HETEROSCEDASTICIDAD

Eco. Douglas C. Ramírez V.

1) INTRODUCCIÓN

Las propiedades de los estimadores mínimos cuadráticos de los coeficientes de regresión dependen de las propiedades del termino de perturbación del modelo. Entre los supuestos asumidos, se postula que los errores tienen una esperanza igual a cero, una varianza constante y son independientes entre sí. Ahora se estudiara los problemas que surgen cuando uno los supuestos es violado. El de varianza constante. Sé vera que los estimadores Mínimo Cuadráticos Ordinarios (MCO) respectivos son poco representativos y resultaría más adecuado utilizar procedimientos alternativos de estimación.

El presente documento pretende resumir los aspectos relacionados con la violación del supuesto de homoscedasticidad, para ello se ha estructurado en siete secciones incluyendo la introducción. En la segunda sección presenta las condiciones para que el método de estimación por mínimos cuadrados ordinarios (MCO) cumpla las condiciones para ser los mejores estimadores lineales e insesgados de acuerdo al Teorema de Gauss-Markov. En la tercera sección se establece las condiciones de homoscedasticidad y se compara con una varianza heteroscedástica. En la cuarta sección se describe las posible causas que pueden originar que los datos presenten un patrón heteroscedástico. En la quinta sección se resume las consecuencias en la confiabilidad y valides de la estimación. En la sexta parte del documento se introduce las medidas remédiales para corregir el problema de una varianza no constante y en la séptima sección se habla de los métodos de detección intuitivos y comprobatorios. Anexo se lista los datos y las instrucciones realizadas para las pruebas escritas en el programa LIMDEP ©.

En este documento la discusión se centrará en la importancia de la condición, cuando puede ser violada, como puede ser detectada y que se puede hacer para ser corregida y así obtener estimadores adecuados.

2) LAS CONDICIONES GAUSS-MARKOV

El método de los Mínimos Cuadrados Ordinarios (MCO) asume que el termino de error en el modelo de regresión satisface las cuatro condiciones básicas del Teorema de Gauss-Markov, las cuales se resumen (Dougherty, 1992) en términos de los siguiente aspectos. A partir del modelo

$$(2.1) \qquad Y_i = \alpha + \beta X_i$$

Donde u_i y X_i son valores de las perturbaciones y de las variables respectivamente. Las condiciones establecidas serían:

- Condición de Insesgadez;

$$(2.2) \qquad E(u_i)=0$$

- Condición de Homoscedasticidad;

$$(2.3) \qquad E(u_i^2)=\sigma^2$$

- Condición de No Autocorrelación Serial;

$$(2.4) \qquad E(u_i, u_j)=0$$

- Condición de Independencia;

$$(2.5) \qquad E(X_i, u_j)=0$$

Si el modelo de regresión es múltiple (de más de una variable explicativa), entonces simplemente las condiciones son las mismas, excepto que la última debe ser satisfecha por cada una de las variables explicativas. Si no se cumple una de las condiciones previstas, exceptuando tal vez la primera que es producto de esencialmente de la definición, no se obtendrían estimadores MELI (Mejores Estimadores Lineales Insesgados).

En este documento se estudiará esencialmente las condiciones de la violación del segundo supuesto, el de varianza constante en toda la muestra observada o supuesto de Homocedasticidad.

3) LA CONDICIÓN DE HOMOSCEDASTICIDAD

3.1. Intuición

Una de las condiciones del Teorema de Gauss-Markov para estimar los parámetros lineales más eficientes establece, entre otros, que *la varianza de los términos de disturbio para cada muestra observada debe ser constante*. Esta afirmación se refiere que para el modelo estimado: $Y_i=\alpha+\beta X_i+u_i$ la condición significa que el término de disturbio; $\{u_1, u_2, \ldots, u_N\}$; en las N observaciones, llamando N el total de observaciones muéstrales, potencialmente tendrían un valor medio o esperanza muestral igual a cero y su varianza debe ser constante e igual para cada una de las muestras (Dougherty, 1992).

Se espera, por tanto, que algunos valores de los errores estimados (y de las perturbaciones) serán positivos y otros negativos pero la mayoría relativamente cercanos a cero. Se espera, a priori, que no exista ningún patrón de comportamiento sistemático de los errores alrededor de su valor esperado o cero y se distribuyan uniformemente presentado "*la misma dispersión"*. Esta condición es conocida como homoscedasticidad.

En algunas muestras, sin embargo, puede ser razonable suponer que la distribución potencial del término de disturbio es diferente en "*diferentes*" observaciones de la "*misma*" muestra. En la figura 1 se ejemplifica como a lo largo de las muestras (una es homoscedástica y la otra heteroscedástica) viendo tres secciones proporcionales de la misma, se observa que va cambiando la dispersión de la misma, en el caso de la distribución uniforme, u homoscedástica, no se altera, en cambio, en la que no sigue una distribución uniforme esta cambia significativamente entre la primera y tercera sección.

Figura 1

Homoscedasticidad vs. Heteroscedasticidad

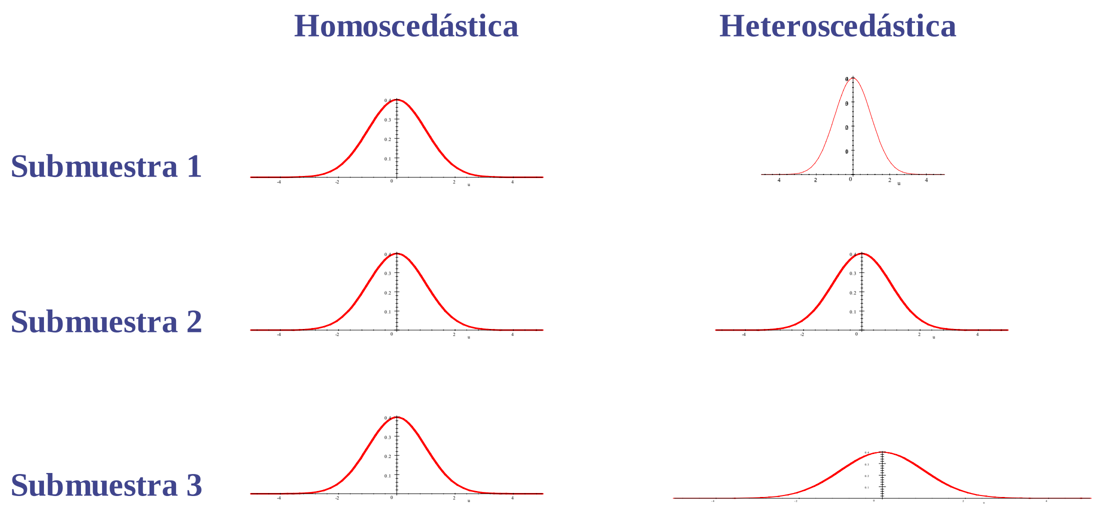
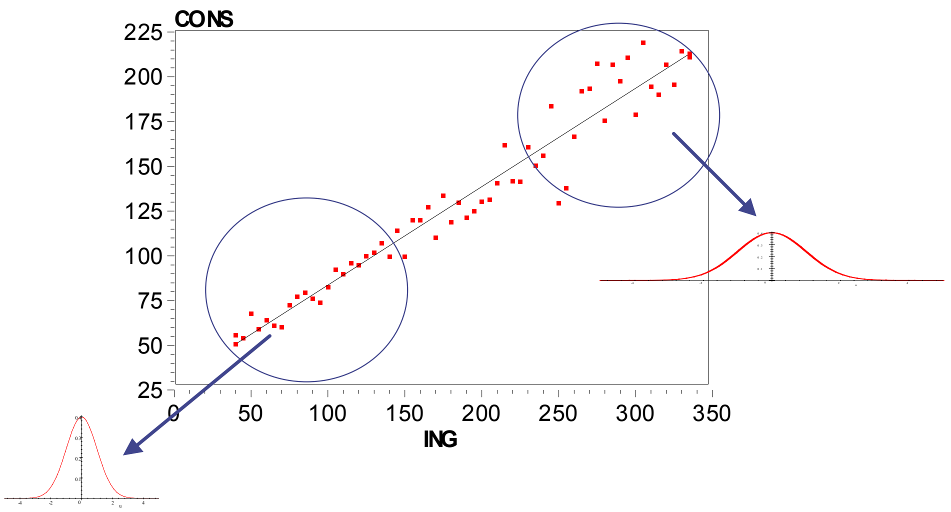


Figura 2

Modelo Con Perturbaciones Heteroscedásticas



En el ejemplo que se muestra en la Figura 2, se relaciona el consumo de diferentes hogares versus el ingreso de los mismo, se puede observar que en el tramo inferior izquierdo, de menor ingreso y menor consumo, se presenta una menor dispersión y en el tramo superior derecho, de mayor ingreso y de mayor consumo, existe una mayor dispersión a lo largo de la línea media del modelo de regresión estimado por MCO. En este ejemplo se indica diferentes dispersiones para diferentes tramos de la muestra e indica a su vez que el comportamiento errático es relativamente más grande al final de la muestra que al principio.

3.2. Planteamiento Matricial

Matemáticamente la varianza homoscedástica y la heteroscedasticidad pueden definirse, en términos del modelo de regresión lineal general (Greene, 1999). En forma compacta se define un vector columna Y de N filas con la variable dependiente, una matriz de rectangular de N filas y K columnas de variables explicativas, un vector de parámetros de K filas y un vector columna de perturbaciones de N filas.

$$(3.1) \qquad \begin{bmatrix} Y_1 \\ \vdots \\ Y_1 \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{(K-1)1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{(K-1)N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ U_N \end{bmatrix}$$

Nótese que se tienen que cumplir las condiciones de conformidad matricial para que permitan resolver el sistema por el método de MCO, para ello la matriz debe ser invertible, en este caso se añade un supuesto adicional que los K vectores columnas deben ser linealmente independiente. Este supuesto se conoce como ausencia de multicolinealidad o condición de rango.

El modelo matricial se escribe a continuación, sin indicar sus dimensiones por simplicidad, como:

$$(3.2) \qquad Y = X\beta + U$$

Las condiciones en términos matriciales serían:

$$(3.3.) \qquad E(U)=0$$

$$(3.4) \qquad E(UU')=\sigma^2\Omega$$

Donde **Ω** es una matriz definida positiva. Aun cuando se supone todavía que las perturbaciones están incorrelacionadas entre observaciones, es decir, la covarianza entre los errores son nulas (supuesto que se levanta cuando se estudia los problemas de autocorrelación serial) y sólo en la diagonal principal de la matriz existen valores, por lo tanto, se puede expresar la última matriz como:

$$(3.5) \qquad \sigma^2\Omega = \begin{bmatrix} \sigma\omega_1 & 0 & \cdots & 0 \\ 0 & \sigma\omega_2 & \cdots & \vdots \\ \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \vdots & \sigma\omega_N \end{bmatrix}$$

La representación que se escribe resulta útil para expresar diferentes patrones de comportamiento de la varianza de la regresión, donde la varianza homoscedástica, bajo los supuestos clásicos, sería un caso entre varios. Por lo cual se puede escribir los elementos de la diagonal de la matriz sigma-omega ($\sigma^2\Omega$)como:

$$(3.6) \qquad \sigma_i^2 = \sigma^2\omega_i$$

Donde la varianza homoscedástica se da cuando; ω_i=1, (y en consecuencia; $\boldsymbol{\sigma\Omega=\sigma I_n}$) para todos los casos. Esta formulación permite visualizar que si se detecta el patrón de comportamiento de los elementos omega sub i (ω_i) que genera la alteración de la varianza a lo largo de la muestra, es posible extraer este efecto y obtener la varianza correcta y así realizar las pruebas estadísticas que permitan validar el modelo. Luego entonces, los estimadores MCO pueden ser expresados como siempre de la forma.

$$(3.7) \qquad b = \beta + \left(X'X\right)^{-1} X'U$$

Siendo insesgados dado que E(U)=0 y por tanto E(b)=β; pero su matriz de varianza y covarianza sería ahora.

$$(3.8) \quad Var(b) = E[(b-\beta)(b-\beta)'] == \{[(X'X)^{-1}X'U]\ [\ (X'X)^{-1}X'U]'\}$$

$$(3.9) \qquad Var(b) = E\ \{(\ (X'X)^{-1}X'UU'X(X'X)^{-1}\}$$

$$(3.10) \qquad Var(b) = \sigma^2\{\ (X'X)^{-1}X'\Omega X(X'X)^{-1}\}$$

Por lo cual la formula tradicional de estimación de la matriz de Varianzas MCO como; $\sigma^2(X'X)^{-1}$, ya no resulta valida y cualquier aplicación de ella arrojará resultados erróneos. Pero adicionalmente si se pudiese utilizar (3.10) para estimar las varianzas muéstrales, el uso el las pruebas de contraste de la bondad de ajuste (pruebas t y F entre otras) estos se invalidarían ya que no cumplirían los supuestos que se usaron para construir las pruebas de inferencia y por la misma razón tampoco cumple la propiedad de varianza mínima de los MCO.

En general la existencia de heterocedasticidad no afecta las condiciones de sesgo de los estimadores pero invalida las pruebas estadísticas de contrastes asociadas a los intervalos de confianza a partir de las distribuciones T de Student y F de Fisher-Snedecor

4) CAUSAS

La Heteroscedasticidad surge normalmente en datos de sección cruzada donde la escala de la variable dependiente y el poder explicativo de la tendencia del modelo varia a lo largo de las observaciones. En muchos casos los datos microeconómicos, como los estudios de gasto de las familias, presentan este comportamiento. Intuitivamente la presencia de diferente varianza para diferentes observaciones se pueden ejemplificar en las siguientes causas[1].

- Razones de escala
- Distribución espacial

[1] Para ver otros ejemplos véase a: Kamenta, 1977, Dougherty, 1992, Gujarati, 1997, Hair et al., 1999; entre otros.

- Aprendizaje de los errores
- Menores restricciones de elección.
- Mejoras en la Información
- Observaciones Influyentes
- Problemas de especificación

4.1. Razones de escala

Sí, por ejemplo, se quiere explicar y cuantificar el gasto en innovación que una empresa realiza, mediante un conjunto de variables explicativas como el capital o el trabajo (mídase como se mida) o se quiere explicar la estructura del empleo o la tasa de accidentes, es muy probable que se produzca una mayor variación en el gasto total de las empresas de mayor tamaño (incluso controlando la regresión por el tamaño) razón por la cual probablemente las perturbaciones de los errores de estimación presenten un comportamiento heterocedastico.

4.2. Distribución espacial

En estudios de corte transversal (por ejemplo datos regionales o comunales) combinados con series de tiempo donde se quiere estudiar por ejemplo el crecimiento económico o la tasa de delito a través de variables como la inversión, el trabajo y otros factores sociales y ambientales. Es posible que se presente un proceso heterocedastico por el hecho no sólo de mayor variabilidad en las regiones grandes sino por la (s) variable (s) que se utilice (n) como deflactor de las series (la población de cada región y el índice de precios).

4.3. Aprendizaje de los errores

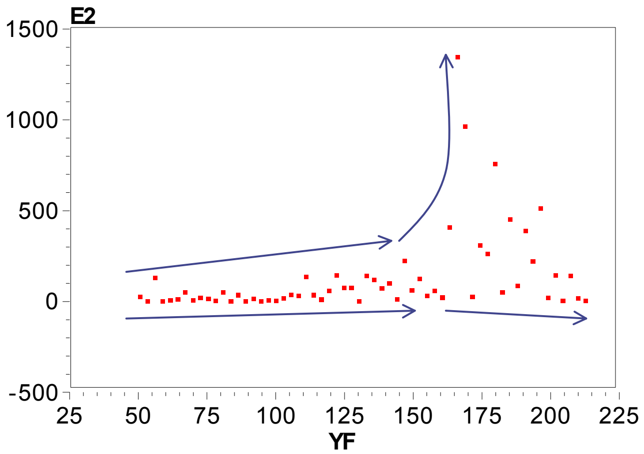
A medida que la gente aprende (por ejemplo, en los procesos productivos en cadena o en la evaluación de la compra) con el tiempo, sus errores de comportamiento se reducen y por tanto se espera que la varianza se reduzca. Sí se piensa en una línea de producción de un bien, las primeras unidades producidas tendrán un mayor costo o un mayor tiempo que las unidades producidas al final del periodo, la experiencia en la practica reduce el tiempo y el costo promedio así como el número de unidades defectuosas.

4.4. Menores restricciones de elección.

En promedio, los hogares con mayores ingresos poseen mayores posibilidades u opciones de elegir respecto a sus decisiones de compra, respecto a los hogares de menores ingresos. En los hogares de mayores ingresos su comportamiento puede variar desde las familias más austeras hasta las familias más dispendiosas. En cambio en los hogares de menores ingresos sólo podrían cubrir lo necesario de su canasta básica, sobre todo si existen restricciones al crédito intertemporal.

Figura 3

Método Grafico



En el ejemplo de ingreso y consumo que se presenta en la figura 3. Aquí se ha construido un grafico con el error al cuadrado en el eje de las ordenadas y el valor estimado del consumo en el eje de abscisas. Lo primero que se observa es un comportamiento de mayor dispersión a medida que aumenta el nivel de ingreso de los hogares, abriéndose al final la figura en forma de abanico.

Al igual que los hogares, las empresas con mayores ganancias pueden tener una mayor variabilidad en su comportamiento en cuanto a sus políticas de dividendo, las empresas con una política de crecimiento tendran una mayor variabilidad en sus tasas de crecimiento que empresas más conservadoras que prefieren mantener sus posiciones.

4.5. Mejoras en la Información

En la medida que mejoran los sistemas de procesos y de recolección de datos, además de las técnicas para explorar y utilizar con mayor eficiencia los mismo, es posible que se reduzca las variabilidades de las perturbaciones en las estimaciones.

4.6. Observaciones Influyentes

Son observaciones que caen fuera de las pautas generales del conjunto de datos o que ejercen una fuerte influencia en los resultados de la regresión. Si no existen errores de diseño, ni de recolección de los datos, estas representan los elementos diferenciadores del conjunto de datos y por tanto no deben ser excluida aunque la inclusión o exclusión de esta (estas) observación (observaciones) puedan alterar sustancialmente los resultados del análisis de regresión, especialmente en los casos de muestras pequeñas. Estas observaciones deben identificarse para estudiar su impacto y considerar su posible causa, entre las cuales se debe evaluar las siguientes:

a) Un error en la entrada de observaciones o datos
b) Una observación valida pero excepcional que es explicable por una situación extraordinaria
c) Una observación excepcional sin una explicación plausible
d) Una observación ordinaria en sus características individuales pero excepcional en su combinación de características.

En el caso a) se debe corregir la observación o eliminar, en el caso b) se debe considerar en el modelo la causa extraordinaria valida, tal vez; a través de una variable ficticia (binaria), en el caso c) no se debe eliminar a pesar de que no se pueda justificar su inclusión y por último, en el caso d) indica que debe modificarse la base teórica o conceptual del modelo de regresión.

4.7. Problemas de Especificación

Este problema hace referencia a la inclusión de variables irrelevantes o a la omisión de variables relevantes o a la forma funcional elegida o a los errores de medida lo cual puede afectar la varianza.

Aunque la inclusión de variables irrelevantes no sesgue el resultado de las otras variables independientes, tiene cierto impacto sobre los resultados ya que, por un lado, reduce la parsimonia del modelo y por la otra, las variables adicionales pueden enmascarar o desplazar los efectos de variables más útiles y por último, hace que las contrastaciones de la significación estadística de las variables relevantes sea menos precisa en su significancia practica. En el caso de los errores de omisión son más difíciles de considerar ya que no se puede evaluar lo que no se ha incluido, sólo resta de contar con un soporte teórico y practico del asunto bajo estudio.

Cuando se elige una forma funcional inadecuada existe procedimientos de búsqueda pero un primer indicio lo da el análisis grafico de los residuos. Los errores de medida afectan también a las variables independientes ya que reduce su poder predictivo en la medida que aumenta el error de medida. El modelo de regresión no tiene métodos directos para corregirlos pero si se sospecha del problema se debe considerar la posibilidad de utilizar modelos estructurales o variables instruméntales.

En general, la heteroscedasticidad surge con mayor probabilidad en datos de corte transversal o en encuestas que en las series de tiempo ya que se tratan con una “población” en un “momento” a través de un “espacio” que pueden originar una variabilidad fuerte con un patrón sistemático que afecta la constancia de la varianza.

5) CONSECUENCIAS

Con presencia de heteroscedasticidad, los estimadores mínimo cuadrático seguirán siendo insesgados dado que **E(U)=0** y las variables explicativas son no estocásticas o independientes de los errores, por lo tanto **E(XU)=0**, pero dado que la varianza al no ser constante se invalidan las pruebas de significancia estadística en especial las pruebas de la distribución t-Student y F de Fisher-Snedecor.

En consecuencia, los principales problemas serían:

- Estimadores ineficientes.
- Perdida de validez de las pruebas de contraste estadísticos.

Dado que la estimación de la varianza estaría sesgada y su estimación no sería además consistente, no se puede, en consecuencia, depender de los resultados obtenidos de los intervalos de confianza, ni de las pruebas puntuales usuales, por tanto, si se detecta su presencia, se deben estimar los parámetros por otros medios más adecuados como; mínimos cuadrados ponderados u otra técnica de estimación. Siempre es necesario comprobar si existe o no un problema significativo antes de proceder a corregirlo. En algunos casos la comprensión de la forma de corregir permite comprender mejor la utilidad de los procedimientos formales de detección.

6) CORRECCIÓN

Se verá en principio, la forma de corregir el problema antes de su detección por cuanto ayudara comprender mejor los métodos de detección y su utilidad para corregir el posible problema. Para ello se introduce en la lógica del método de estimación por mínimos cuadrados generalizados factibles, para luego formular las medidas de solución de acuerdo a los supuestos sobre el patrón de comportamiento de los errores.

6.1. La Matriz de Transformación

Supóngase ahora que sé premultiplica el modelo descrito en la ecuación (3.2) por alguna matriz de transformación no singular T de orden *n x n* que cumpla los siguientes requisitos E(TU)=0 y $T\Omega T'=In^{2}$; (siendo In la matriz identidad) entonces sería posible que estimar el modelo

$$(6.1) \qquad TY=TX\beta+TU$$

Donde cada elemento del Vector TY será una combinación lineal de los elemento de Y, siendo lo mismo para los otros vectores, con esta transformación si se utiliza los MCO se obtendría estimaciones consistentes de los parámetros y de la matriz de varianzas como

$$(6.2) \qquad b_g= (X'T'TX)^{-1}X'T'TY$$

$$(6.3) \qquad Var(b_g)=\sigma^2(X'T'TX)^{-1}$$

El estimador b_g se define como el estimador por mínimos cuadrados generalizados (MCG) o de Aitken (1935), obviamente el procedimiento es simple si se conoce cual es la transformación adecuada, lo que ocurre en casos muy especiales.

6.2. Medidas remédiales desconociendo la especificación de la varianza

Para entender un poco más este punto se va a ejemplificar con un modelo simple como el de (2.1) y que será extensible al modelo matricial además permitirá comprender los métodos de detección. Sea el modelo con un patrón de heterocedasticidad como el descrito en (3.6), $\sigma_i^2 = \sigma^2\,\omega_i$.

$$(6.4) \qquad Y_i = \alpha + \beta\,X_i + u_i$$

[2] Nótese que si $T\Omega T'=In$ entonces $\Omega^{-1}=T'T$

La corrección se haría por un caso especial de los mínimos cuadrados generalizados como son los mínimos cuadrados ponderados, donde cada uno de los vectores del modelo sería multiplicado por un ponderador $w_i=(1/\sigma_i)$ lo cual permitiría obtener una varianza constante libre de los efectos del patrón heterocedástico.

$$(6.5) \qquad (Y_i /\sigma_i) = (\alpha/\sigma_i) + (\beta X_i /\sigma_i) + (u_i /\sigma_i)$$

Es decir si el patrón es $\sigma^2_i=\sigma^2\omega_i$ y por lo tanto la transformación adecuada sería entonces la raíz cuadrada de la parte variable ($\sqrt{\omega_i}$ $_i$).

$$(6.6) \qquad (Y_i /\sqrt{\omega_i}\,) = (\alpha /\sqrt{\omega_i}\,) + (\beta X_i /\sqrt{\omega_i}\,) + (u_i /\sqrt{\omega_i}\,)$$

Esto permitiría obtener una varianza homocedastica que cumpliría los supuestos y permitiría la inferencia estadística y realizar las pruebas de los intervalos de confianza. Con la transformación de las variables ahora la varianza sería.

$$(6.7) \qquad E(u_i /\sqrt{\omega_i}\,)^2 = (1/\omega_i)E(u_i)^{\,2} = (1/\omega_i)\,\sigma^2\omega_i = \sigma^2$$

Si se desconoce la varianza y su patrón de cambio, tal como sucede en la mayoría de los casos, entonces deberá formularse algún supuesto sobre su comportamiento. Dependiendo de los supuestos que se adopten, corresponderá una transformación para estimar por mínimos cuadrados ponderados, por ello se asumirán varios supuestos razonables para ejemplificar el proceso de corrección.

Algunos textos recomiendan que a priori se transforme el modelo en términos logarítmicos ya que con ello reduce el nivel de dispersión y comprime la escala. Como señala Gujarati (1997), si 80 es 10 veces el numero 8 el ln 80(=4,3280) es casi el doble del ln 8(=2,0794). Pero en muchos casos si bien tiene ventajas no siempre es posible aplicar esta transformación ya que la estructura de datos o del modelo no lo permite o no es adecuada.

Supuesto 1

Se asume que la varianza del error es proporcional al cuadrado de la variable explicativa; $E(u_i)^2 = \sigma^2 X_i^2$.

Se utiliza la siguiente transformación:

$$\frac{Y_i}{X_i} = \frac{\alpha}{X_i} + \beta + \frac{u_i}{X_i} \quad (6.8)$$

$$= \frac{\alpha}{X_i} + \beta + \upsilon_i \quad (6.9)$$

Donde (ypsilon) v es él término del error transformado que cumple todas propiedades para los estimadores mínimo cuadrático (son idénticos e igualmente distribuidos iid).

Entonces se tiene que: $E(v_i)^2 = E(u_i / X_i)^2 = (1/X_i^2)E(u_i)^2 = (1/X_i^2)\,\sigma^2 X_i^2 = \sigma^2$. Por tanto, la varianza de v_i es homocedástica.

Supuesto 2

Se asume que la varianza del error es proporcional a la variable explicativa; $E(u_i)^2 = \sigma^2 X_i$. Para ello se utiliza la siguiente transformación:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\alpha}{\sqrt{X_i}} + \beta\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} \quad (6.10)$$

$$= \frac{\alpha}{\sqrt{X_i}} + \beta\sqrt{X_i} + \upsilon_i \quad (6.11)$$

Ahora se tiene que la varianza de ypsilon sería: $E(v_i)^2 = E(u_i / \sqrt{X_i})^2 = (1/\sqrt{X_i})E(u_i)^2 = (1/X_i)\,\sigma^2 X_i = \sigma^2$. Por tanto, la varianza de v_i es homocedástica.

Supuesto 3

Se asume que la varianza del error es proporcional al cuadrado del valor esperado de la variable explicada; $E(u_i)^2 = \sigma^2 E[Y_i]^2$.la transformación sería ahora:

$$\frac{Y_i}{E[Y_i]} = \frac{\alpha}{E[Y_i]} + \frac{\beta X_i}{E[Y_i]} + \frac{u_i}{E[Y_i]} \qquad (6.12)$$

$$= \frac{\alpha}{E[Y_i]} + \frac{\beta X_i}{E[Y_i]} + \upsilon_i \qquad (6.13)$$

La varianza de v_i es homocedástica como se ve, $E(v_i)^2 = E(u_i /E[Y_i])^2 = \sigma^2$. Por tanto se satisface los supuestos clásicos y se puede realizar las pruebas de confianza de los parámetros estimados del modelo.

Como es obvio no se cuenta con el valor medio esperado sino con los valores estimados ($\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$) y a ellos se recurre como sustituto de $E[Y_i]$. Para ello se procede en dos etapas. Primero se estima el modelo por MCO, sin considerar el problema de heteroscedasticidad y se obtiene los $\hat{Y}_i$ para ser utilizados en la segunda etapa. Luego se corre la regresión utilizando los $\hat{Y}_i$ como ponderadores y se transforma el modelo como.

$$\frac{Y_i}{\hat{Y}_i} = \frac{\alpha}{\hat{Y}_i} + \frac{\beta X_i}{\hat{Y}_i} + \frac{u_i}{\hat{Y}_i} \qquad (6.14)$$

$$= \frac{\alpha}{\hat{Y}_i} + \frac{\beta X_i}{\hat{Y}_i} + \upsilon_i \qquad (6.15)$$

De hecho una formulación más general (Kamenta,1971) es considerar un patrón que incluye dos parámetros, σ^2 y δ y una variable ω_i.

$$E(\upsilon_i)^2 = \sigma_i^2 = \sigma^2 \omega_i^{\delta} \qquad (6.16)$$

Donde el parámetro δ mide el grado de heteroscedasticidad, cuanto más pequeño sea menor será la diferencia entre las varianzas individuales y cuando $\delta=0$, el modelo sería homoscedástico. La variable (ω_i) sería cualquier variable que causa el problema, de hecho se recomienda utilizar el valor estimado de la variable dependiente, $\left(\hat{Y}_i\right)$, cuando hay más de una variable que causa el problema de heteroscedasticidad. Normalmente los valores de δ se buscan en un rango que va desde cero a dos ($0 \leq \delta \leq 2$).

6.3. Un ejemplo.

A continuación se muestra un ejemplo con el programa LIMDEP ©, en un estudio sobre bienes que no se tranza en el mercado[3], se estudiaba la disposición a pagar para proteger una colonia de delfines y que características hacían más propenso el pago por deducción en la cuenta del agua. La muestra final fue de 282 observaciones validas realizadas en la Ciudades de La Serena y Coquimbo de la IV Región de Chile, en el segundo semestre del 2000.

La variable dependiente (Y) es la disposición de pago en pesos chilenos a julio del 2000, y las variables explicativas del modelo eran las variables dicotómicas del conocimiento de la existencia de la colonia de delfines nariz de botella (CD; Si=1, No=0). Si había visitado la Isla Damas donde estaba la colonia de delfines nariz de botella (PVI; Si=1, No=0), el sexo del entrevistado (Sexo; hombre =1, mujer =0), la edad en años (Edad), los años de estudio (Educ) y el nivel de ingreso en pesos que declara en la entrevista (Ing).

En el resultado de la estimación inicial sólo resultaron tres parámetros significativos de siete. al 5% de error tipo I (95% de confianza), el resto (4 parámetros) no eran significativos a ese nivel. Sólo las variables de visita a la Isla Damas (PVI), Edad y Educación fueron significativas, las variables menos significativas fueron el ingreso y el conocimiento de la colonia de delfines (CD) y el sexo. El coeficiente de determinación ajustado llega al 7.4% de explicación de la variabilidad total. Las instrucciones listadas para la estimación fueron:

[3] Ramírez; D. & R Helo y C. Rivera (2000)"Evaluación de bienes ambientales que no se transan en el mercado: Caso delfines Nariz de Botellas Islas de Choros, IV Región". . Universidad Católica del Norte, Coquimbo, Chile.

```
Namelist   ; X=ONE, CD, PVI, SEXO, EDAD, EDUC, ING$
REGRESS    ; LHS=Y ;RHS= X; Keep=BX ;res=ER0$
```

La salida del primer listado de instrucciones fue;

```
+-----------------------------------------------------------------------+
| Ordinary    least squares regression    Weighting variable = none     |
| Dep. var. = Y        Mean=   651.5957447    , S.D.=    851.1699754    |
| Model size: Observations =     282, Parameters =   7, Deg.Fr.=    275 |
| Residuals:  Sum of squares= 184453448.0    , Std.Dev.=      818.98706 |
| Fit:        R-squared=  .093959, Adjusted R-squared =         .07419 |
| Model test: F[  6,    275] =    4.75,    Prob value =         .00013 |
| Diagnostic: Log-L =  -2288.2718, Restricted(b=0) Log-L =   -2302.1843 |
|             LogAmemiyaPrCrt.=   13.441, Akaike Info. Crt.=     16.279 |
| Autocorrel: Durbin-Watson Statistic =   1.99210,   Rho =       .00395 |
+-----------------------------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 Constant  252.9661444      263.82111         .959   .3385
 CD        17.85207305      104.16576         .171   .8640   .62765957
 PVI       349.2599848      125.90186        2.774   .0059   .20921986
 SEXO      149.0031948      102.13755        1.459   .1457   .56028369
 EDAD     -9.932205411      4.2599321       -2.332   .0204   38.911348
 EDUC      41.25347989      15.158500        2.721   .0069   13.202128
 ING       .2817488262E-03  .23509862E-03    1.198   .2318   258067.38
```

Significativos
al 5%

Se realizaron pruebas que detectaron el patrón de heteroscedasticidad y luego de comprobado el problema, la transformación elegida fue $E(\upsilon_i)^2 = \sigma_i^2 = \sigma^2 \hat{Y}_i^2$, con dicha corrección se redujo el problema de heteroscedasticidad.

Las instrucciones para la corrección se muestran a continuación con los resultados de la salida correspondiente

```
CREATE     ; W = 1 / BX ^ 2 $
REGRESS    ;     Lhs = Y ;Rhs = X ; Wts = W ;res=E; Keep = BX $
```

```
+--------------------------------------------------------------------------+
| Ordinary    least squares regression    Weighting variable = W          |
| Dep. var. = Y        Mean=   355.0380144    , S.D.=   405.6342111       |
| Model size: Observations =     282, Parameters =   7, Deg.Fr.=    275   |
| Residuals:  Sum of squares= 21594882.94    , Std.Dev.=      280.22642  |
| Fit:        R-squared=  .532937, Adjusted R-squared =          .52275  |
| Model test: F[  6,    275] =   52.30,    Prob value =          .00000  |
| Diagnostic: Log-L =  -2344.3117, Restricted(b=0) Log-L =   -2451.6538  |
|             LogAmemiyaPrCrt.=   11.296, Akaike Info. Crt.=     16.676  |
| Autocorrel: Durbin-Watson Statistic =   1.80280,   Rho =       .09860  |
+--------------------------------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 Constant  1030.359296      157.12548        6.558   .0000
 CD         149.9108487      70.380422        2.130   .0341  .86155350
 PVI        404.6595845     164.11495         2.466   .0143  .10905988E-01
 SEXO      -268.8388319      51.654338       -5.205   .0000  .39036270
 EDAD      -16.02714786      2.3355126       -6.862   .0000  48.403722
 EDUC       32.94264757      8.5668900        3.845   .0001  5.2204400
 ING       -.1163879946E-02  .27470142E-03   -4.237   .0000  86046.301
```

En esta última salida puede verse que todos los parámetros eran significativos al 95% de confianza (5% de error tipo uno) y como puede verse la varianza estimada, comparando con la primera salida, obtenida con la estimación corregida de los regresores fue en unos casos sobrestimada o subestimada en la primera salida y el signo del parámetro de ingreso cambio. Todos los parámetros son estadísticamente significativos al 95% de confianza. El coeficiente de determinación ahora explica el 52,3% lo cual es sustancialmente mayor al 7,4% anterior de la estimación no corregida.

7) DETECCIÓN

Anteriormente se mostró, en el ejemplo de la figura 2, la relación entre el consumo e ingreso con presencia de heteroscedasticidad y se dijo que si se ordenan las variables de mayor a menor ingreso, la varianza del grupo de mayor ingreso presentaba una dispersión mayor de la línea de regresión que el grupo de menor ingreso. Intuitivamente la detección pasa por demostrar que las varianzas no son iguales. Para ello se utilizan métodos informales (gráficos e intuitivos) y métodos formales o comprobatorios, en estos últimos existen dos tipos de procedimientos a saber; constructivos y no constructivos (Greene, 1999). Los métodos no constructivos sólo pretenden identificar la existencia del problema y los constructivos pretenden identificar el posible patrón del proceso heterocedastico.

Diagrama 1
Métodos Detección

$$\text{Métodos}\begin{cases}\text{Gráficos e intuitivos}\begin{cases}\text{Ex - ante}\\ \text{Ex - post}\end{cases}\\ \text{Comprobato rios}\begin{cases}\text{No Constructi vos}\\ \text{Constructi vos}\end{cases}\end{cases}$$

Las pruebas no constructivas poseen una mayor potencia estadística que los métodos constructivos pero tienen la desventaja de no describir el tipo de proceso el cual debe ser extraído para resolver el problema de los datos.

Se ejemplificara los diversos métodos con un ejemplo sencillo de ingreso consumo. Las instrucciones listadas son realizadas con el programa[4] Limdep 7.0, y también se incluyen ejemplos de pruebas con el programa[5] Eviews 3.1.

7.1. Métodos Informales

Estos métodos se sustentan en desarrollar un buen “sentido común” o “intuición educada” (*educated guess-work*). Este tipo de intuición permite tener evidencias pero no pruebas. En principio se puede “sospechar” que el tipo de datos, pueden presentar un problema de heterocedasticidad dado las experiencias de estudios anteriores sobre el tema.

[4] LIMDEP 7.0 © Es una marca registrada de W. H. Greene.
[5] Eviews 3.1 © Es una marca registrada de QSM.

El supuesto fundamental del análisis multivariante es la normalidad de los datos ya que es el punto de referencia de los métodos estadísticos utilizados. Si la variación respecto de la distribución normal es bastante amplia, las pruebas estadísticas de contraste no son validas, puesto se requiere la normalidad para las pruebas t y F.

Si una variable es normal multivariante, es también será una normal univariante, aunque lo contrario no es necesariamente cierto, es decir una combinación de dos o más variables normales univariantes no conforman necesariamente una normal multivariante pero en una situación donde todas las variables exhiben una normal univariante ayudaría a obtener una variable normal multivariante aunque no lo garantiza.

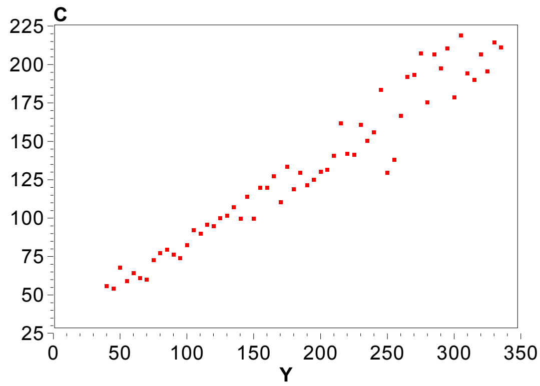
La normalidad multivariante es más difícil de contrastar, por eso en este documento se centrará en el análisis de la normalidad univariante de los residuos que son la expresión de la salud del modelo.

7.1.1. Análisis Ex- ante

Antes de realizar la regresión se deben analizar los datos y ver cuan cercanos se encuentran del supuesto de normalidad uno de las pruebas es realizar los análisis gráficos, el primero fue el que se mostro en la figura 2 donde deben realizarse para cada una de las variables explicativas contra la variable dependiente.

Como se puede observa en el gráfico, que sigue al listado de instrucciones, la dispersión aumenta a medida que aumenta la renta, nótese que este gráfico es el mismo que mostró en la figura 2.

```
?Grafico de dispersion de los datos
PLOT;Lhs=Y;Rhs=C$
```



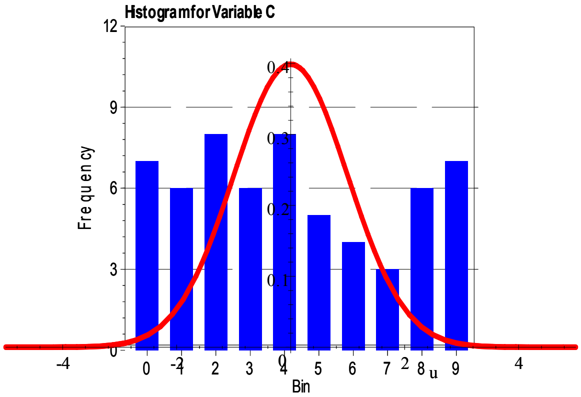
Estos gráficos son llamados diagramas de dispersión y permite ver como es el comportamiento de la nube de puntos. Como se muestra, la primera línea de instrucción es un comentario y la segunda línea es el comando de graficación (Plot) que realiza la nube de puntos del consumo (Rhs=C) contra el ingreso (Lhs=Y). Es relevante tener cuidado con la escala ya que esta puede hacer más o menos clara la visualización del comportamiento de los datos.

Otros análisis, ex–ante, son: el *gráfico de normalidad* a través del *histograma* y los *gráficos QQ de Normalidad*[6]. Los histogramas son más simple y ellos deberían mostrar la típica forma de campana de la distribución de Gauss (gaussiana o normal).

Aquí se construye para el consumo dividiendo los datos en 10 proporciones iguales. La instrucción para construir el histograma y la salida correspondiente se listan a continuación. Nótese que presentan una distribución asimétrica.

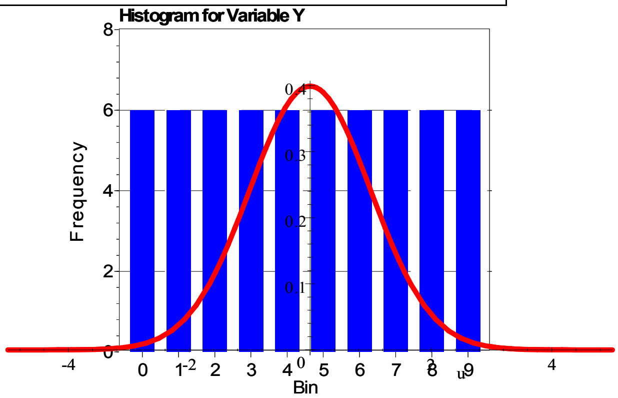
```
? Histograma del consumo
HISTOGRAM; Rhs =C ; Int =10$
```

[6] Para el uso de esto diagramas de normalidad se recomienda revisar la obra de Hair, Anderson, Tatham y Black, (1999). Estos diagramas hoy en día se encuentran en todos los programas de uso estadístico como el SPSS ©, ademas del LIMDEP 7.0 © y el Eviews 3.1©



Para el caso del histograma del ingreso se utiliza la misma instrucción cambiado solo la variable de consumo (C) por la del ingreso (Y) manteniendo los 10 intervalos, la instrucción listada y su salida se muestran seguidamente a continuación. Se puede observar que los datos son tomados en el mismo tamaño para los distintos niveles de ingreso y por tanto presenta una distribución uniforme

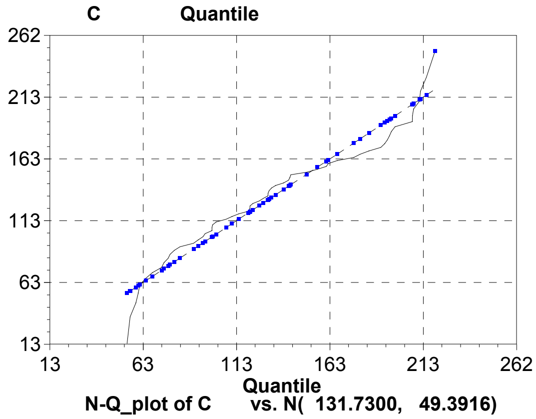
```
? Histograma del ingreso
HISTOGRAM; Rhs =Y ; Int =10$
```

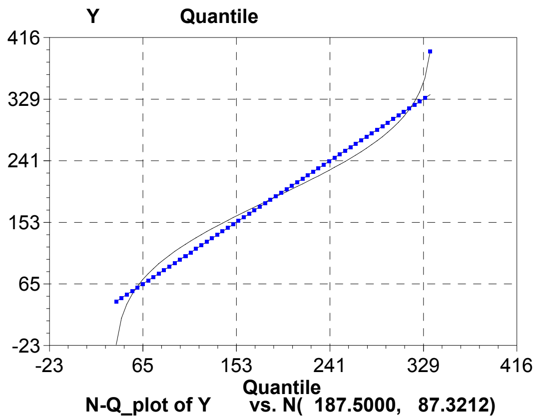


Se puede observar que ninguno de los dos histogramas posee la típica distribución gaussiana, el primero, el consumo, presenta una distribución asimétrica positiva no puntiaguda y el segundo, el ingreso, una distribución tipo uniforme. Esto se confirma con los gráficos QQ de Normalidad para ello se lista la siguiente instrucción (Dstat;) que arroja tanto los gráficos QQ-N, como la estadísticas descriptivas básicas (media, desviación, asimetría o

Skewness, Kurtosis o concentración, valores máximos, mínimos y total de observaciones).

```
DSTAT; Rhs = C, Y; Plot; Output=3$
```





Arriba, el primer gráfico de la izquierda del lector, muestra el diagrama QQ-N para el consumo y el segundo a la derecha, el QQ-N para el ingreso. Los graficos QQ-N confirman lo mostrado por el histograma individual realizado. Este procedimiento trata de averiguar si los datos proceden de una distribución normal y para ello se traza un diagrama con los «*Quantiles reales»* observados versus los «*Quantiles teóricos»* de una *Normal* (QQ-N). Los «Quantiles» dividen la muestra en una serie de grupos de igual tamaño. Si los puntos están más próximos a la recta teórica el ajuste es aceptable y la serie sería normal y lo contrario, si es al revés.

También es posible realizar una prueba estadística de normalidad a través de los coeficientes de asimetría y concentración. El valor estadístico de la normal estandarizada (z) para la asimetría se calcula como: $z_{asimetría} = \frac{Skewness}{\sqrt{6/N}}$. Donde N es el tamaño de la muestra. También se puede obtener un valor de la normal estandarizada o z calculado para la Kurtosis como: $z_{curtosis} = \frac{Kurtosis}{\sqrt{24/N}}$. Si el valor de la z calculada excede al valor crítico de tabla, entonces se rechaza la hipótesis nula de que la distribución es normal para esas características.

Por último, la prueba estadística de normalidad de los datos de Jarque-Bera[7] (JB) está dada por; $JB = \frac{N-K}{6}\left[Skewness^2 + \frac{1}{4}(Kurtosis-3)^2\right]$. Donde N es él número de observaciones, K es cero si es una variable que se estudia y sí se analizan los residuos es igual al numero de variables explicativas de la regresión. El estadístico JB se distribuye como una Chi cuadrado (χ^2) con dos grados de libertad bajo la hipótesis nula de normalidad.

7.1.2. Análisis Ex-post

El análisis expost se realiza con los residuos de la regresión realizada. Para ello se construye un tipo de prueba grafica a partir de los errores estimados de la regresión elevados al cuadrado contra la(s) variable(s) sobre la que se "sospecha" o la variable dependiente estimada. Como se muestra en la figura 3, en la cual se muestra una abertura tipo abanico de los datos.

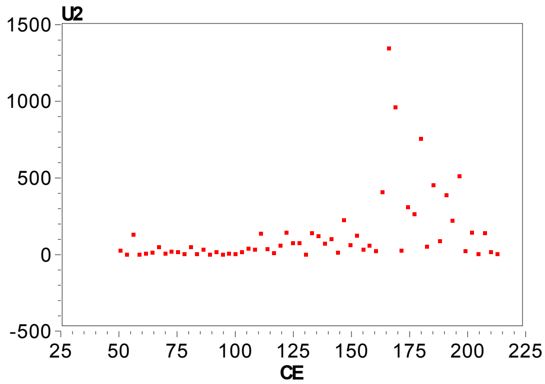
El procedimiento a realizar es el siguiente; se estima la regresión inicial por MCO, guardando los errores estimados en una variable (Res=U) y los valores estimados del consumo (Keep=CE) en otra. Luego utilizando el comando de transformación (Create) se elevan al cuadrado los errores

[7] Jarque, C. and A. Bera (1980)

multiplicando elemento del vector por sí mismo y creando un nuevo vector columna de perturbaciones cuadráticas (U2=U*U). Por último se trazan los errores al cuadrado (U2) en la ordenada y el consumo estimado (CE) en el eje de abscisas.

Ejemplo listado de instrucciones en LIMDEP© y la salida se muestran a continuación:

```
/*Se realiza la primera regresión del modelo guardando
 los errores en el vector u y los valores estimados
 en el vector CE */
Regress; Lhs=C; Rhs =one, Y ; Res =U; Keep=CE$
? Se construye el vector de errores al cuadrado
Create; U2 = U*U $
? Se grafican los residuos
PLOT; Lhs=C; Rhs=U2$
```



Otros gráficos similares pueden ser construidos tomando los errores estimados originales o los valores absolutos de los mismos contra las variables dependiente e independientes de la regresión estimada. En la sección siguiente se verán los métodos más formales.

7.2. Métodos Formales

Como se señalo, en los métodos formales se trabaja con dos tipos: procedimientos "*No Constructivos*" y procedimientos "*Constructivos*". Los primeros presentan una mayor potencia estadística y en algunos casos no son pruebas exigentes respecto al supuesto de normalidad. Los segundos poseen una menor potencia estadística en cuanto no rechazar la hipótesis nula siendo esta falsa, por lo cual pueden arrojar resultados dudosos pero pueden mostrar el posible patrón de corrección del modelo. No se debe olvidar que todas las pruebas de contraste se construyen bajo la hipótesis nula de homocedasticida versus la hipótesis nula de no homocedasticidad, esto, aun cuando en muchos casos es obvio, por ser tan obvio se pasa por alto cuando se quiere concluir en los resultados. Se verán primero los "No Constructivos" y luego los "Constructivos"

7.2.1 Métodos No Constructivos

En los métodos no constructivos se verán cuatro pruebas, la de Levene (1960), de Goldfeld-Quandt (1965 y 1972), la de Breusch-Pagan (1979) - Godfrey (1978) y, finalmente, la prueba de White (1980).

7.2.1.1 Prueba de Levene

La prueba de Levene[8](1960) o de homogeneidad de varianza, contrasta hasta que punto los distintos niveles del factor tiene una varianza homogénea en la variable dependiente. Tiene la ventaja de no ser tan exigente respecto a la distribución de normalidad de la variable dependiente. En este caso la variable dependiente será el error al cuadrado[9] y la variable por la cual se categoriza (factor) es la variable dependiente o cualquiera de las variables explicativas de la regresión. El valor se obtiene calculando para cada sujeto la diferencia en valores absolutos entre la puntuación y la media del grupo o nivel del factor y realizando después un ANOVA con estas diferencias. Se contrasta con una prueba F bajo la hipótesis nula de Homoscedastidad.

[8] Esta prueba se encuentra en el menú de varios programas especialmente en SPSS© y en el Eviews©. También existe otra prueba de contraste de homogeneidad de la varianza de Kendall y Stuart (1961) descrita por Johnston (1987) es parecida a la prueba de Levene.
[9] También se puede utilizar para la variable original o transformada en valor absoluto u otra transformación monótona si lo que se quiere probar es la varianza de la misma y no del error.

Aquí se ejemplifica con una salida realizada en el programa Eviews © con los datos del ejemplo de consumo, la salida mostrada a continuación indica la presencia de heteroscedasticidad. En la figura adjunta (Tabla1) se muestra el resultado obtenido a través del programa Eviews© 3.1.

Views ➔Test for Descriptive Stat ➔Equality Tests by Classification ➔Ing

Tabla 1

Prueba de Levene

Test for Equality of Variances of E2
Categorized by values of ING
Date: 06/02/05 Time: 20:15
Sample: 1 60
Included observations: 60

$F^{5\%}_{tabla(3,56)} = 2.769$

Method	df	Value	Probability
Bartlett	3	22.05070	0.0001
Levene	(3, 56)	7.952121	0.0002
Brown-Forsythe	(3, 56)	5.481001	0.0023

Category Statistics

ING	Count	Std. Dev.	Mean Abs. Mean Diff.	Mean Abs. Median Diff.
[0, 100)	12	5.397101	4.107516	4.107516
[100, 200)	20	7.117652	5.907751	5.642876
[200, 300)	20	17.26311	14.24416	13.82843
[300, 400)	8	12.13553	8.764134	8.764134
All	60	11.73481	8.707357	8.480490

Bartlett weighted standard deviation: 11.93446

En el caso del programa Eviews© el procedimiento es el siguiente: Se selecciona la variable a estudiar (en el ejemplo, E2), la prueba se encuentra en el menú "Views" submenú "Test for Descriptive Stat" y se elige la opción "Equality Tests by Classification" y en la ventana se pone el nombre de la serie con la que se construye los intervalos (en el ejemplo; ING). Los grados de libertad son el número de factores menos uno en el numerador (En este caso 4 intervalos de la variable ingreso menos 1 grado de libertad por calculo de la media =3) y en el denominador, el numero de observaciones menos el numero de factores (En este caso 60 observaciones menos 4 intervalos =56). Si el F calculado es mayor que el F de tabla se rechaza la hipótesis nula. En este caso el estadístico de Levene es 7.9521 que es mayor al valor en tabla al 5% de error tipo uno (95% de confianza) que es de 2.7694.

7.2.1.2 Goldfeld-Quandt (G-Q)

La prueba de Goldfeld y Quandt (1965 y 1972) se utiliza para muestras pequeñas. Para realizar esta prueba se debe, *primero* ordenar toda la muestra en orden decreciente, de mayor a menor, utilizando la variable dependiente como variable de ordenación, también se puede utilizar una de las variables explicativas que presenta un comportamiento claramente creciente; como el ingreso en el consumo de los hogares.

En *segundo* lugar, se divide la muestra en tres submuestras, la muestra del medio debe ser menor igual a un tercio y mayor a un quince por ciento de la muestra total. Las submuestras extremas (la superior e inferior) deben ser mas o menos iguales (esto es deseable mas no es necesario). La submuestra superior corresponde a los valores grandes y la submuestra inferior corresponde a los valores pequeños de la regresión.

Luego se estima la regresión de la submuestra superior (omitiendo el resto de las observaciones) y se obtienen los residuos, con ellos se calcula la suma cuadrada de los residuos superior o primer grupo (SCR1), se procede a realizar lo mismo, pero para la submuestra inferior, o segundo grupo y se obtienen los residuos del grupo inferior y se calcula la suma cuadrada de los residuos del segundo grupo (SCR2). Bajo la hipótesis nula de homocedasticidad se realiza una prueba contraste F, con m_1-k_1 grados de libertad en el numerador y m_2-k_2 grados de libertad en el denominador

$$Ho: \sigma_1=\sigma_2$$

$$Ho: \sigma_1\neq\sigma_2$$

(7.1) $$F_C = \frac{SCR1/(m1-k1)}{SCR2/(m2-k2)} \rightarrow F^{\alpha}_{m1-k1;m2-k2}$$

Donde m_1; es el número de observaciones de la submuestra superior, k_1; es el número de parámetros de la submuestra superior, m_2; es el número de observaciones de la submuestra inferior, k_2; es el número de parámetros de la submuestra inferior. Si el F calculado es mayor al valor del F de tablas se rechaza la hipótesis nula de homocedasticidad.

Ejemplo de la prueba G-Q

El listado de instrucciones para LIMDEP©

```
? Ordena las variables de mayor a menor
SORT ; LHS=Y; RHS=C ;DESCENDING$
?Toma los valores de submuestra superior
SAMPLE; 1-25$
? Estima la regression y guarda los residuos
REGRESS; LHS=C; RHS=ONE, Y
      ; RES=E1; KEEP= C1$
? Guarda el numero de grados de libertad de la regresion
Calc; list; DG1=DEGFRDM $
? Obtiene el vector de residuos al cuadrado
Create; E21=E1*E1$
? Obtiene la suma de los residuos al cuadrado
Calculate; list; SCR1=Sum(E21) $
? Toma la submuestra inferior
SAMPLE; 35-60$
? Estima la regression y guarda los residuos
REGRESS; LHS=C; RHS=ONE, Y
      ; RES=E2; KEEP= C1$
? Guarda el numero de grados de libertad de la regresion
Calc; List; DG2=DEGFRDM $
? Obtiene el vector de residuos al cuadrado
Create; E22=E2*E2$
? Obtiene el vector de residuos al cuadrado
Calculate; List; SCR2=Sum(E22) $
/* Calcula el F para la prueba de G-Q */
Calculate; List; FCalc=   (SCR1/DG1)/(SCR2/DG2)$
/* Arroja el valor de tabla de la F con 5% de error tipo I*/
Calc; List; Ftb(.95,DG1,DG2)$
```

El resultado final arroja un F calculado de 10,34 mayor al 1,99 de tabla.

```
Calculate; List; FCalc=   (SCR1/DG1)/(SCR2/DG2)$
    FCalc    =  .10340607254351180D+02
Calc; List; Ftb(.95,DG1,DG2)$
    Result  =  .19932391349600010D+01
```

En algunos casos esta prueba de contraste no arroja resultados claros, ya que puede ser afectada por la forma funcional del modelo y por el tipo de distribución de los datos[10]. Por ejemplo, en un modelo con transformaciones logarítmicas en las variables, la prueba de G-Q puede arrojar como resultado el no rechazo de la hipótesis nula a pesar del que los datos presenten un patrón de heterocedasticidad, por esta razón se utilizan, como pruebas de mayor potencia estadística, las pruebas de Breusch-Pagan (1979) y Godfrey (1978) que no dependen de la forma funcional y la prueba de White que es independiente de la forma funcional y del tipo de distribución de los datos.

7.2.1.3 Breusch-Pagan-Godfrey (BPG)

En los trabajos de Breusch y Pagan (1979) y en el de Godfrey (1978) se construyo una prueba que no depende del numero de observaciones omitidas, ni de la identificación de la variable que genera la heteroscedasticidad para realizar el ordenamiento, ni de la forma funcional. Esta prueba de contraste cubre un amplio espectro, es simple y se basa en los residuos MCO. Se asume para ello el modelo

$$(7.2) \qquad Y=X\beta+U$$

Donde las perturbaciones u_i se distribuyen como una normal, con una varianza no constante.

$$(7.3) \qquad \sigma_i^2 = h(z_i\theta)$$

Donde h(.) es una forma funcional cualquiera no especificada, θ (tetha) es un vector de coeficientes (px1) no relacionados con los β y z_i es un vector de variables (px1) que presumiblemente causan la heteroscedasticidad, donde p es el numero de

[10] Véase a Johsnton (1989) quien cita a Harvey y Phillips (1973) quienes realizaron experimentos para validar la robustez y potencia de las pruebas de contraste del supuesto de Homocedasticidad.

regresores o variables que se incluyen en la regresión auxiliar. Se asume que él primer elemento es igual a uno y por tanto la hipótesis nula de homoscedasticidad es

$$Ho: \theta_2 = \theta_3 = \theta_4 = \ldots = \theta_p = 0$$

Las variables z_i pueden ser una o todas las variables de la regresión que pueden afectar estructuralmente a la varianza de la regresión. La prueba se construye de la siguiente forma.

i) Estime el modelo Y=Xβ+U (6.2) y obtenga los residuos MCO estimados $\hat{u}_i$ $(\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_N)$.

ii) Calcule la varianza estimada como; $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N}$; (nótese que este es el estimador MV y no el estimador MCO de la varianza).

iii) Construya la variable (xi) ξ_i como; $\xi_i = \frac{\hat{u}_i^2}{\hat{\sigma}^2}$; véase que es el residuo estimado MCO elevado al cuadrado y dividido por la varianza estimada.

iv) Estime la regresión; $\xi_i = \theta_1 + \theta_2 z_2 + \cdots + \theta_p z_p + \upsilon_i$; donde (ypsilon) υ_i; es un error normal con media cero y varianza constante.

v) Calcule la Suma Cuadrada Explicada (SCE) de la regresión auxiliar y calcúlese; $\Theta = (½)$ SCE.

vi) Bajo la hipótesis nula de homoscedasticidad el estadístico $\Theta \sim \chi^2_{p-1}$; se distribuye (asintoticamente) como una *Chi* cuadrada con p-1 grados de libertad, donde p es el numero de variables incluidas en la regresión auxiliar.

vii) Si el estadístico calculado es mayor al valor de tabla se rechaza la hipótesis nula.

Ejemplo del listado de instrucciones para los datos de consumo-ingreso

```
/* Prueba de Breusch Pagan y Godfrey */
? Estima la regresión inicial
     Regress; Lhs=C; Rhs=One, Y ; Res=U;Keep=CE$ BPG
? Crea el Vector de errores al cuadrado
     Create; U2=U*U$
? Calcula la suma de los errores al cuadrado
     Calc; List; SU2=sum(U2)$
? Calcula la varianza de la prueba
     Calc; List; VMV=SU2/NREG$
? Calcula el vector de la variable dependiente
```

```
     Create; GI=U2/VMV$
? Estima la regresión auxiliar
     REGRES; LHS=GI ; RHS= ONE, Y; Res=e; Keep=GIE$
? Calcula la Suma de cuadrados explicada
     Matrix; List; XCX=GIE'*GIE$
          Calc; List; SCE=XCX-NGI$
? Estima el Estadístico
     CALC; List; TBPG=0.5*(SCE)$
? Calcula el valor en Tabla
     Calc; List; Ctb(.95, KREG-1)$
```

Con la muestra de 60 datos sobre consumo-ingreso del ejemplo, se obtiene que la suma cuadrada explicada es de: SCE=68.63 y el estadístico de Breusch-Pagan y Godfrey es: Θ=34.315; para una Chi cuadrada de tabla con un grado de libertad y 95% de confianza (χ^2_1=3.84<Θ=34.315) se rechaza la hipótesis nula de Homocedasticidad. Es un grado de libertad ya que sólo hay una variable explicativa en la regresión auxiliar. Las salidas de las tres últimas líneas se muestran a continuación:

```
Calc; List; SCE=XCX-NGI$
    SCE      =  .68630387375018730D+02
CALC; List; TBPG=0.5*(SCE)$
    TBPG     =  .34315193687509360D+02
Calc; List; Ctb(.95, KREG-1)$
    Result   =  .38414591508300020D+01
```

7.2.1.4 Prueba de White.

White (1980) derivo la estimación de una matriz de varianzas y covarianzas consistente que provee una prueba de contraste de los coeficientes con presencia de un patrón no conocido de heteroscedasticidad, la matriz de Varianza-Covarianza de White esta dada por

$$\hat{\Sigma}_W = \frac{N}{N-k}\left(X'X\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i^T\right)\left(X'X\right)^{-1} \qquad (7.4)$$

Donde *N* es el número de observaciones, *k* es el número de regresores y los $\hat{u}_i^2$ son los residuos al cuadrado de la regresión estimado por MCO. El estimador convencional MCO homoscedástico es

$$V=\hat{\sigma}^2\left[X'X\right]^{-1} \quad (7.5)$$

Donde la varianza de los errores de la regresión por MCO ($\hat{\sigma}^2$) esta dada por;

$$\hat{\sigma}^2=\sum\hat{u}_i^2/(N-k) \quad (7.6)$$

Si no hay heteroscedasticidad (7.5) será un estimador consistente de la matriz de varianza y covarianza de los parámetros (Var(b)), Ahora, si existe, no lo será. Partiendo de esto White construyo una prueba de contraste a través de obtener el coeficiente de determinación (R^2) de una regresión auxiliar de los errores contra las variables explicativas del modelo original, sus cuadrados y sus productos cruzados[11], luego se construye un estadístico que se distribuye, como una Chi cuadrada con p-1 grados de libertad ($N*R^2 \sim \chi^2_{p-1}$)[12]. Donde p es él número de regresores de la regresión auxiliar sin incluir la constante. El estadístico provee una prueba de hipótesis de que todas las pendientes de la regresión auxiliar son cero, es decir;

$$Ho: \theta_1=\theta_2=\theta_3=\ldots=\theta_p=0$$

A su vez White señala que es una prueba general de especificación del modelo, dado que la hipótesis nula se fundamenta en que los errores son **homoscedásticos** e **independientes** de los regresores y que la **especificación lineal es la correcta**. Sí la hipótesis nula se rechaza, una o más de estas condiciones violadas conducen a una prueba significativa. De lo contrario, la ausencia de significancia estadística de la prueba, puede implicar que ninguna de las tres condiciones es violada.

La prueba de White se ejemplificará con un modelo de dos variables explicativas, la construcción de la prueba se realiza bajo la hipótesis nula de homoscedasticidad, independencia y linealidad. El procedimiento es el siguiente:

a) Se estima el modelo inicial; y se obtienen los errores del modelo estimado como
b) $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i$.

[11] Si el modelo tiene variables dummy y censuradas se omite los productos cruzados y algunos cuadrados.
[12] En el Eviews © se construye una prueba de White usando las distribuciones F y Chi.

c) Con los errores al cuadrado como variable dependiente y los cuadrados y los productos cruzados de las variables explicativas se utilizan en una regresión auxiliar como

d) $\hat{u}_i^2 = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \theta_3 X_{2i} X_{1i} + \theta_4 X_1^2 + \theta_5 X_2^2 + \nu_i$;

e) De esta regresión auxiliar se obtiene el coeficiente de determinación (R^2_{aux}) que se multiplica por él numero de observaciones (N) que se tiene en la muestra y se obtiene el estadístico de White como una Chi cuadrado con grados de libertad al número de regresores de la regresión auxiliar (exceptuando la constante). Para la ecuación del ejemplo (véase b y d) serían 5 grados de libertad entonces el Chi cuadrado calculado del estadístico de White (χ^2_{TW}) bajo la hipótesis nula sería:

f) $\chi^2_{TW} = N\text{-}R^2_{aux} \sim \chi^2_{(Kaux-1)}$

g) Si el Chi cuadrado calculado es mayor que el Chi cuadrado de tabla dado el nivel de *Error Tipo I*[13], se rechaza la hipotes nula de homoscedasticidad.

Al construir la prueba debe tenerse en cuenta si hay variables explicativas dicotómicas o "dummies" ya que los productos cruzados pueden crear una multicolinealidad artificial en la regresión auxiliar.

7.2.2 Métodos Constructivos

Ninguna de las pruebas anteriores provee una estimación especifica de la forma del proceso heteroscedástico que afecta al comportamiento de los errores y que se pueda insertar en el modelo y obtener estimadores Mínimo Cuadráticos Generalizados para ello se utilizan dos pruebas de contrastes de forma funcional del proceso heteroscedástico como las pruebas de Glesjer (1969) y de Park (1966).

[13] Por convención se utiliza el 5% o 1% de error tipo I esto arroja un 95% a un 99% de nivel de confianza-

7.2.2.1 Glejser

La prueba de contraste de Glejser explora la naturaleza de la heteroscedasticidad de manera muy cercana. Aquí se puede asumir supuestos más relajados que el de la varianza del error de perturbación o su desviación (σ), sea proporcional a alguna X_i y se puede, entonces, investigar formas funcionales más apropiadas como por ejemplo:

$$(7.7) \qquad \sigma_i = \alpha + \delta Z_i^{\gamma} + v_i$$

Donde ypsilon (v_i) es el término de perturbación estocástico que cumple todas las propiedades iid y Z_i es la variable que se sospecha que produce el patrón heteroscedástico, gamma (γ) es el parámetro que mide el grado de heteroscedasticidad y sigma (σ_i) es la desviación de cada error con respecto al modelo.

El procedimiento utilizado es el siguiente; se estima la regresión del modelo por MCO se obtiene los residuos, en su valor absoluto, $|u_i|$, como sustituto de σ_i, y se ajustan al modelo (7.7) asumiendo diferentes valores para cada gamma (γ). Esta prueba provee diferentes estimaciones para cada γ seleccionado. En cada caso la hipótesis nula de homoscedasticidad es rechazada sí el estimador *δ (delta)* es significativamente distinto de cero. Si más de un modelo tiene pendientes significativas la guía es seleccionar aquel que se ajusta mejor a la naturaleza de la heteroscedasticidad.

Ejemplo de la prueba de Glejser

Usando los datos del ejemplo se ajusta un modelo utilizando los valores de gamma (γ) de –1, –½, ½ y 1,5 ($-1 \leq \gamma \leq 1{,}5$) y se realizan cuatro regresiones que se resumen en la tabla 2, para los datos del ejemplo el mejor ajuste es para (γ=½) gamma igual a ½ dado que los valores de los criterios de Amemiya y Akaike son los menores y es el que tiene el mayor ajuste dado por el coeficiente de determinación, luego el patrón estimado seria: $\sigma_i = \alpha + \delta Z_i^{1/2}$

Tabla 2

Prueba de Glejser para valores de -1, -1/2 , 1/2 y 1.5

	Coeff.	t-ratio	R^2	Amemiya	Akaike
Constante	13.212	8.127	16.2%	3.961	6.799
Y^-1	- 613.731	- 3.353			
Constante	19.547	6.494	19.7%	3.919	6.757
Y^-(1/2)	- 132.629	- 3.771			
Constante	- 5.866	- 1.650	23.6%	3.869	6.707
Y^(1/2)	1.099	4.233			
Constante	3.031	1.825	22.1%	3.888	6.726
Y^(1.5)	0.002	4.054			

El listado de instrucciones

```
/* Prueba de Glejser
 Se ajusta un modelo utilizando los valores
de  gamma = -1 , -1/2,  1/2, 1,5    */
? Regresion inicial
Regress; Lhs=C; Rhs=one, Y ; Res=U; Keep=YE$ GLEJSER
? Valor absoluto de la desviación
Create;  AU=Abs(U)$
? Variables de las regresiones auxiliares
Create;  Z1=Y^(-1); Z2=Y^-(1/2); Z3=Y^(1/2); Z4=Y^(1.5)$
Regress; Lhs=AU; Rhs=one, Z1 $
Regress; Lhs=AU; Rhs=one, Z2 $
Regress; Lhs=AU; Rhs=one, Z3 $
Regress; Lhs=AU; Rhs=one, Z4 $
```

7.2.2.2 Park

La prueba de Park formaliza la intuición gráfica sugiriendo una forma funcional al patrón de los residuos que se vio en la sección 7.1.2, para ello se sugiere una forma funcional de la heteroscedasticidad como:

$$\sigma_i^2 = \sigma^2 Z_i^{\theta} e^{\upsilon_i} \qquad (7.8)$$

Donde v_i es el término de perturbación iid., e, es él número neperiano y Z_i es la variable que se sospecha que causa la heteroscedasticidad.

Dado que no se conoce σ^2_i, se utilizan los errores estimados de la regresión del modelo estimado por mínimos cuadrados $(\hat{u}_i)$ y se estima en la siguiente regresión.

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \theta \ln Z_i + \upsilon_i \qquad (7.9)$$

La hipótesis nula de homocedasticidad se rechaza sí el parámetro de la pendiente de la regresión (θ) es significativamente distinto de cero.

LISTADO PARA EJERCICIO

```
Reset$
/*   C: Consumo de los hogares
     Y: Ingreso de los hogares */
Read ; Nobs = 20 ; Nvar =3 ; Names =obs, C, Y$
1    19.9 22.3
2    31.2 32.3
3    31.8 36.6
4    12.1 12.1
5    40.7 42.3
6    6.1  6.2
7    38.6 44.7
8    25.5 26.1
9    10.3 10.3
10   38.8 40.2
11   8    8.1
12   33.1 34.5
13   33.5 38
14   13.1 14.1
15   14.8 16.4
16   21.6 24.1
17   29.3 30.1
18   25   28.3
19   17.9 18.2
20   19.8 20.1
```

BIBLIOGRAFÍA

- Aitken, A., (1935). «*On Least Squares and Linear Combinations of Observations*». **Proceedings of the Royal Statistical Society**, 55
- Breusch, T. S. y A. R. Pagan., (1979). «*A Simple Test for Heteroscedaticity and Random Coefficient Variation*». **Econometrica**, vol. 47. pp. 1287-1294
- Dougherty, Christopher. (1992). **Introduction to Econometrics**. Oxford University Press, Oxford y New York
- Glesjer, H., (1969). «A New Test for Heteroscedasticity», **Journal of the American Statistical Association**, vol. 64, pp. 316-323
- Godfrey, L. (1978) «*Testing For Multiplicative Heteroscedasticity*» **Journal of Econometrics**, vol. 8.
- Goldfeld S. M. y R. E. Quandt., (1965). «Some Tests for Homoscedasticity», **Journal of the American Statistical Association**, vol. 60, pp. 539-547
- Goldfeld S. M. y R. E. Quandt., (1972). **Nonlinear Methods in Econometrics**. North-Holland, Amsterdam,
- Greene, W.H., (1999). **Análisis Econométrico**. Prentice Hall IBERIA. Madrid. Tercera Edición
- Gujarati, Damodar.,(1997). **Econometría**. McGraw Hill INTERAMERICANA. Santafé de Bogotá, Colombia
- Hair, J. F, R. E. Anderson, R. L. Tatham y W. C. Black, (1999) **Análisis Multivariante**. Prentice Hall Madrid
- Harvey, A. C. y G. D. Phillips. (1973). «*A Comparison of the Power of Some Tests for Heteroscedasticity in General Linear Model*», **Journal of Econometrics**, vol. 2, 132
- Jarque, C. and A. Bera (1980) "Efficient Tests for Normality, Homoskedasticity, and Serial Independence of Regression Residuals," **Economics Letters**, 6, 255–259.
- Johnston, J. (1989). **Métodos de Econometría**. Editorial Vincens-Vives, Barcelona.
- Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee (1985) **The Theory and Practice of Econometrics**, 2nd edition, John Wiley & Sons
- Kamenta, Jan. (1977). **Elementos de Econometría**. Editorial Vicens-Vives Barcelona Kendall, M. G. y A. Stuart. (1961). **The Advanced Theory of Statistics**. Griffin, Londres, vol. 2.
- Levene, H. (1960) «Robust Tests for the Equality of Variances» en I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann (eds) **Contribution to Probability and Statistics**, Stanford University Press.
- Park, R. E. (1966) «*Estimation with Heteroscedastic Error Term*», **Econometrica**, vol. 34, No 4, pp. 461-465
- Ramírez; D. & R Helo y C. Rivera. (2000) «Evaluación de bienes ambientales que no se transan en el mercado: Caso delfines Nariz de Botellas Islas de Choros, IV Región». Documento Mimeografiado **Universidad Católica del Norte**, Coquimbo, Chile.
- White, Halbert (1980). «A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity». **Econometrica**, No. 48, pp. 817–838.