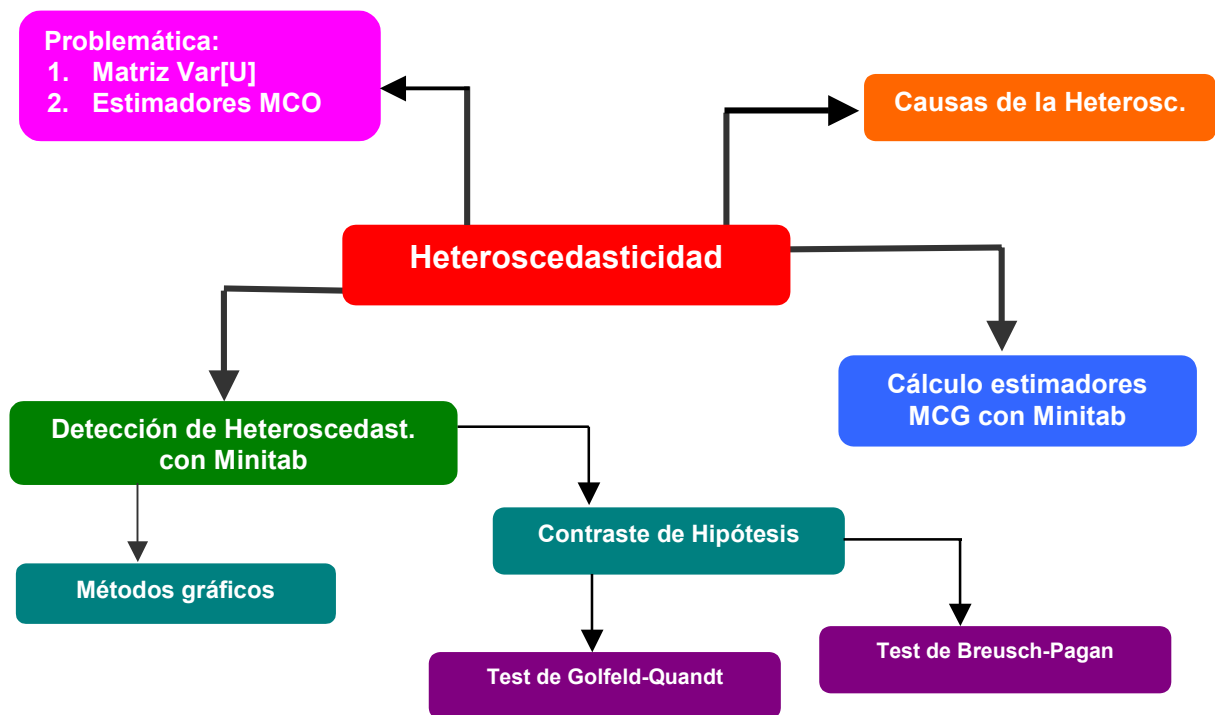


HETEROSCEDASTICIDAD

Autores: Ángel Alejandro Juan Pérez (ajuanp@uoc.edu), Renatas Kizys (rkizys@uoc.edu), Luis María Manzanedo Del Hoyo (lmanzanedo@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

En el *math-block* **Introducción al MRLG** vimos qué ocurría cuando fallaban las hipótesis de esfericidad en el término de perturbación. Nos centraremos ahora en la hipótesis de homoscedasticidad (el supuesto de varianza constante para el término de perturbación), y supondremos que el resto de las hipótesis de esfericidad sí se cumplen.

En este *math-block* aprenderemos a detectar la presencia de heteroscedasticidad (término de perturbación con varianza no constante) en el modelo, analizaremos algunas de sus posibles causas, y mostraremos cómo es posible resolver dicha problemática a fin de obtener estimadores de calidad.

OBJETIVOS

- Entender en qué consiste el problema de heteroscedasticidad y cómo afecta éste a la matriz de varianzas y covarianzas y a los estimadores MCO.
- Saber obtener los estimadores MCG (en el caso de heteroscedasticidad) con ayuda de Minitab.
- Entender las causas que pueden provocar el incumplimiento de la hipótesis de homoscedasticidad.
- Saber detectar, con ayuda de Minitab, la presencia de heteroscedasticidad en un modelo, tanto por medios gráficos como a través de contrastes de hipótesis.

CONOCIMIENTOS PREVIOS

Aparte de estar iniciado en el uso del paquete estadístico Minitab, resulta muy conveniente haber leído con profundidad los siguientes *math-blocks*:

- Regresión Lineal Múltiple
- Introducción al MRLG

CONCEPTOS FUNDAMENTALES

□ El problema de la Heteroscedasticidad

Como se comentó en el *math-block* **Introducción al MRLG**, cuando se utiliza el modelo de regresión lineal múltiple (donde usamos la notación $X_1 = 1$ para la “variable” que acompaña al término independiente):

$$Y = \beta_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + u$$

resulta habitual suponer que el término de perturbación presenta varianza constante (hipótesis de **homoscedasticidad**), i.e.:

$$Var[u_i] = Var[u_j] = \sigma^2 \quad \forall i \neq j$$

Por lo que a la matriz de varianzas y covarianzas se refiere, esta hipótesis se traduce en el hecho de que todos los términos de la diagonal principal serán iguales entre sí:

$$Var[U] = \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \dots & \\ & & & \sigma^2 \end{bmatrix}$$

Cuando no se cumpla la hipótesis de varianza constante para el término de perturbación, diremos que el modelo presenta problemas de **heteroscedasticidad**. En tal caso, no todos los términos de la diagonal principal de la matriz de varianzas y covarianzas serán iguales.

En presencia de heteroscedasticidad, y suponiendo que sí se cumple la hipótesis de No Autocorrelación (i.e.: que los términos fuera de la diagonal principal son ceros), la matriz de varianzas y covarianzas será de la forma:

$$Var[U] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \sigma^2 \cdot \begin{bmatrix} \gamma_1^2 & 0 & \dots & 0 \\ 0 & \gamma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \gamma_n^2 \end{bmatrix} = \sigma^2 \cdot \Omega_n$$

donde $Var[u_i] = \sigma_i^2$, y σ es un factor de escala (y, por tanto, la matriz Ω no es única).

En tales condiciones, el estimador MCO de **B** es insesgado y consistente, pero no es eficiente (es decir, ya no será el de mínima varianza, por lo que si usamos el estimador MCO en lugar del eficiente para hallar intervalos de confianza estaremos perdiendo precisión ya que obtendremos intervalos más grandes de los que proporcionaría el estimador eficiente). Además, el estimador de la varianza del término de perturbación, $\hat{\sigma}_u^2$, será sesgado.

□ Aplicación del MCG para modelos con Heteroscedasticidad

Como acabamos de ver, cuando el modelo presente problemas de heteroscedasticidad, el método MCO no nos proporciona buenos estimadores, y será necesario recurrir al método de **mínimos cuadrados ponderados** o **generalizado** (MCG) para obtener estimadores de calidad:

$$\hat{B}_{MCG} = (X' \cdot \Omega^{-1} \cdot X)^{-1} \cdot (X' \cdot \Omega^{-1} \cdot Y)$$

Por tanto, resulta imprescindible conocer la matriz Ω^{-1} (la inversa, salvo factor escalar, de la matriz de varianzas y covarianzas) o, alternativamente, la matriz $T = P^{-1}$ tal que $\Omega = P \cdot P'$ (recordemos que otra opción alternativa para hallar los estimadores MCG es aplicar MCO sobre el modelo transformado mediante T).

La forma concreta de la matriz Ω^{-1} (determinada en el caso de heteroscedasticidad por los elementos que componen su diagonal principal) dependerá de la expresión funcional que tome la varianza del término de perturbación. Más concretamente, si la forma funcional de la varianza es:

$$Var[u_i] = \sigma_i^2 = \sigma^2 \cdot f(i)$$

(donde f es una función que sólo depende de la variable i), entonces las matrices T y Ω^{-1} serán de la forma:

$$T = \begin{bmatrix} \frac{1}{\sqrt{f(1)}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{f(2)}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{f(n)}} \end{bmatrix} \quad \Omega^{-1} = \begin{bmatrix} \frac{1}{f(1)} & 0 & \dots & 0 \\ 0 & \frac{1}{f(2)} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{f(n)} \end{bmatrix}$$

En el *math-block* **introducción al MRLG** ya vimos cómo era posible usar Excel para hallar los estimadores MCG mediante operaciones con matrices. En el caso de heteroscedasticidad, sin embargo, Minitab nos facilita bastante la obtención de los estimadores MCG. Para ello, tan sólo será necesario introducir en el programa una columna de pesos compuesta por los elementos que configuran la diagonal principal de la matriz Ω^{-1} . Veámoslo con un ejemplo:

Ejemplo: Supongamos que pretendemos explicar el consumo familiar (C) en función del nivel de renta (R) mediante el modelo siguiente:

$$C = \beta_1 + \beta_2 \cdot R + u$$

Para ello disponemos de las siguientes observaciones:

↓	C1	C2
	C	R
1	15	10
2	23	12
3	13	14
4	23	16
5	27	18
6	15	20
7	21	22
8	55	24
9	47	26
10	55	28
11	53	30
12	33	32
13	13	34
14	50	36

Al realizar la regresión por MCO, obtenemos la siguiente ecuación:

Regression Analysis

The regression equation is
 $C = 6,2 + 1,10 R$

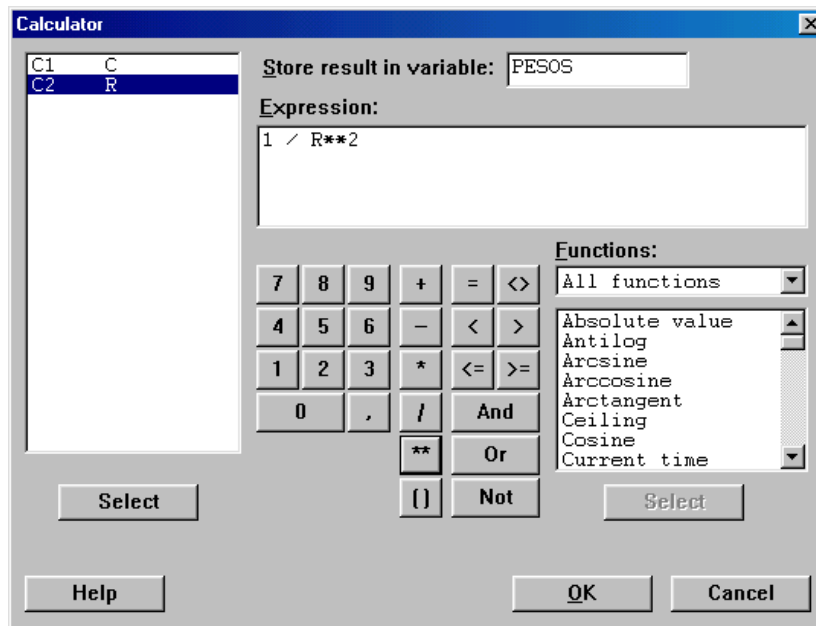
Sin embargo, tras haber realizado una serie de análisis, sospechamos que el modelo presenta problemas de heteroscedasticidad. En concreto, pensamos que la varianza del término de perturbación es directamente proporcional al cuadrado de la renta, i.e.:

$$Var[u_i] = \sigma^2 \cdot R_i^2$$

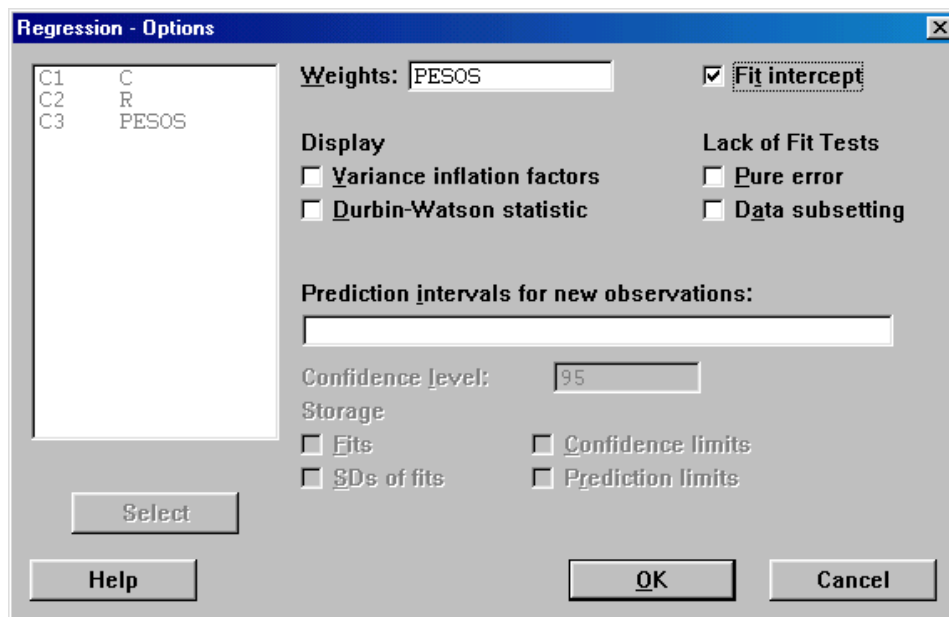
Así pues, la matriz Ω contendrá los términos R_i^2 en su diagonal principal (fuera de la diagonal principal sólo habrá ceros, ya que estamos suponiendo que no hay problemas de autocorrelación).

Crearemos una nueva columna, la de pesos, la cual contendrá los valores $1 / R_i^2$:

Calc > Calculator:



Finalmente, realizaremos la regresión por MCG. Para ello, deberemos indicar -en el menú de opciones de regresión- la columna que contiene los pesos a usar:



Obtendremos el siguiente “output”, en el cual se muestran los estimadores MCG para los coeficientes del modelo (observar que difieren bastante de los estimadores MCO que habíamos obtenido anteriormente):

Regression Analysis

Weighted analysis using weights in PESOS

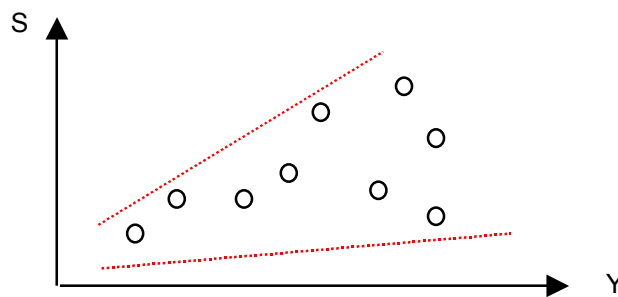
The regression equation is
 $C = 2,72 + 1,26 R$

□ **Causas de la Heteroscedasticidad**

- **Causas estructurales:** se producen al trabajar con modelos de corte transversal cuando las observaciones muestran un comportamiento muy heterogéneo. Consideremos el siguiente ejemplo, en el que se usan observaciones de corte transversal para explicar el ahorro mensual de las familias (S) en función de sus ingresos mensuales (Y):

$$S = \beta_1 + \beta_2 \cdot Y + u$$

Es de esperar que la varianza en S (y, por tanto, en el término de perturbación) dependa del nivel de ingresos Y: en aquellos casos de familias con bajos ingresos, los niveles de ahorro serán bastante similares (puesto que una gran proporción de los ingresos estarán destinados a cubrir las necesidades básicas). Sin embargo, en el caso de familias con ingresos elevados, encontraremos mucha más variedad (desde familias que dedican gran parte de sus ingresos al ahorro, hasta familias que dedican al consumo inmediato la práctica totalidad de lo que ingresan):



- **Causas muestrales:** en muchas ocasiones, no será posible disponer de todos los datos que componen una muestra, sino que sólo se tendrá acceso a valores medios o a valores agregados correspondientes a submuestras de la muestra total. Así, p.e., si estuviéramos interesados en realizar un estudio sobre el uso que de Internet hacen las familias españolas, podría ocurrir que no dispusiésemos de la información para cada una de las familias que componen la muestra total y que nos tuviésemos que conformar con los valores medios o agregados por provincias.

Pues bien, el trabajo con submuestras de diferentes tamaños propiciará la aparición de heteroscedasticidad en el modelo. Más concretamente, se puede demostrar que:

1. En el caso de usar submuestras de valores medios, $Var[u_i^m] = \sigma^2 \cdot \frac{1}{n_i}$
2. En el caso de usar submuestras de valores agregados, $Var[u_i^a] = \sigma^2 \cdot n_i$

(donde n_i es el número de observaciones que componen la muestra i -ésima).

- **Causas espurias:** finalmente, también es posible que la presencia de heteroscedasticidad sea consecuencia de la existencia de otros problemas en el modelo. Tal es el caso de una especificación errónea del modelo (p.e. la omisión de alguna variable explicativa relevante, lo que causaría que el término de perturbación absorbiese el efecto de dicha variable), o la existencia de cambios estructurales en el modelo (p.e. un cambio estacional en una serie temporal).

□ Detección de Heteroscedasticidad por métodos gráficos con Minitab

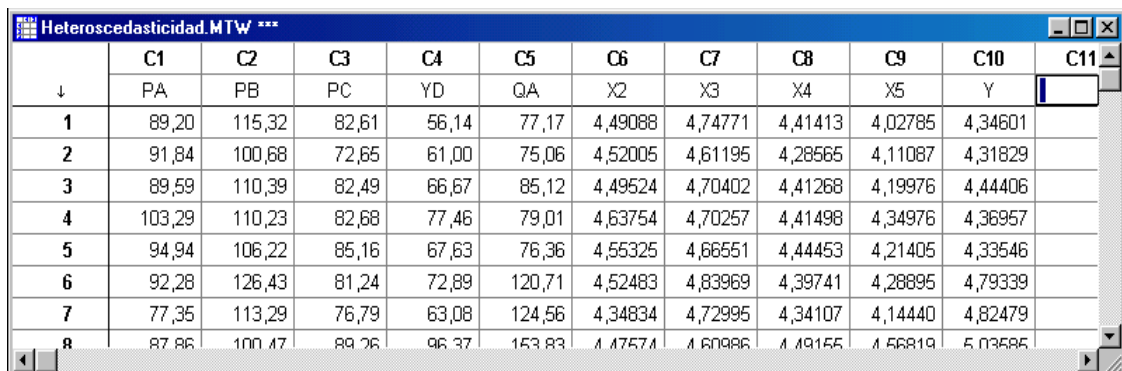
Para analizar la posible presencia de heteroscedasticidad en el modelo se suele recurrir a dos técnicas complementarias: (1) el análisis gráfico de los residuos (obtenidos al realizar la regresión por MCO), y (2) los contrastes de hipótesis específicos (test de Breusch-Pagan, test de Golfeld-Quandt, test de White, test de Glesjer, etc.).

Los métodos gráficos se basan en el hecho de que los residuos MCO son estimadores consistentes de los términos de perturbación. Por tanto, si se aprecian diferencias claras entre las varianzas de los residuos para diferentes niveles de una variable explicativa, podremos afirmar que los términos de perturbación no cumplen la hipótesis de homoscedasticidad.

Minitab nos permite representar gráficamente los residuos (o, alternativamente, los residuos estandarizados) frente a los valores estimados de la variable dependiente. También resulta conveniente representar los residuos (o los residuos estandarizados) frente a cada uno de las variables explicativas. Si se cumple el supuesto de homoscedasticidad, en todos los gráficos anteriores encontraremos variaciones similares en los residuos (eje vertical) para cualquier nivel del eje horizontal (donde se sitúa la variable explicativa o la estimación de la variable dependiente). En caso contrario, el modelo presentará problemas de heteroscedasticidad.

Ejemplo: El archivo **Heteroscedasticidad.mtw** contiene 30 observaciones de corte transversal para cada una de las siguientes variables:

VARIABLE	DESCRIPCIÓN	VARIABLE	DESCRIPCIÓN
PA	Precio (en euros) del producto A	X2	Ln(PA)
PB	Precio (en euros) del producto B	X3	Ln(PB)
PC	Precio (en euros) del producto C	X4	Ln(PC)
YD	Ingresos disponibles (en euros)	X5	Ln(YD)
QA	Cantidad demandada del producto A	Y	Ln(QA)



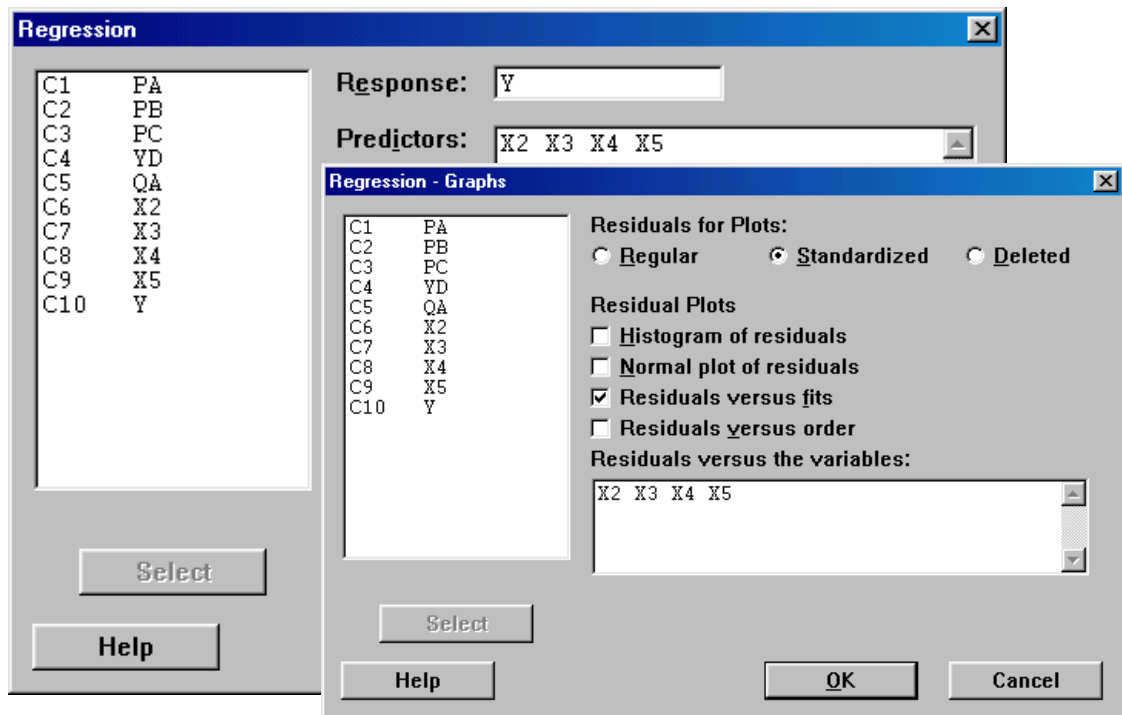
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
↓	PA	PB	PC	YD	QA	X2	X3	X4	X5	Y	
1	89,20	115,32	82,61	56,14	77,17	4,49088	4,74771	4,41413	4,02785	4,34601	
2	91,84	100,68	72,65	61,00	75,06	4,52005	4,61195	4,28565	4,11087	4,31829	
3	89,59	110,39	82,49	66,67	85,12	4,49524	4,70402	4,41268	4,19976	4,44406	
4	103,29	110,23	82,68	77,46	79,01	4,63754	4,70257	4,41498	4,34976	4,36957	
5	94,94	106,22	85,16	67,63	76,36	4,55325	4,66551	4,44453	4,21405	4,33546	
6	92,28	126,43	81,24	72,89	120,71	4,52483	4,83969	4,39741	4,28895	4,79339	
7	77,35	113,29	76,79	63,08	124,56	4,34834	4,72995	4,34107	4,14440	4,82479	
8	87,86	100,47	89,26	96,37	153,83	4,47574	4,60986	4,49155	4,56819	5,03585	

A fin de comprender cómo varía la demanda del producto A en función de los precios asociados a cada uno de los tres productos y de los ingresos disponibles, usaremos el siguiente modelo lineal:

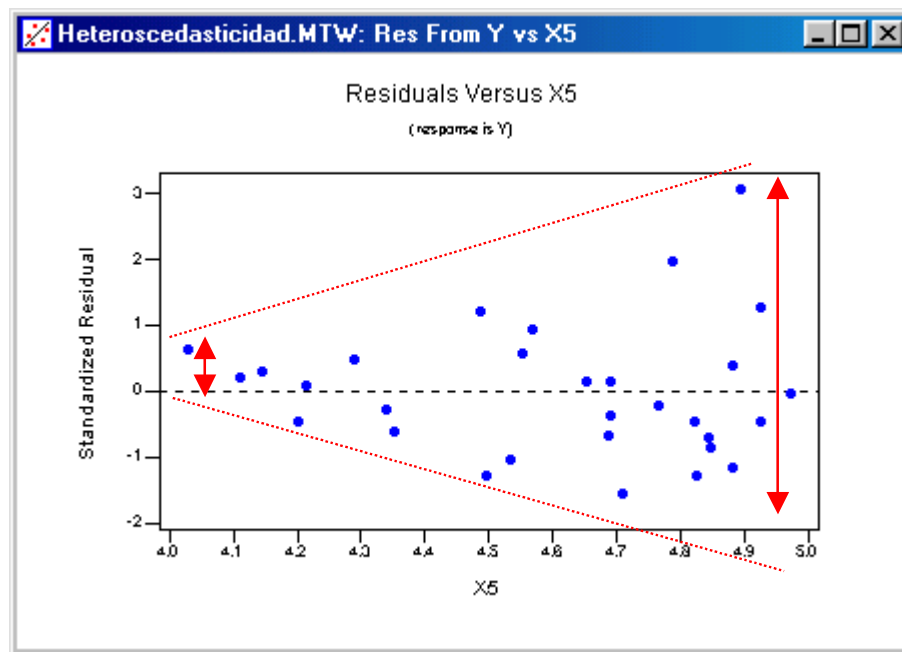
$$Y = \beta_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + u$$

Emplearemos Minitab para realizar la estimación por MCO y el análisis gráfico de los residuos estandarizados a fin de determinar si existe o no heteroscedasticidad en el modelo (le pediremos al programa que nos muestre un gráfico de los residuos estandarizados frente a los valores estimados de la variable dependiente, así como un gráfico de los residuos estandarizados frente a cada una de las variables explicativas):

Stat > Regression > Regression:



Al analizar algunos de los gráficos resultantes, podemos observar indicios claros de que la hipótesis de homoscedasticidad no se cumple. Así, p.e., en el gráfico siguiente se aprecia claramente cómo la varianza de los residuos va aumentando conforme mayor es el valor de la variable explicativa X5:



□ **Detección de Heteroscedasticidad mediante contraste de hipótesis**

A continuación presentaremos una serie de tests que se utilizan, de forma complementaria al análisis gráfico de los residuos, para realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \text{se cumple la hipótesis de homoscedasticidad} \\ H_1 : \text{NO se cumple la hipótesis de homoscedasticidad} \end{cases}$$

- **Test de Golfeld-Quandt:** este test se suele utilizar cuando se sospecha que la varianza del término de perturbación es directa o inversamente proporcional al valor de una de las variables explicativas, i.e.: cuando pensemos que el modelo:

$$Y = \beta_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + u$$

verifica una de las siguientes cuatro relaciones:

$$\begin{array}{ll} \text{Proporción directa:} & \text{Var}[u] = \delta \cdot X_j \quad \text{o} \quad \text{Var}[u] = \delta \cdot X_j^2 \\ \text{Proporción inversa:} & \text{Var}[u] = \delta \cdot \frac{1}{X_j} \quad \text{o} \quad \text{Var}[u] = \delta \cdot \frac{1}{X_j^2} \end{array}$$

donde $\delta > 0$ y X_j es alguna de las variables explicativas del modelo.

Los pasos a seguir para obtener el estadístico G-Q, el cual seguirá -bajo la hipótesis nula- una distribución F de Snedecor, son:

1. Ordenar según la variable X_j las observaciones correspondientes a todas las variables del modelo. Si la proporción es directa, ordenaremos de menor a mayor valor de la variable (orden creciente), mientras que si la proporción es indirecta lo haremos de mayor a menor valor (orden decreciente).
2. Dividir las observaciones ya ordenadas en tres grupos o submuestras, de forma que la submuestra central contenga las c observaciones centrales, donde c representa aproximadamente la cuarta parte del total de observaciones.
3. Aplicar regresión MCO a las submuestras primera y tercera (las cuales contendrán, respectivamente, los valores más pequeños y los más grandes de la variable X_j).
4. Obtener la suma de cuadrados del error (Error SS) para cada una de las regresiones anteriores (las denotaremos por Error SS1 y Error SS3 respectivamente).
5. Calcular el estadístico de Golfeld-Quandt:

$$GQ = \frac{\text{ErrorSS3}}{\text{ErrorSS1}} \approx (\text{bajo } H_0) \approx F\left(\frac{n-c}{2} - k, \frac{n-c}{2} - k\right)$$

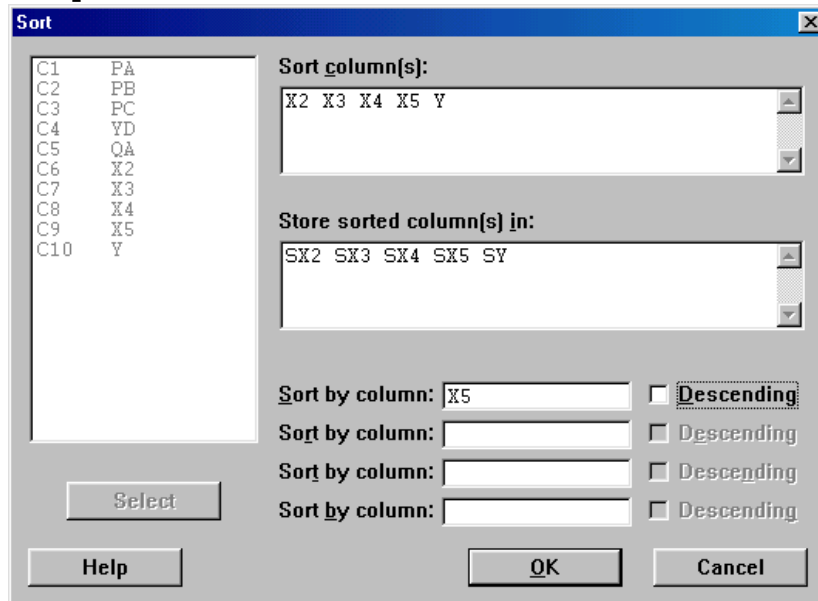
Una vez calculado el estadístico GQ, y para un nivel de significación α dado, usaremos la siguiente regla de decisión:

$$\text{Rechazamos } H_0 \text{ (i.e.: existe heterosced.)} \Leftrightarrow \text{si } GQ > F\left(\frac{n-c}{2} - k, \frac{n-c}{2} - k; \alpha\right)$$

Ejemplo: Continuando con el ejemplo anterior, el gráfico de los residuos no sólo parecía indicar la presencia de heteroscedasticidad en el modelo, sino que además apuntaba la idea de que la varianza del término de perturbación era directamente proporcional al valor de la variable X5 (a mayor valor de la variable X5, mayor varianza se observaba en el valor de los residuos). Aplicaremos ahora el test de Golfeld-Quandt para contrastar empíricamente la existencia de heteroscedasticidad:

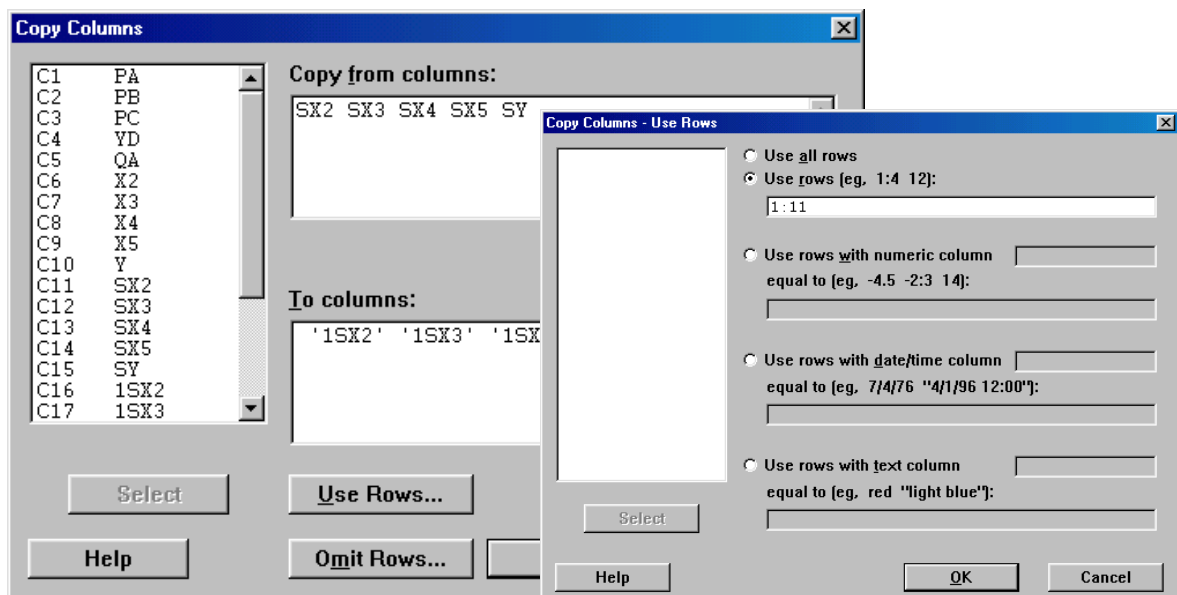
En primer lugar, ordenaremos –de menor a mayor según los valores de la variable X5- las observaciones de las variables X2, X3, X4, X5 e Y, guardando los datos resultantes en las nuevas variables SX2, SX3, SX4, SX5 e SY:

Manip > Sort:



A continuación, dividiremos las observaciones en 3 grupos: el primer grupo estará formado por las filas 1 a 11, y el tercero por las filas 20 a 30 (el grupo central abarcará las filas 12 a 19, por lo que descartamos un total de $c = 8$ filas de observaciones centrales). Las columnas del primer grupo las denotaremos por 1SX2, 1SX3, 1SX4, 1SX5 y 1SY, mientras que a las columnas del tercer grupo las denotaremos por 3SX2, 3SX3, 3SX4, 3SX5, y 3SY respectivamente:

Manip > Copy Columns...:



	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
↓	1SX2	1SX3	1SX4	1SX5	1SY	3SX2	3SX3	3SX4	3SX5	3SY	
1	4,49088	4,74771	4,41413	4,02785	4,34601	4,55671	4,48108	4,80254	4,78607	4,91523	
2	4,52005	4,61195	4,28565	4,11087	4,31829	4,68333	4,48458	4,76140	4,82101	4,49290	
3	4,34834	4,72995	4,34107	4,14440	4,82479	4,69465	4,22902	4,71850	4,82615	4,15324	
4	4,49524	4,70402	4,41268	4,19976	4,44406	4,62830	4,58996	4,70066	4,84458	4,79040	
5	4,55325	4,66551	4,44453	4,21405	4,33546	4,68417	4,53646	4,64947	4,84545	4,63986	
6	4,52483	4,83969	4,39741	4,28895	4,79339	4,65586	4,52548	4,80016	4,88182	4,75755	
7	4,64227	4,69857	4,45353	4,33860	4,34264	4,47961	4,56674	4,73321	4,88204	5,06203	
8	4,63754	4,70257	4,41498	4,34976	4,36957	4,68435	4,48706	4,73392	4,89425	5,02428	
9	4,63142	4,69839	4,44230	4,48503	4,75677	4,75935	4,41171	4,88015	4,92421	4,32028	
10	4,65072	4,63269	4,64083	4,49658	4,23077	4,83898	4,42185	4,90705	4,92500	4,33047	
11	4,52613	4,60906	4,57223	4,53400	4,59714	4,73602	4,50921	4,97176	4,97217	4,52958	
12											

El nuevo paso es realizar la regresión por MCO para cada uno de los grupos, a fin de obtener el correspondiente Error SS. En el caso del primer grupo, obtenemos el siguiente "output", con un valor de Error SS1 = 0,02891:

Regression Analysis					
The regression equation is					
1SY = 4,96 - 1,96 1SX2 + 1,59 1SX3 - 1,35 1SX4 + 1,62 1SX5					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	4	0,43741	0,10935	22,70	0,001
Residual Error	6	0,02891	0,00482		
Total	10	0,46632			

Al realizar la regresión por MCO sobre el tercero de los grupos, obtendremos un Error SS3 = 0,20030.

Así pues, el estadístico GQ = Error SS3 / Error SS1 = 6,93.

Por otra parte, para un nivel de significación $\alpha = 0,05$, el estadístico F con $(n-c)/2 = 11$ grados de libertad en numerador y denominador es:

Inverse Cumulative Distribution Function	
F distribution with 11 DF in numerator and 11 DF in denominator	
P(X <= x)	x
0,9500	2,8179

Dado que $GQ = 6,93 > F(11,11;0,05) = 2,82$ se sigue que hay indicios suficientes como para rechazar la hipótesis nula, i.e.: el test confirma la sospecha de que el modelo presenta heteroscedasticidad, y que la variable X5 (que era el logaritmo neperiano de la variable ingresos disponibles) tiene mucho que ver en ello.

- **Test de Breusch-Pagan:** el test de Goldfeld-Quandt se basaba en el supuesto de que la heteroscedasticidad venía provocada por una única variable. Cuando sean varias las variables causantes de tal problema, se deberá recurrir a otros tests. En concreto, el test de Breusch-Pagan -el cual sólo es estrictamente válido cuando se dispone de muestras suficientemente grandes- presupone que es posible expresar la varianza del término de perturbación como una combinación lineal de p variables explicativas, i.e.:

$$Var[u] = \alpha_0 + \alpha_1 \cdot Z_1 + \alpha_2 \cdot Z_2 + \dots + \alpha_p \cdot Z_p$$

Los pasos a seguir para obtener el estadístico BP, el cual seguirá -bajo la hipótesis nula- una distribución χ^2 con p grados de libertad, son:

1. Estimar por MCO el modelo original $Y = \beta_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + u$, guardando los residuos resultantes.
2. Definir $\hat{\sigma}^2 = \frac{ErrorSS}{n}$
3. Estimar por MCO el modelo auxiliar $\frac{e^2}{\hat{\sigma}^2} = \alpha_0 + \alpha_1 \cdot Z_1 + \alpha_2 \cdot Z_2 + \dots + \alpha_p \cdot Z_p + v$
4. Definir el estadístico $BP = \frac{RegressionSS}{2} \approx (\text{bajo } H_0) \approx \chi_p^2$

Una vez calculado el estadístico BP, y para un nivel de significación α dado, usaremos la siguiente regla de decisión:

$$\text{Rechazamos } H_0 \text{ (i.e.: existe heterosced.)} \Leftrightarrow \text{si } BP > \chi_{p,\alpha}^2$$

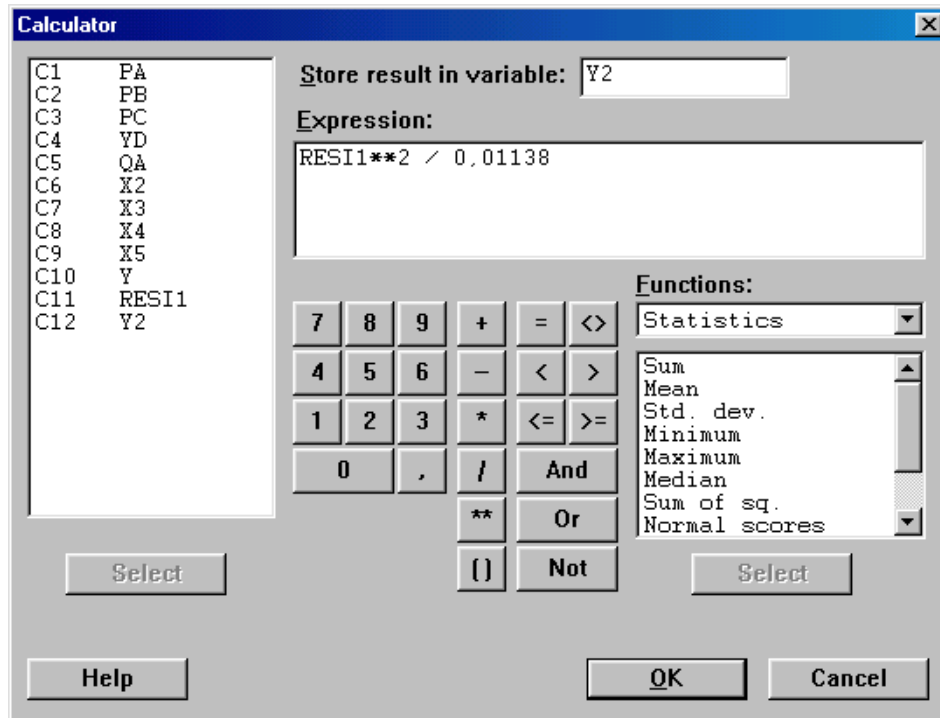
Ejemplo: Volviendo a nuestro ejemplo anterior, utilizaremos el test BP para determinar la posible existencia de heteroscedasticidad en el modelo causada por la variable X5. Para ello, lo primero será hacer la regresión por MCO del modelo inicial, guardando los residuos resultantes en la columna RES11:

Regression Analysis					
The regression equation is					
Y = 5,14 - 2,04 X2 + 1,17 X3 - 0,802 X4 + 1,57 X5					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	4	1,71925	0,42981	31,48	0,000
Residual Error	25	0,34139	0,01366		
Total	29	2,06063			

Observar que Error SS = 0,34139. Por tanto, $\hat{\sigma}^2 = \frac{ErrorSS}{n} = 0,34139 / 30 = 0,01138$.

Ahora, construiremos la nueva columna $\frac{e^2}{\hat{\sigma}^2}$:

Calc > Calculator:



Lo siguiente será realizar la regresión por MCO de esta nueva variable respecto a la variable X5:

Regression Analysis					
The regression equation is					
Y2 = - 8,20 + 2,00 X5					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	9,011	9,011	2,49	0,126
Residual Error	28	101,259	3,616		
Total	29	110,270			

El "output" nos proporciona un valor de Regression SS = 9,011. Ello significa que el estadístico de Breusch-Pagan será $BP = 9,011 / 2 = 4,51$.

Por otro lado, el valor crítico para un nivel de significación $\alpha = 0,05$ y teniendo en cuenta que sólo se ha usado una variable ($p = 1$ grado de libertad), será:

Inverse Cumulative Distribution Function	
Chi-Square with 1 DF	
P(X <= x)	x
0,9500	3,8415

Se puede concluir pues que sí hay indicios que nos llevan a rechazar la hipótesis nula de homoscedasticidad, ya que $BP = 4,51 > 3,84$.

CASOS PRÁCTICOS CON SOFTWARE

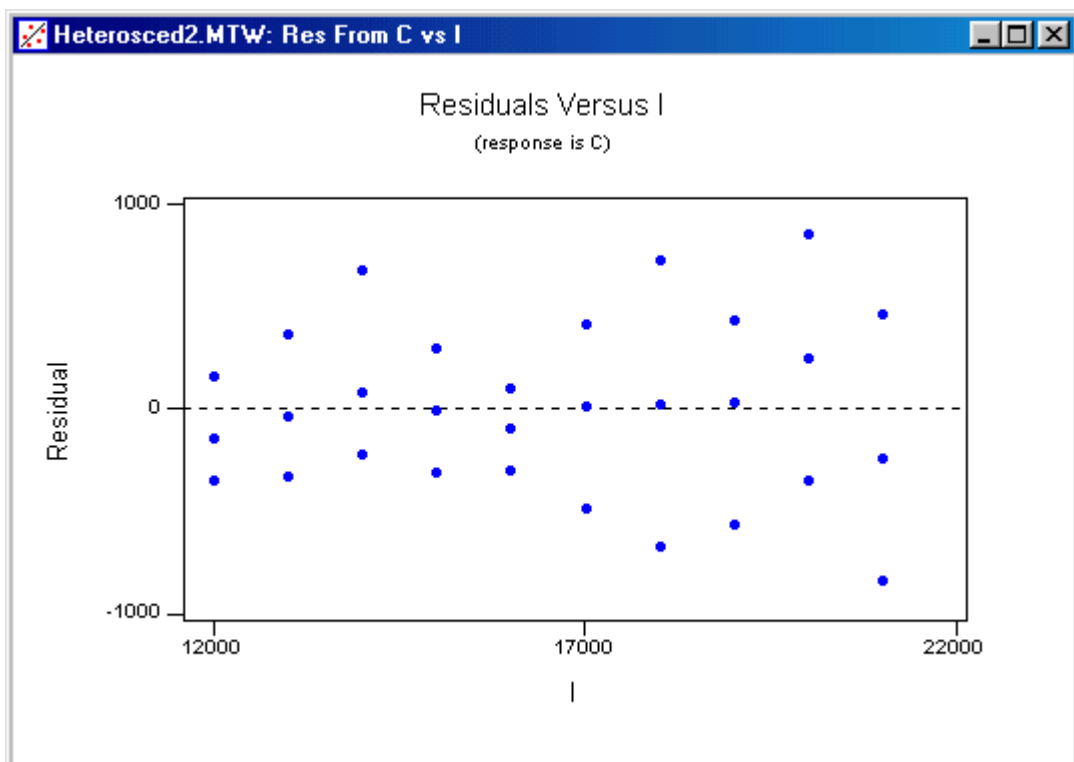
❑ **Detección de heteroscedasticidad en el modelo**

El archivo **Heterosced2.mtw** contiene el gasto en consumo (C) y los ingresos disponibles (I) para una muestra de 30 familias.

Al realizar la regresión por MCO obtenemos el “output” que se muestra a continuación:

Regression Analysis					
The regression equation is					
C = 1480 + 0,788 I					
Predictor	Coef	StDev	T	P	
Constant	1480,0	449,6	3,29	0,003	
I	0,78848	0,02685	29,37	0,000	
S = 422,3		R-Sq = 96,9%		R-Sq(adj) = 96,7%	
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	153872818	153872818	862,69	0,000
Residual Error	28	4994182	178364		
Total	29	158867000			

Sin embargo, el siguiente gráfico de los residuos frente a la variable explicativa parece indicar la existencia de heteroscedasticidad en el modelo:



Para salir de dudas, decidimos realizar -con ayuda de Minitab- el test de Golfeld-Quandt, eliminando un total de c = 6 observaciones centrales. Tras ordenar los datos y crear los tres

grupos de observaciones (de forma análoga a como se ha explicado anteriormente), obtenemos los siguientes “outputs” de regresión MCO para los grupos primero y tercero:

Regression Analysis

The regression equation is
1SC = 847 + 0,837 1SI

Predictor	Coef	StDev	T	P
Constant	847	1144	0,74	0,476
1SI	0,83667	0,08442	9,91	0,000

S = 327,0 R-Sq = 90,8% R-Sq(adj) = 89,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10500167	10500167	98,22	0,000
Residual Error	10	1069000	106900		
Total	11	11569167			

Regression Analysis

The regression equation is
3SC = 2307 + 0,747 3SI

Predictor	Coef	StDev	T	P
Constant	2307	2916	0,79	0,447
3SI	0,7467	0,1493	5,00	0,001

S = 578,3 R-Sq = 71,4% R-Sq(adj) = 68,6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8362667	8362667	25,01	0,001
Residual Error	10	3344000	334400		
Total	11	11706667			

Así pues, Error SS1 = 1.069.000 y Error SS3 = 3.344.000, de lo cual se sigue que el estadístico GQ = Error SS3 / Error SS1 = 3.13.

Dado que $n = 30$, $c = 6$, y $k = 2$, bajo la hipótesis H_0 (homoscedasticidad), el estadístico GQ se distribuirá según una F de Snedecor con $(30-6)/2 - 2 = 10$ grados de libertad en el numerador y los mismos grados de libertad en el denominador.

Por su parte, para un nivel de significación $\alpha = 0,05$, tenemos que el valor crítico será $F(10,10;0,05)$, el cual podemos hallar con ayuda de Minitab:

Inverse Cumulative Distribution Function

F distribution with 10 DF in numerator and 10 DF in denominator

P(X <= x)	x
0,9500	2,9782

Como $GQ = 3,13 > 2,98$ rechazaremos la hipótesis nula, i.e.: hay indicios claros de heteroscedasticidad en el modelo.

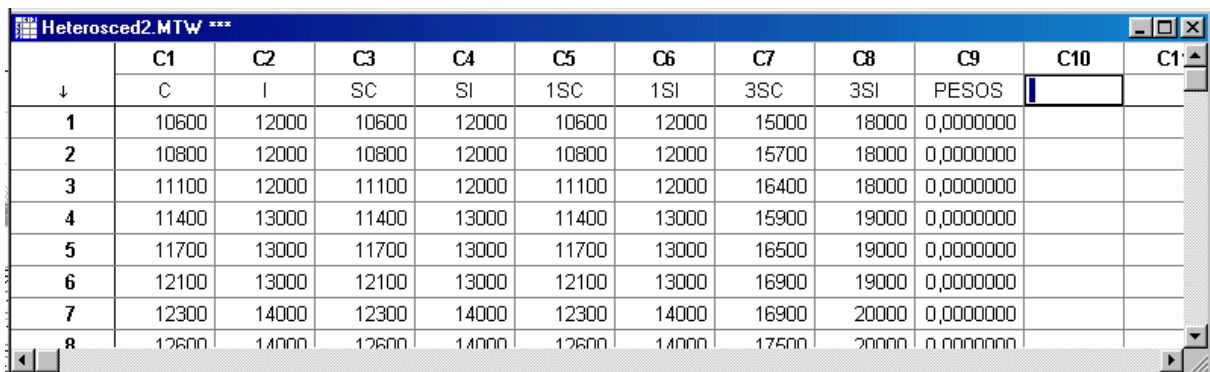
❑ **Corrección del problema de heteroscedasticidad: regresión por MCG**

Una vez confirmado que no se verifica la hipótesis de homoscedasticidad, aplicaremos el método de mínimos cuadrados ponderados (MCG) para obtener estimadores válidos.

Supondremos que la varianza del error es proporcional al cuadrado de los ingresos, i.e.:

$$Var[u_i] = \sigma^2 \cdot I_i^2$$

Así las cosas, la columna de pesos contendrá los valores $1 / I_i^2$:



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
↓	C	I	SC	SI	1SC	1SI	3SC	3SI	PESOS		
1	10600	12000	10600	12000	10600	12000	15000	18000	0,0000000		
2	10800	12000	10800	12000	10800	12000	15700	18000	0,0000000		
3	11100	12000	11100	12000	11100	12000	16400	18000	0,0000000		
4	11400	13000	11400	13000	11400	13000	15900	19000	0,0000000		
5	11700	13000	11700	13000	11700	13000	16500	19000	0,0000000		
6	12100	13000	12100	13000	12100	13000	16900	19000	0,0000000		
7	12300	14000	12300	14000	12300	14000	16900	20000	0,0000000		
8	12600	14000	12600	14000	12600	14000	17500	20000	0,0000000		

Al realizar la regresión por MCG (recordar incluir la columna de pesos en el menú de opciones), se obtiene el siguiente resultado:

Regression Analysis				
Weighted analysis using weights in PESOS				
The regression equation is				
C = 1421 + 0,792 I				
Predictor	Coef	StDev	T	P
Constant	1421,3	395,5	3,59	0,001
I	0,79210	0,02514	31,51	0,000
S = 0,02447		R-Sq = 97,3%		R-Sq(adj) = 97,2%

BIBLIOGRAFÍA

- [1] Artís, M.; Suriñach, J.; et al (2002): "Econometría". Ed. Fundació per a la Universitat Oberta de Catalunya. Barcelona.
- [2] Carter, R.; Griffiths, W.; Judge, G. (2000): "Using Excel for Undergraduate Econometrics". ISBN: 0-471-41237-6
- [3] Doran, H. (1989): "Applied Regression Analysis in Econometrics". Ed. Marcel Dekker, Inc. ISBN: 0-8247-8049-3
- [4] Gujarati, D. (1997): "Econometría básica". McGraw-Hill. ISBN 958-600-585-2
- [5] Johnston, J. (2001): "Métodos de econometría". Ed. Vicens Vives. Barcelona. ISBN 84-316-6116-X
- [6] Kennedy, P. (1998): "A Guide to Econometrics". Ed. MIT Press. ISBN: 0262611406
- [7] Novalés, A. (1993): "Econometría". McGraw-Hill. ISBN 84-481-0128-6
- [8] Pulido, A. (2001): "Modelos econométricos". Ed. Pirámide. Madrid. ISBN 84-368-1534-3
- [9] Uriel, E. (1990): "Econometría: el modelo lineal". Ed. AC. Madrid. ISBN 84-7288-150-4
- [10] Wooldridge, J. (2001): "Introducción a la Econometría: un enfoque moderno". Ed. Thomson Learning. ISBN: 970-686-054-1

ENLACES

- ❑ <http://www.uam.es/departamentos/economicas/econapli/pdf/heterocedasticidad.pdf>
Heterocedasticidad con E-Views
- ❑ <http://www.feweb.vu.nl/econometriclinks/index.html>
The Econometrics Journal On-Line
- ❑ <http://www.elsevier.com/hes/books/02/menu02.htm>
Libro on-line: Handbook of Econometrics Vols. 1-5
- ❑ <http://elsa.berkeley.edu/users/mcfadden/discrete.html>
Libro on-line: Structural Analysis of Discrete Data and Econometric Applications
- ❑ http://www.oswego.edu/~kane/econometrics/stud_resources.htm
Online Resources for Econometric Students
- ❑ <http://www.econ.uiuc.edu/~morillo/links.html>
Econometric Sources: a collection of links in econometrics and computing. University of Illinois
- ❑ <http://www.econometrics.net/>
Econometrics, Statistics, Mathematics, and Forecasting
- ❑ <http://ideas.uqam.ca/EDIRC/ectrix.html>
Economics Departments, Institutes and Research Centers in the World: Econometrics, Mathematical Economics.