.221674	0.2395
RENTAPC	0.025278	0.000213	118.6913	0.0000

R-squared	0.999224	Mean dependent var	421.8983
Adjusted R-squared	0.999078	S.D. dependent var	278.2593
S.E. of regression	8.447007	Akaike info criterion	7.282358
Sum squared resid	1141.631	Schwarz criterion	7.481504
Log likelihood	-68.82358	F-statistic	6867.346
Durbin-Watson stat	2.376763	Prob(F-statistic)	0.000000

Referencias bibliográficas

GOLFEDLD,SM Y QUANDT (1972): *Non Linnear Methods in Econometrics.* North Holland, pag. 280.

MARTÍN-GUZMÁN Y MARTÍN PLIEGO (1985): Curso básico de Estadística Económica. Editorial AC

NOVALES, A. (1993): Econometría. Editorial M'c Graw Hill, segunda edición. Madrid.

OTERO, JM (1993): Econometría. Series temporales y predicción. Editorial AC, libros científicos y técnicos. Madrid.

PULIDO, A. y PÉREZ, J. (2001): Modelos Econométricos. Editorial Pirámide, SA. Madrid.