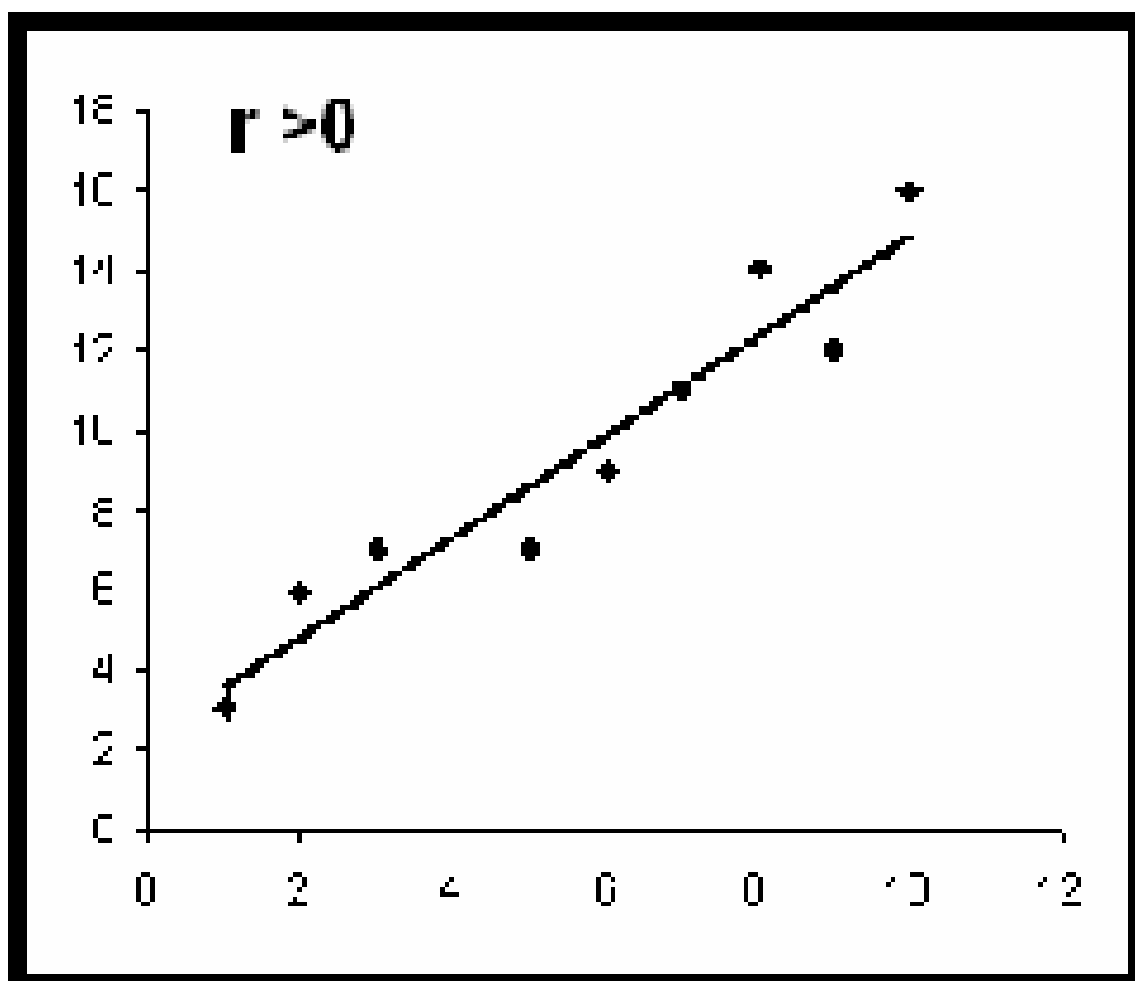


# MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE CON CORRECCIONES DE ESPECIFICACIÓN, USOS DE STATA 9.0, STATA 10.0, EVIEWS 5.0, SSPS 11.0



**Autores:**

**Rafael David Escalante Cortina  
Wilson Mayorga Mogollon  
Juan Carlos Vergara Schmalbach**

## Contenido

MODELO DE REGRESION LINEAL MULTIPLE.....	3
Etapas para realizar una regresión Lineal Múltiple.....	4
SUPESTOS DE MINIMOS CUADRADOS ORDINARIOS.....	5
PRUEBAS ESTADISTICAS PARA MEDIR LA SIGNIFICANCIA DEL MODELO Y DE LOS PARAMETROS. ....	7
Propiedades de los estimadores .....	7
Regresión con Variables Dummy .....	10
PROBLEMAS DE ESPECIFICACIÓN DEL MODELO DE REGRESIÓN MULTIPLE.....	33
MULTICOLINEALIDAD .....	33
HETEROCEDASTICIDAD .....	42
AUTOCORRELACION.....	45
APLICACIÓN DE MODELO LOGISTICOS EN SPSS 11.0.....	57
BIBLIOGRAFIA .....	67

## MODELO DE REGRESION LINEAL MULTIPLE

El modelo de regresión lineal múltiple busca una función de regresión poblacional a partir de una función de regresión muestral. La idea de la regresión es mostrar un análisis cuantitativo los fenómenos financieros y económicos combinados con inferencia de la variable explicada.

Este modelo une a la teoría económica, la estadística, y la matemática para establecer relaciones entre una variable dependiente llamada “y” y una o varias variables explicativas llamadas “x”, con el fin establecer un predicción o en su defecto el impacto que tienen las variables explicativas sobre la variable explicada.

$$\begin{array}{l}
 y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \beta_4 x_{41} \dots \dots \dots \beta_k x_{k1} + u_1 \\
 y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \beta_4 x_{42} \dots \dots \dots \beta_k x_{k2} + u_2 \\
 \vdots \\
 y_n = \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \beta_4 x_{4n} \dots \dots \dots \beta_k x_{kn} + u_n
 \end{array}$$

Para este modelo se definen dos ecuaciones:

$$Y = X\beta + U$$

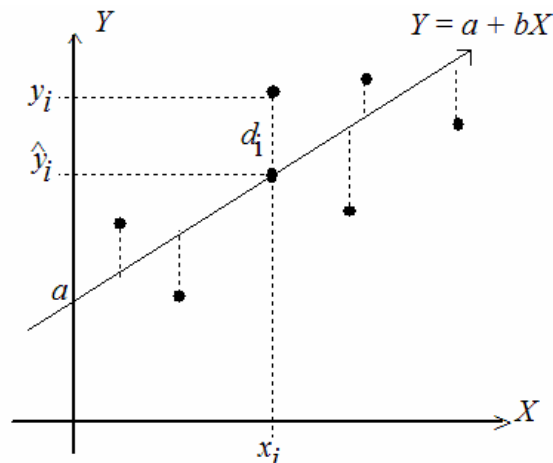
Esta función va ser llamada la regresión poblacional#

$$\hat{Y} = \hat{X} \hat{\beta} + \hat{U}$$

Esta función va ser llamada la regresión muestral

Existen dos razones para aplicar la regresión múltiple:

- **Explicativa:** A través de este modelo la variable explicada “y” se pueden medir los impactos que tienen sus variables explicativas, es decir que los betas que acompañan la regresión pueden ser positivos o negativos los cuales no dicen si una variable explicativa ayuda a crecer o disminuir la variable dependiente.
- **Predicción:** por medio de la regresión y reemplazando las variables explicativas “X” por valores numéricos, la variable dependiente “Y” puede tomar diferentes valores.



Bajo el enfoque matricial:

$$\hat{Y} = \hat{X} \hat{\beta} + \hat{U}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}, u = \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix}_{n \times 1}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{k \times 1}, x = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix}_{n \times k}$$

$$y(na) = x(na) * \beta(na) + u(na)$$

Demostración:

- **Criterio del Minimos Cuadrados Ordinarios:**

$$\begin{aligned} MinSCE = u'u = y'y - \hat{y}'\hat{y} &= Min \sum u_i^2 = \sum y_i^2 - \sum \hat{y}_i^2 = 0 \\ U'U = (Y - X\hat{\beta})'(Y - X\hat{\beta}) &= Y'Y - Y'X\hat{\beta} + \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} = Y'Y - Y'X\hat{\beta} = \\ &= Y'Y - \hat{\beta}'(X'X)\hat{\beta} = Y'Y - \hat{Y}'\hat{Y} \end{aligned}$$

## Etapas para realizar una regresión Lineal Múltiple

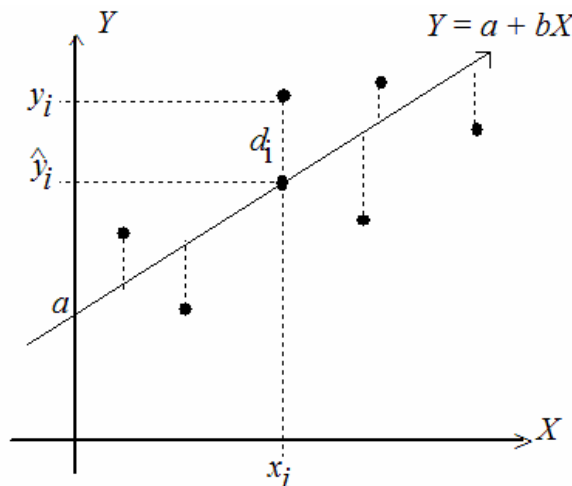
1. Plantear una hipótesis: Es importante encontrar un problema que se quiere estudiar el cual debe ser basado en la teoría financiera o económica.
2. Especificar modelo matemático de la teoría económica: Se debe buscar en teoría y a través de experimentos anteriores cual ha sido los resultados, en caso de que la hipótesis nunca ha sido medida se debe buscar un modelo similar para hacer la comparación respectiva.
3. Especificar modelo econométrico de la teoría económica: se busca el modelo teórico y sus variables iniciales.
4. Obtención de datos: se obtienen los datos que a través de las encuestas o en su defecto se busca la base datos. En este punto se hace una aclaración referente a la estructura de datos que podemos encontrar. Los datos pueden ser
  - Sección Cruzada: cuando los datos se toman en un momento en el tiempo, como una fotografía en el tiempo.
  - Series de Tiempo: Cuando los datos se toman en diferentes momentos del tiempo de una misma unidad.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

- Combinaciones datos sección Cruzada: Se refiere cuando los datos tienen unidades diferentes y en diferentes momentos en el tiempo.
5. Estimación del modelo: Se calcula el modelo en el paquete estadísticos, entre los más usados se encuentran, Excel, Stata, SPSS windows, Stata, SAS, R, etc.
  6. Prueba de hipótesis: se verifican si los “betas” de la regresión, son adecuados.
  7. Pronóstico o predicción: Con la ecuación ajustada a los betas consistentes y confiables se reemplazan los valores a las variables explicativas y se obtiene el pronóstico de la variable explicada.
  8. Uso de modelos para recomendación: se hacen los ajustes dados por los investigadores.

### SUPESTOS DE MINIMOS CUADRADOS ORDINARIOS.

1. El modelo es lineal en los parámetros (los betas).
2. Las variables explicativas toma valores fijos en muestreo repetitivo.
3. La esperanza del error es nula es de decir,  $\sum e_i = 0$



4. No hay autocorrelaciones en los errores, es decir que la esperanza de la covarianzas entre los errores es nula.  $E(u_i u_j) = 0$  para todo  $i \neq j$
5. Homocedasticidad:, que significa que la varianza de los errores es constante.

$$\text{Var} (u) = \sigma^2 I_{n \times n}$$

6. El modelo esta correctamente especificado.

7. No hay relación lineal perfecta entre las variables explicativas.
8. El número de observaciones es mayor que los parámetros estimados.
9. Los errores “U” siguen distribución normal. (0,  $\sigma^2$ ). Es decir media cero y varianza constante.
10. Las Variables explicativas “X” son exógenas.

**Variabilidad de los errores en la regresión lineal:**

1.- Variabilidad total:  $SCT = \sum (y_i - \bar{y})^2$ <sup>1</sup>; Sumatoria de los cuadrados totales.

2.- Variabilidad explicada:  $SCE = \sum \left( \hat{y}_i - \bar{y} \right)^2$  Sumatoria de los cuadrados de errores

3.- Variabilidad no explicada:  $SCR = \sum u_i^2$  Sumatoria de los cuadrados de la regresión

En el caso en el que el modelo hay una constante la  $SCT = SCE + SCR$ .

La bondad de ajuste muestra qué tan bien se ajusta la línea de regresión a los datos. El coeficiente de determinación  $r^2$  para la regresión lineal simple y el  $R^2$  para la regresión múltiple, es una medida que nos dice qué tan bien se ajusta la línea de regresión muestral a los datos.

$$R^2 = 1 - \frac{\sum \hat{U}_i^2}{\sum (Y_i - \bar{Y}^2)} = 1 - \frac{SCE}{SCT} = 1 - \frac{\text{Suma de Residuales Cuadrados}}{\text{Suma Total de Cuadrados}}$$

Una propiedad importante del  $R^2$  es que es una función no decreciente del número de variables explicativas o de regresores presentes en el modelo; a medida que aumenta el número de regresores, el  $R^2$  aumenta. Es por esto que al comparar dos modelos de regresión con la misma variable dependiente pero con un número diferente de variables X, se debe tener mucho cuidado al escoger el modelo con el  $R^2$  más alto. Para comparar dos términos  $R^2$ , se debe tener en cuenta el número de variables presentes en el modelo. Esto puede hacerse con el coeficiente de determinación ajustado  $\bar{R}^2$ , el cual está ajustado por los grados de libertad asociados con la suma de los cuadrados.

$$\bar{R}^2 = 1 - \frac{\sum \hat{U}_i^2 / (n - k)}{\sum (Y_i - \bar{Y}^2) / (n - 1)}$$

<sup>1</sup> La “Y” barra se refiere al promedio ponderado de la variable explicada.

## PRUEBAS ESTADÍSTICAS PARA MEDIR LA SIGNIFICANCIA DEL MODELO Y DE LOS PARÁMETROS.

- La prueba “t” de student es utilizada para medir la significancia estadística de los parámetros del modelo, es decir los betas. El estadístico “t” (t-statistic) que se calcula como cociente entre el estimador y su error estándar  $\frac{\hat{\beta}}{\hat{\sigma}}$  permite contrastar la hipótesis de que el coeficiente es igual a cero ( $H_0 : \beta = 0$  frente a  $H_a : \beta \neq 0$ ) y por lo tanto la variable en cuestión no es individualmente significativa para explicar el comportamiento de la variable endógeno. Para que la variable
- La prueba estadística “F” de Fisher puede medir la significancia global del modelo es decir que el modelo de regresión múltiple es estadísticamente significativo.

Ambos caso se busca un nivel de confianza del 95%, con un p-valor entre cero y 0,05 para que el modelo y los parámetros sean estadísticamente significativos.

### Propiedades de los estimadores

#### 1. Lineales: es una función lineal

es la matriz de proyección.

#### 2. Insesgados: La esperanza del estimador coincide con el beta poblacional.

$$E(\beta) = E[(X'X)^{-1}X'Y] = E[(X'X)^{-1}X'(X\beta + u)] = \beta + (X'X)^{-1}X'E(u) = \beta / E(u) = 0$$

3. Además de estimadores lineales e insesgados, también son los de mínima varianza (de Gauss-Markov) que son los mínimo cuadrados. El nuevo estimador sería  $Var(\hat{\beta}) = \sigma^2[(X'X)^{-1}]$ .

A continuación se presentan la forma de la lectura de los estimadores con respecto a los resultados obtenidos en la variable independiente.

Modelo	Ecuación	Pendiente = $\frac{dY}{dX}$	Elasticidad = $\left(\frac{dY}{dX} \frac{X}{Y}\right)$	Características
Lineal	$Y = \beta_1 + \beta_2 X + u$	$\beta_2$	$\beta_2 \left(\frac{X}{Y}\right)$	Pendiente constante. Elasticidad variable (depende del valor de X y Y). Interpretación $\beta_2$ : un cambio en una unidad de X genera un cambio en $\beta_2$ unidades de Y. Es decir cambio absoluto genera cambio absoluto.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Log - log	$LnY = \beta_1 + \beta_2 LnX + u$	$\beta_2 \left( \frac{Y}{X} \right)$	$\beta_2$	Pendiente variable (depende del valor de X y Y). Elasticidad constante. Interpretación $\beta_2$ : un cambio en un 1% de X genera un cambio en $\beta_2$ por ciento de Y (Ojo: no se multiplica por 100). Este es un cambio porcentual genera cambio porcentual.
Log - lin	$LnY = \beta_1 + \beta_2 X + u$	$\beta_2(Y)$	$\beta_2(X)$	Pendiente variable (depende del valor de X y Y). Elasticidad variable (depende del valor de X y Y). Interpretación $\beta_2$ : un cambio en una unidad de X genera un cambio en $100 * \beta_2$ por ciento de Y. Este es un cambio absoluto genera cambio porcentual. También se interpreta como una tasa de crecimiento.

### EJEMPLO (paquete estadístico *EViews*): Modelo regresión lineal Múltiple

Se desea estimar el efecto de la tasa de desempleo  $X_1(\%)$ , y la tasa de inflación esperada  $X_2(\%)$ , sobre la tasa de inflación observada  $Y(\%)$ .

Dependent Variable: Y

Method: Least Squares

Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.193357	1.594789	4.510538	0.0011
$X_1$	-1.392472	0.305018	-4.565214	0.0010
$X_2$	1.470032	0.175786	8.362633	0.0000
R-squared	0.876590	Mean dependent var	7.756923	
Adjusted R-squared	0.851907	S.D. dependent var	3.041892	
S.E. of regression	1.170605	Akaike info criterion	3.352092	
Sum squared resid	13.70316	Schwarz criterion	3.482465	
Log likelihood	-18.78860	F-statistic	35.51521	
Durbin-Watson stat	2.225465	Prob(F-statistic)	0.000029	



En este modelo, se puede observar que la tasa de inflación esperada  $X_2$  (%), los signos de los coeficientes de las variables explicativas son los esperados (Curva de Phillips ampliada). El intercepto muestra que si  $X_2$  y  $X_3$  hubiese sido cero durante el periodo muestral, la tasa promedio de inflación observada habría estado cercana al 7.19%. El coeficiente de regresión parcial de -1.392472 significa que al mantener constante la tasa de inflación esperada, la tasa de inflación observada en promedio se redujo en cerca del 1.4% por cada aumento del 1% de la tasa de desempleo. De igual manera, al mantener la tasa de desempleo constante, el valor del coeficiente de 1.470032 implica que la tasa de inflación observada en promedio, aumentó en cerca de 1.5% por cada aumento del 1% en la tasa de inflación esperada.

El  $R^2$  es alto, e indica que las dos variables explicativas, en su conjunto, son la causa de cerca del 88% de la variación en la tasa de inflación observada.

## Regresión con Variables Dummy

Al realizar análisis de regresión, la variable dependiente<sup>2</sup> y las independientes no solamente pueden estar dadas por variables cuantitativas, existen otros tipos de variables de carácter cualitativo. Dichas variables se conocen comúnmente como Variables: Dummies, categóricas, dicótomas, binarias, ficticias o cualitativas.

Usualmente, dichas variables indican la presencia o ausencia de una cualidad o atributo, como por ejemplo sexo, raza, color, religión, etc. Estas son variables que toman valor de 1 en una submuestra y 0 en el resto de la muestra. Si el número de submuestras es mayor a dos, se define una variable ficticia para cada una de ellas, tomando valor de 1 en dicha submuestra y el valor de 0 en el resto de las observaciones muestrales. Las variables ficticias pueden ser nominales y ordinales.

Al incluir variables dummies en un modelo de regresión, estimar sus coeficientes y llevar a cabo los contrastes de significancia de las variables, es equivalente a estimar los modelos restringido y no restringido (Novales, 1993), explicados en la clase anterior.

A continuación, se relacionan ejemplos de modelos estimados con variables cualitativas, con el fin de analizar significancia individual de los betas, y mencionar la interpretación de los coeficientes estimados. Asimismo, se pretende analizar los cambios que se presentan en el intercepto y pendiente de la regresión estimada.

### Ejercicios:

#### 1. Modelo de regresión simple con una sola variable dummy

La expresión general de este modelo, esta dada por:

$$Y_i = \beta_1 + \beta_2 D_i + u_i$$

Donde:

$Y$ , es la variable dependiente.

$D_i$ , es la variable explicativa dummy.

Estimamos en stata el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 female + u$$

Donde:

---

<sup>2</sup> Los modelos de regresión con variable dependiente dummy, presentan comúnmente 4 enfoques de estimación: El modelo Lineal de Probabilidad (MLP), El modelo Logit, El modelo Probit, y el Modelo Tobit. Sin embargo, para esta clase trabajaremos con modelos de regresión con variables independientes de esta clase.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

*WAGE* , es el salario/hora en dólares.

*female* , es la variable dummy que representa el sexo de las personas, y toma los valores de:

*female* = 1 cuando la persona es mujer.

*female* = 0 cuando la persona es hombre.

$\beta_2$ , es el parámetro que define la diferencia entre el salario/hora de las mujeres y hombres. Si el coeficiente  $\beta_2 < 0$ , las mujeres ganan (dado que la categoría base es hombre), en promedio, menos que los hombres.

**NOTA:** Si una variable dummy tiene  $m$  categorías, se debe introducir en el modelo a estimar,  $m - 1$  variables cualitativas. Lo anterior, con el fin de evitar la trampa de la variable dicótoma, es decir, la situación de multicolinealidad perfecta. En el ejemplo a realizar, la variable *female* tiene dos categorías (hombre ó mujer) y, por lo tanto, se introducirá solamente una variable dummy. Teniendo en cuenta lo anterior, siempre se deberá tomar una categoría como base, con el fin de comparar las estimaciones realizadas con respecto a esa categoría.

**Salida en stata:** `reg wage female`

Source	SS	df	MS	Number of obs =	526
-----+-----				F( 1, 524) =	68.54
Model	828.220467	1	828.220467	Prob > F	= 0.0000
Residual	6332.19382	524	12.0843394	R-squared	= 0.1157
-----+-----				Adj R-squared =	0.1140
Total	7160.41429	525	13.6388844	Root MSE	= 3.4763
-----					
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
female	-2.51183	.3034092	-8.28	0.000	-3.107878 -1.915782
_cons	7.099489	.2100082	33.81	0.000	6.686928 7.51205
-----					

De la salida anterior se puede decir varias cosas.

**Primero:**

$$E[WAGE / female = 0] = \beta_1 = 7,1$$

La intercepción es el salario promedio de los hombres de la muestra (cuando  $female = 0$ ), de modo que ellos, en promedio, ganan 7.1 dólares por hora.

**Segundo:**

$$E[WAGE / female = 1] = \beta_1 + \beta_2 = 4,59$$

El salario promedio de las mujeres es 4.59 (7.1-2.51) dólares por hora.

**Y tercero:**

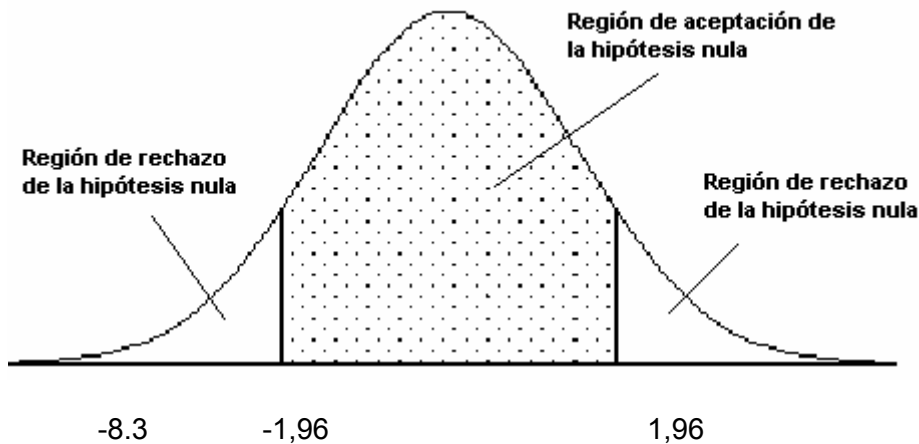
$$E[WAGE / female = 1] - E[WAGE / female = 0] = \beta_1 + \beta_2 - \beta_1 = \beta_2 = -2,51$$

El coeficiente de  $female$ , es la diferencia en el salario promedio entre hombres y mujeres, es decir la mujeres, en promedio, ganan 2.51 dólares menos que los hombres. A este coeficiente se le puede llamar **coeficiente de intercepción** porque dice que tanto difiere el valor del termino de intercepción de la categoría que recibe el valor de 1, del coeficiente del intercepción de la categoría base.

Sin embargo, para saber si el  $\beta_2$  es estadísticamente significativo, y así poder concluir con certeza que los hombres ganan, en promedio, más que las mujeres, realizamos la prueba de significancia individual de la variable  $female$ :

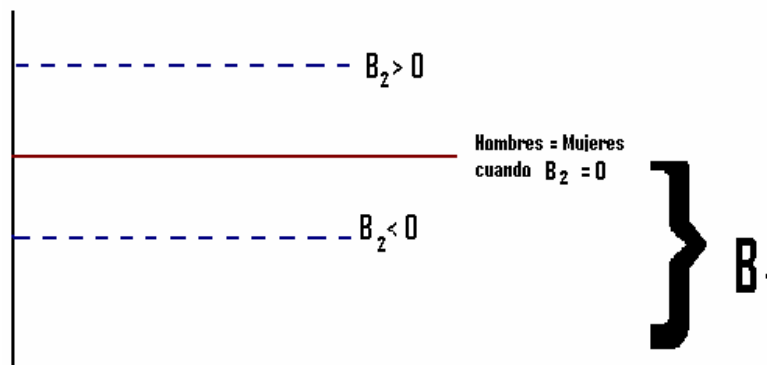
$$\begin{aligned} H_o : \beta_2 &= 0 \\ H_a : \beta_2 &\neq 0 \end{aligned} \quad t_c = \frac{\hat{\beta}_i - \delta}{\sqrt{Var(\hat{\beta}_i)}} = t_c = \frac{-2.51}{0,30} = -8.3$$

$$t_{tabla0,025;526} = \pm 1,96$$



Ahora, dado que el  $t$  tabla con un nivel de significancia de 5% ( $\alpha = 0.05, \alpha/2 = 0.025$ ) y 526 grados de libertad es  $\pm 1.96$ , se puede apreciar en el gráfico, que el  $t_c$  (-8.3) cae en la región de rechazo, por lo tanto hay evidencia suficiente para rechazar la hipótesis nula, es decir, el  $\beta_2$  es estadísticamente diferente a cero y la variable es significativa, luego se puede decir que las mujeres, en promedio, ganan 2.5 dólares por hora menos que los hombres. De igual forma, en la salida en stata se puede evidenciar que el p-valor de la variable *female* es 0.0000, lo cual indica que la variable es estadísticamente significativa al 1%.

**NOTA:** En general, la regresión simple con una constante y una variable dummy es una forma directa de comparar las medidas de dos grupos; para este ejemplo, entre el salario de las mujeres y los hombres. Geométricamente, se podría dar que:



## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Para el caso de este modelo, se observa un cambio en intercepto. Como el beta que acompaña la variable *female* es negativo ( $\beta_2 < 0$ ), se tiene un menor nivel de salario por hora de las mujeres con respecto a los hombres.

Al correr el modelo anterior, tomando como la categoría base a las mujeres, es decir, dando el valor de 1 a los hombres y 0 a las mujeres se obtiene:

**Salida en stata:** `reg wage hombre`

Source	SS	df	MS	Number of obs	=	526
-----+-----				F( 1, 524)	=	68.54
Model	828.220467	1	828.220467	Prob > F	=	0.0000
Residual	6332.19382	524	12.0843394	R-squared	=	0.1157
-----+-----				Adj R-squared	=	0.1140
Total	7160.41429	525	13.6388844	Root MSE	=	3.4763
-----						
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
hombre	2.51183	.3034092	8.28	0.000	1.915782	3.107878
_cons	4.587659	.2189834	20.95	0.000	4.157466	5.017852
-----						

Se puede observar, que el coeficiente estimado para la variable *hombre*, presenta signo contrario al modelo anterior estimado con la variable *female*. De igual forma, se evidencia que el intercepto ha cambiado, ahora el intercepto representa el salario/hora de las mujeres (cuando *hombre* = 0).

Retomando el concepto de la trampa de variable dicótoma, existe una forma de evitarla diferente a incluir  $m - 1$  categorías dentro del modelo. Para esto, consideremos el siguiente modelo de regresión a través del origen:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 hombre + u$$

Donde:

*WAGE*, es el salario/hora en dólares.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

*female* , es una variable dummy que toma los valores de:

*female* = 1 cuando la persona es mujer.

*female* = 0 cuando la persona es hombre.

*hom bre* , es una variable dummy que toma los valores de:

*hom bre* = 1 cuando la persona es hombre.

*hom bre* = 0 cuando la persona es mujer.

**Salida en stata:** `reg wage female hombre exper, noconstant`

Source	SS	df	MS	Number of obs = 526		
-----+-----				F( 3, 523) = 534.06		
Model	19184.0401	3	6394.68002	Prob > F = 0.0000		
Residual	6262.25231	523	11.9737138	R-squared = 0.7539		
-----+-----				Adj R-squared = 0.7525		
Total	25446.2924	526	48.3769817	Root MSE = 3.4603		
-----						
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
female	4.145462	.2845875	14.57	0.000	3.586387	4.704537
hombre	6.626882	.2862475	23.15	0.000	6.064546	7.189218
exper	.0269163	.0111369	2.42	0.016	.0050379	.0487948
-----						

Como se puede observar en la salida, ahora el  $\beta_2$  y  $\beta_3$  no representan el efecto diferencial entre el salario de las mujeres y hombres. Los valores de los coeficientes estimados, son el salario promedio/hora de las mujeres y hombres respectivamente.

Sin embargo, es importante tener en cuenta que la practica común es asignar las variables dicótomas de tal manera que si una variable tiene  $m$  categorías, se introducen solamente  $(m - 1)$  variables dicotómicas (Gujarati, 2001). Lo anterior, teniendo en cuenta que al utilizar dicho enfoque se obtiene un valor usual del  $R^2$ , mientras que con el modelo

sin intercepto<sup>3</sup>, se tendría que calcular lo que se conoce como el  $R^2$  simple. Asimismo, cuando se considera a priori un modelo sin intercepto, se puede estar cometiendo un error de especificación, violando así uno de los supuestos del modelo clásico de regresión lineal.

### 2. Modelo de regresión múltiple con una sola variable dummy y una variable cuantitativa.

La expresión general de este modelo, esta dada por:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + u_i$$

Donde:

$Y$ , es la variable dependiente.

$D_i$ , es la variable explicativa dummy.

$X_i$ , es la variable explicativa cuantitativa.

Para realizar un ejemplo, trabajaremos con la base de datos WAGE1.RAW.

Estimamos en stata el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 exper + u$$

Donde:

$WAGE$ , es el salario/hora en dólares.

$female$ , es la variable dummy que representa el sexo de las personas, y toma los valores de:

$female = 1$  cuando la persona es mujer.

$female = 0$  cuando la persona es hombre.

$exper$ , es la variable que representa los años de experiencia de la persona.

$\beta_2$ , es el parámetro que define la diferencia entre el salario/hora de las mujeres y hombres, dado los mismos niveles de experiencia. Si, manteniendo constante la experiencia, el coeficiente  $\beta_2 < 0$ , las mujeres ganan, en promedio, menos que los hombres, dado un mismo nivel de experiencia.

---

<sup>3</sup> Cuando se estima un modelo de regresión con intercepto cero, se habla de un modelo de regresión a través del origen.



## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Salida en stata: *reg wage female exper*

Source	SS	df	MS	Number of obs =	526
-----+-----				F( 2, 523) =	37.51
Model	898.161983	2	449.080991	Prob > F	= 0.0000
Residual	6262.25231	523	11.9737138	R-squared	= 0.1254
-----+-----				Adj R-squared =	0.1221
Total	7160.41429	525	13.6388844	Root MSE	= 3.4603
-----					
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
female	-2.48142	.3022793	-8.21	0.000	-3.07525 -1.887589
exper	.0269163	.0111369	2.42	0.016	.0050379 .0487948
_cons	6.626882	.2862475	23.15	0.000	6.064546 7.189218
-----					

De la salida anterior se puede decir varias cosas.

### Primero:

$$E[WAGE / exper, female = 0] = \beta_1 + \beta_3 \text{ exper} = 6,62 + 0,026 \text{ exper}$$

Por cada año de experiencia que tengan los hombres (cuando *female* = 0 ), su salario promedio, aumenta en 6.64 (6.62+0.02) dólares hora

### Segundo:

$$E[WAGE / exper, female = 1] = (\beta_1 + \beta_2) + \beta_3 \text{ exper} = 6,64 + 0,026 \text{ exper}$$

Por cada año de experiencia que tengan las mujeres (cuando *female* = 1 ), su salario promedio, aumenta en 4.16 (6.62+0.02-2.48) dólares hora. Es decir, que las mujeres ganan menos que los hombres, para un mismo nivel de experiencia.

### Y tercero:

$$E[WAGE / exper, female = 1] - E[WAGE / exper, female = 0] = \beta_1 + \beta_2 + \beta_3 \text{ exper} - \beta_1 - \beta_3 \text{ exper} = \beta_2$$

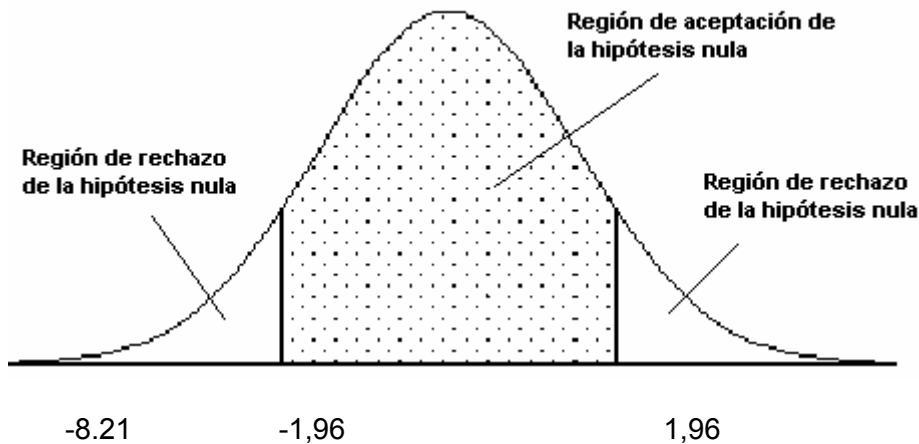
## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

El coeficiente de *female*, es la diferencia en el salario promedio entre hombres y mujeres manteniendo un mismo nivel de experiencia, es decir las mujeres, en promedio, ganan 2.51 dólares menos que los hombres.

Sin embargo, para saber si el  $\beta_2$  y  $\beta_3$  son estadísticamente significativos, y así poder concluir con certeza que un año de experiencia de trabajo en los hombres aporta más a su salario, que un año de experiencia de las mujeres, tenemos que realizar la prueba de significancia individual de las variables *female* y *exper*.

$$\begin{aligned} H_o : \beta_2 &= 0 \\ H_a : \beta_2 &\neq 0 \end{aligned} \quad t_c = \frac{\hat{\beta}_i - \delta}{\sqrt{\text{Var}(\hat{\beta}_i)}} = t_c = \frac{-2.48}{0.3022} = -8.21$$

$$t_{\text{tabla } 0,025;526} = \pm 1,96$$

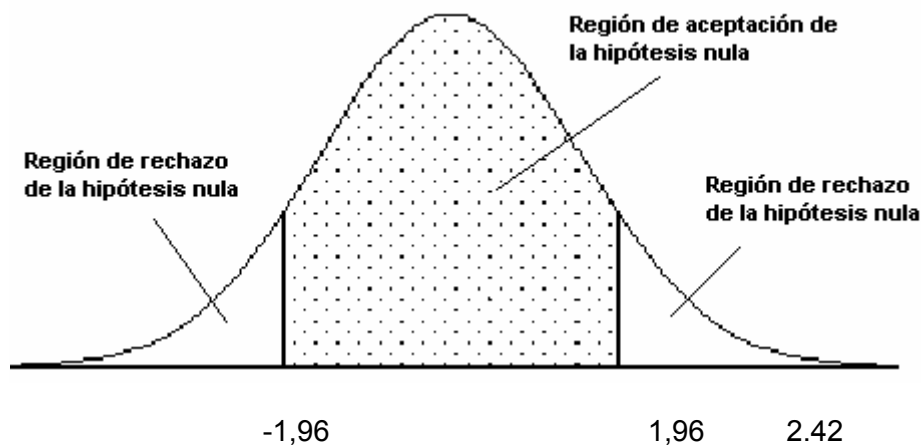


Ahora, dado que el t tabla con un nivel de significancia de 5% ( $\alpha = 0.05, \alpha/2 = 0.025$ ) y 526 grados de libertad es  $\pm 1.96$ , se puede apreciar en el gráfico, que el  $t_c$  (-8.21) cae en la región de rechazo, por lo tanto hay evidencia suficiente para rechazar la hipótesis nula, es decir, el  $\beta_2$  es estadísticamente diferente a cero y la variable es significativa. De igual forma, en la salida en stata se puede evidenciar que el p-valor de la variable *female* es 0.0000, lo cual indica que la variable es estadísticamente significativa al 1%.

Ahora:

$$\begin{aligned} H_o : \beta_3 &= 0 \\ H_a : \beta_3 &\neq 0 \end{aligned} \quad t_c = \frac{\hat{\beta}_i - \delta}{\sqrt{\text{Var}(\hat{\beta}_i)}} = t_c = \frac{0.026}{0.011} = 2.42$$

$$t_{\text{tabla}0,025;526} = \pm 1,96$$



Ahora, dado que el  $t$  tabla con un nivel de significancia de 5% ( $\alpha = 0.05, \alpha/2 = 0.025$ ) y 526 grados de libertad es  $\pm 1.96$ , se puede apreciar en el gráfico, que el  $t_c$  (2.42) cae en la región de rechazo, por lo tanto hay evidencia suficiente para rechazar la hipótesis nula, es decir, el  $\beta_3$  es estadísticamente diferente a cero y la variable es significativa. De igual forma, en la salida en stata se puede evidenciar que el p-valor de la variable *exper* es 0.016, lo cual indica que la variable es estadísticamente significativa al 1%.

Al igual que en el ejemplo anterior, se puede estimar un modelo de regresión sin intercepto con el fin de enviar la trampa de la variable dicotómica. Considérese el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 \text{female} + \beta_3 \text{hombre} + \beta_3 \text{exper} + u$$

**Salida en Stata:** `reg wage female hombre exper, noconstan`

Source	SS	df	MS	Number of obs =	526
-----+-----					
Model	19184.0401	3	6394.68002	F( 3, 523) =	534.06
Residual	6262.25231	523	11.9737138	Prob > F =	0.0000
-----+-----					
				R-squared =	0.7539
				Adj R-squared =	0.7525
Total	25446.2924	526	48.3769817	Root MSE =	3.4603
-----					

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
female	4.145462	.2845875	14.57	0.000	3.586387	4.704537
hombre	6.626882	.2862475	23.15	0.000	6.064546	7.189218
exper	.0269163	.0111369	2.42	0.016	.0050379	.0487948
-----						

Se puede observar que los coeficientes estimados representan el salario promedio de las mujeres y los hombres, dado un nivel de experiencia.

### 3. Modelo de regresión múltiple con una sola variable dummy y dos o más variables cuantitativas

Al modelo del ejemplo anterior, le agregaremos una variable explicativa cuantitativa.

Nuestro modelo a estimar ahora será:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 exper + \beta_4 tenure + \beta_5 exper^2 + \beta_6 tenure^2 + u$$

Donde:

$WAGE$ , es el salario/hora en dólares.

$female$ , es la variable dummy que representa el sexo de las personas, y toma los valores de:

$female = 1$  cuando la persona es mujer.

$female = 0$  cuando la persona es hombre.

$exper$ , es la variable que representa los años de experiencia de la persona.

$tenure$ , es la variable que representa la antigüedad de la persona en el trabajo.

$exper^2$ , es la variable que representa los años de experiencia de la persona al cuadrado.

$tenure^2$ , es la variable que representa la antigüedad de la persona en el trabajo al cuadrado.

**Salida en stata:** `reg wage female exper tenure expersq tenursq`

Source	SS	df	MS	Number of obs =	526
--------	----	----	----	-----------------	-----

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

-----+-----					F( 5, 520) = 37.84
Model		1910.28805	5	382.057611	Prob > F = 0.0000
Residual		5250.12624	520	10.0963966	R-squared = 0.2668
-----+-----					Adj R-squared = 0.2597
Total		7160.41429	525	13.6388844	Root MSE = 3.1775
-----					
wage		Coef.	Std. Err.	t	P> t  [95% Conf. Interval]
-----+-----					
female		-1.998933	.2837282	-7.05	0.000 -2.556328 -1.441539
exper		.2003255	.039522	5.07	0.000 .1226832 .2779678
tenure		.2521445	.0542485	4.65	0.000 .1455714 .3587176
expersq		-.0049574	.0008445	-5.87	0.000 -.0066166 -.0032983
tenursq		-.0037726	.0018635	-2.02	0.043 -.0074335 -.0001117
_cons		4.79956	.347897	13.80	0.000 4.116104 5.483017

De acuerdo con la salida en Stata, todos los betas estimados que acompañan a las variables explicativas, son estadísticamente significativos. Se observa, que los  $t$  calculados son caen en la región de rechazo comparado con un  $t$  tabla con 526 g.l. y  $\alpha/2 = 0.025$  igual a  $\pm 1.96$ , es decir, se rechaza la hipótesis nula, por tanto se puede decir que los betas son estadísticamente diferentes de cero.

En cuanto a la interpretación de los coeficientes estimados, se puede decir que las mujeres ganan en promedio menos que los hombres manteniendo constantes las demás variables; un año de experiencia y/o antigüedad, proporciona mayor salario/hora a los hombres y las mujeres, manteniendo constantes el resto de variables.

La interpretación de las variables estimadas al cuadrado, permite establecer que la experiencia y la antigüedad presentan rendimientos marginales decrecientes, es decir, que a medida que aumenta la experiencia y antigüedad de las personas, aumenta el salario/hora, hasta un punto en el cual después disminuye.

#### 4. Modelo de regresión múltiple con dos variables dummy y una variable cuantitativa.

Continuando con la base de datos WAGE.RAW, estimaremos el siguiente modelo:

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

$$WAGE = \beta_1 + \beta_2 female + \beta_3 tenure + \beta_4 married + u$$

Donde:

*WAGE* , es el salario/hora en dólares.

*female* , es la variable dummy que representa el sexo de las personas, y toma los valores de:

*female* = 1 cuando la persona es mujer.

*female* = 0 cuando la persona es hombre.

*married* , es la variable dummy que representa el estado civil de las personas, y toma los valores de:

*married* = 1 cuando la persona es casada.

*married* = 0 cuando la persona no casada.

*tenure* , es la variable que representa la antigüedad de la persona en el trabajo.

**Salida stata:** `reg wage female tenure married`

Source	SS	df	MS	Number of obs =	526
-----+-----				F( 3, 522) =	46.41
Model	1507.68441	3	502.561468	Prob > F	= 0.0000
Residual	5652.72989	522	10.8289845	R-squared	= 0.2106
-----+-----				Adj R-squared =	0.2060
Total	7160.41429	525	13.6388844	Root MSE	= 3.2907
-----					
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
female	-1.976333	.2953181	-6.69	0.000	-2.556491 -1.396176
tenure	.1354407	.0207625	6.52	0.000	.0946523 .1762291
married	.9131324	.3051762	2.99	0.003	.313608 1.512657
_cons	5.596056	.2989656	18.72	0.000	5.008732 6.183379

De acuerdo con la salida en Stata, los  $\beta_2$ ,  $\beta_3$  y  $\beta_4$  que acompañan a las variables *female*, *tenure* y *married* respectivamente, son estadísticamente significativos, pues vemos que su p-valor asociado al estadístico t son 0.00. Además se evidencia que los t calculados son -6.69, 6.52 y 2.99 respectivamente, caen en la región de rechazo comparado con un t tabla con 526 g.l. y  $\alpha/2 = 0.025$  igual a  $\pm 1.96$ , es decir se rechaza la hipótesis nula, por tanto estadísticamente el  $\beta_2$  y  $\beta_3$  son diferentes de cero.

Como se puede observar en la salida anterior, el p valor asociado a la  $F$  es de 0.0000, lo cual indica que el modelo presenta una significancia global inclusive al 1%.

En cuanto a la interpretación de los coeficientes estimados, se tiene que las mujeres, ganan en promedio, menos que los hombres, independientemente si son casadas o no, manteniendo constante la antigüedad; y las personas casadas, gana en promedio, mas que los solteros, ya sean de sexo masculino ó femenino, dado un mismo nivel de antigüedad.

Teniendo en cuenta lo anterior, tenemos que:

El salario promedio de un hombre soltero dado un nivel de antigüedad en el trabajo, esta dado por:

$$E[WAGE / female = 0, married = 0, tenure] = \beta_1 + \beta_3 tenure = 5.59 + 0.135tenure$$

El salario promedio de una mujer soltera dado un nivel de antigüedad en el trabajo, esta dado por:

$$E[WAGE / female = 1, married = 0, tenure] = (\beta_1 + \beta_2) + \beta_3 tenure = 3.62 + 0.135tenure$$

El salario promedio de un hombre casado, dado un nivel de antigüedad en el trabajo, esta dado por:

$$E[WAGE / female = 0, married = 1, tenure] = (\beta_1 + \beta_4) + \beta_3 tenure = 6.5 + 0.135tenure$$

El salario promedio de una mujer casada, dado un nivel de antigüedad en el trabajo, esta dado por:

$$E[WAGE / female = 1, married = 1, tenure] = (\beta_1 + \beta_2 + \beta_4) + \beta_3 tenure = 4.53 + 0.135tenure$$

**5. Modelo de regresión múltiple con varias variables dummy (con una variable cualitativa de tres o cuatro categorías)**

Considérese el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 married + \beta_4 northcen + \beta_5 south + \beta_6 west + u$$

Donde:

*WAGE* , es el salario/hora en dólares.

*female* , es la variable dummy que representa el sexo de las personas, y toma los valores de:

*female* = 1 cuando la persona es mujer.

*female* = 0 cuando la persona es hombre.

*married* , es la variable dummy que representa el estado civil de las personas, y toma los valores de:

*married* = 1 cuando la persona es casada.

*married* = 0 cuando la persona no casada.

*northcen* , es la variable dummy que representa si la persona vive en el norte, y toma los valores de:

*northcen* = 1 cuando la persona vive en el norte.

*northcen* = 0 cuando la persona no vive en el norte.

*south* , es la variable dummy que representa si la persona vive en el sur, y toma los valores de:

*south* = 1 cuando la persona vive en el sur.

*south* = 0 cuando la persona no vive en el sur.

*west* , es la variable dummy que representa si la persona vive en el occidente, y toma los valores de:

*west* = 1 cuando la persona vive en el occidente.

*west* = 0 cuando la persona no vive en el occidente.

**NOTA:** La categoría base para el sector donde viven las personas son las personas que viven en el oriente.



## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Salida: *reg wage female married northcen south west*

Source	SS	df	MS	Number of obs =	526
-----+-----				F( 5, 520) =	21.55
Model	1228.99076	5	245.798152	Prob > F	= 0.0000
Residual	5931.42353	520	11.4065837	R-squared	= 0.1716
-----+-----				Adj R-squared =	0.1637
Total	7160.41429	525	13.6388844	Root MSE	= 3.3774
-----					
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
female	-2.337965	.2994471	-7.81	0.000	-2.92624 -1.74969
married	1.417395	.3068326	4.62	0.000	.8146113 2.020179
northcen	-.6532592	.4281555	-1.53	0.128	-1.494386 .1878678
south	-1.161885	.398185	-2.92	0.004	-1.944134 -.379636
west	.3794599	.4747887	0.80	0.425	-.5532799 1.3122
_cons	6.666695	.3917518	17.02	0.000	5.897084 7.436305
-----					

De la salida anterior se puede observar varias cosas:

- Dado que el intercepto representa la categoría base, se incluyen variables dummies solo para 3 de las 4 categorías del sector donde viven las personas, con el fin de no caer en la trampa de la variable dicótoma (multicolinealidad perfecta).
- Las variables *northcen* y *west* no son significativas al 10%; por lo tanto, se puede decir que no existe diferencia en el salario promedio, entre las personas que viven en estos sectores y los que viven en el oriente (la categoría base).
- Teniendo en cuenta que la variable *south* es estadísticamente significativa (p-valor 0.004), se puede decir que las personas que viven en el sur, ganan en promedio, 1.16 dólares menos que los que viven en el oriente, manteniendo constantes las demás variables.

## 6. Modelo de regresión múltiple con una interacción de dummies.

Considérese el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 educ + \beta_4 exper + \beta_5 tenure + u$$

Con:

*WAGE* , es el salario/hora en dólares.

*female* , es la variable dummy que representa el sexo de las personas, y toma los valores de:

*female* = 1 cuando la persona es mujer.

*female* = 0 cuando la persona es hombre.

*married* , es la variable dummy que representa el estado civil de las personas, y toma los valores de:

*married* = 1 cuando la persona es casada.

*married* = 0 cuando la persona no casada.

*exper* , es la variable que representa los años de experiencia de la persona.

*tenure* , es la variable que representa la antigüedad de la persona en el trabajo.

**Salida stata:** `reg wage female educ exper tenure`

Source	SS	df	MS	Number of obs =	526
-----+-----				F( 4, 521) =	74.40
Model	2603.10658	4	650.776644	Prob > F	= 0.0000
Residual	4557.30771	521	8.7472317	R-squared	= 0.3635
-----+-----				Adj R-squared =	0.3587
Total	7160.41429	525	13.6388844	Root MSE	= 2.9576
-----					
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
female	-1.810852	.2648252	-6.84	0.000	-2.331109 -1.290596

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

educ		.5715048	.0493373	11.58	0.000	.4745803	.6684293
exper		.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure		.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons		-1.567939	.7245511	-2.16	0.031	-2.991339	-.144538

-----

De acuerdo con la salida en Stata, todos los betas estimados que acompañan a las variables explicativas, son estadísticamente significativos. Se observa, que los  $t$  calculados son caen en la región de rechazo comparado con un  $t$  tabla con 526 g.l. y  $\alpha/2 = 0.025$  igual a  $\pm 1.96$ , es decir, se rechaza la hipótesis nula, por tanto se puede decir que los betas son estadísticamente diferentes de cero.

En cuanto a la interpretación de los coeficientes estimados, se puede decir que las mujeres ganan en promedio menos que los hombres manteniendo constantes las demás variables; un año de experiencia, educación y/o antigüedad en el trabajo, proporciona mayor salario/hora tanto a los hombres como las mujeres, manteniendo constantes el resto de variables.

De otro lado, en este modelo esta implícito el supuesto de que el efecto diferencial de la variable dummy *female* es constante a través del posible estado civil de las personas (casado y no casado). Por ejemplo, en la salida anterior se observa que las mujeres ganan en promedio menos que los hombres, esto se da ya sea casada o no.

En muchas aplicaciones, este supuesto puede ser imposible de mantener, una mujer puede ganar mas cuando es casada que cuando es soltera. Con el fin de observar este efecto dentro del modelo, se adiciona la variable *femalemarried*, que representa la interacción entre las variables *female* y *married*.

Por lo tanto, se estima ahora el siguiente modelo:

$$WAGE = \beta_1 + \beta_2 female + \beta_3 educ + \beta_4 exper + \beta_5 tenure + \beta_6 femalemarried + u$$

Donde:

*femalemarried*, es la variable dummy que representa la interacción entre las variables sexo y estado civil de las personas, y toma los valores de:

*femalemarried* = 1 cuando la persona es mujer casada.

*femalemarried* = 0 para el resto de la muestra.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

**Salida stata:** `reg wage female educ exper tenure femalemarried`

Source	SS	df	MS	Number of obs = 526		
-----+-----				F( 5, 520)	=	60.41
Model	2630.9083	5	526.181659	Prob > F	=	0.0000
Residual	4529.50599	520	8.71058845	R-squared	=	0.3674
-----+-----				Adj R-squared	=	0.3613
Total	7160.41429	525	13.6388844	Root MSE	=	2.9514
-----						
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
female	-1.447194	.3335762	-4.34	0.000	-2.102517	-.7918717
educ	.5808792	.0495127	11.73	0.000	.4836097	.6781487
exper	.0297398	.0117984	2.52	0.012	.0065613	.0529182
tenure	.1398767	.0211268	6.62	0.000	.0983724	.181381
femalemarr~d	-.6826192	.3820906	-1.79	0.075	-1.43325	.0680118
_cons	-1.756788	.7307182	-2.40	0.017	-3.19231	-.3212652
-----						

De la salida anterior, ahora se tiene una interpretación adicional, la de la variable con efecto interacción:

$$E[WAGE / female = 1, married = 1, educ, exper, tenure] = (\beta_1 + \beta_2 + \beta_3 educ + \beta_4 exper + \beta_5 tenure + \beta_6)$$

$$E[WAGE / female = 1, married = 1, educ, exper, tenure] = (-1.75 - 1.44 + 0.58educ + 0.029 exper + 0.13tenure - 1.75)$$

La expresión anterior, permite establecer el salario/hora promedio de las mujeres casadas, manteniendo constante el resto de variables.

### Cambios en intercepto y cambios en pendientes

Considérese en siguiente modelo:

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 D_1 + \beta_4 D_1 X_1 + u$$

Donde:

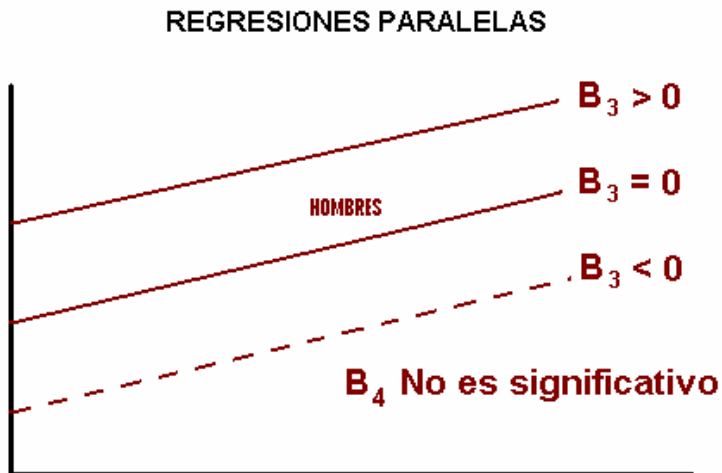
$Y$  Salario de las personas

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

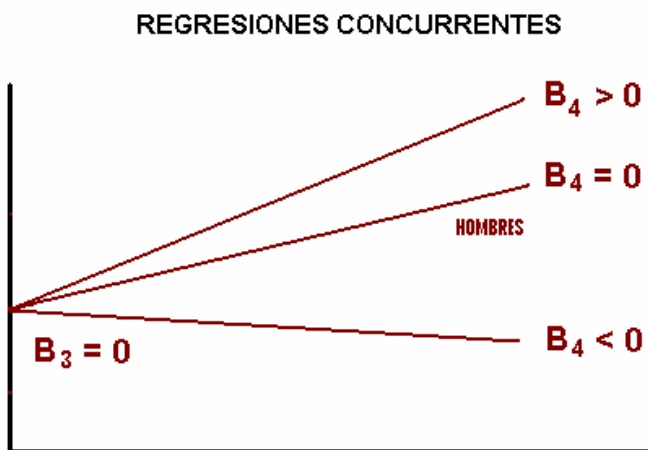
$X_1$  Nivel de escolaridad

$D_1$  Sexo, 1 = Mujeres, 0 los hombres

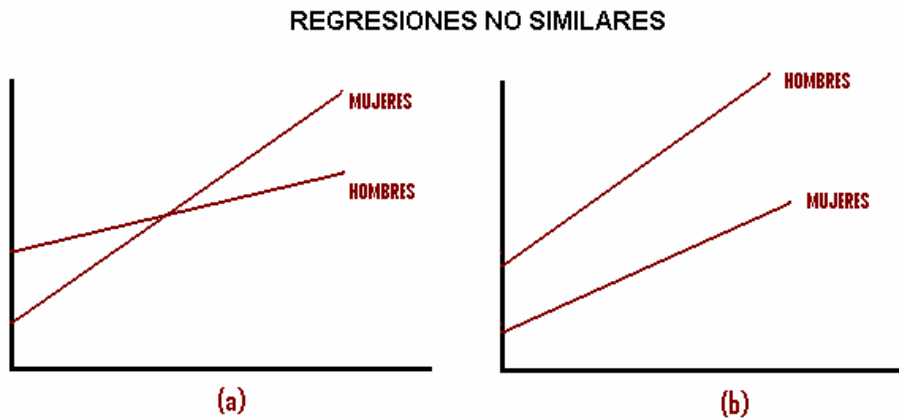
**Regresiones paralelas:** Cambio en intercepto e igual pendiente: Cuando  $\beta_3 < 0$  ó  $\beta_3 > 0$ , y  $\beta_2$  significativos,  $\beta_4$  no es significativo.



**Regresiones concurrentes:** Igual intercepto, cambio en pendiente: Cuando  $\beta_4 < 0$  ó  $\beta_4 > 0$  significativos, y  $\beta_3 = 0$ .



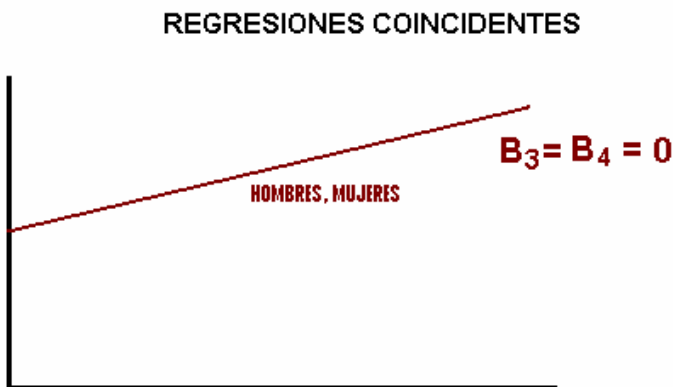
**Regresiones no similares:** Cambio en intercepto y pendiente:  $\beta_2$ ,  $\beta_3$  y  $\beta_4$  significativos.



En la gráfica (a), la intercepción de las mujeres es menor a la de los hombres, pero la pendiente es mayor para las mujeres. Esto significa que las mujeres ganan menos que los hombres en los niveles inferiores de educación, pero la brecha se cierra a medida que aumenta el grado de escolaridad. En algún punto los hombres y las mujeres ganan igual, y después las mujeres ganan mas que los hombres, dados los mismos niveles de educación.

La gráfica (b), muestra el caso en el que la intercepción de las mujeres es menor que la de los hombres, y la pendiente de la línea de las mujeres es menor que la de los hombres. Esto significa que las mujerees ganan menos que los hombres en cualquier nivel de escolaridad

**Regresiones coincidentes:** igual intercepto y pendiente:  $\beta_3 = 0$ ,  $\beta_4 = 0$ .



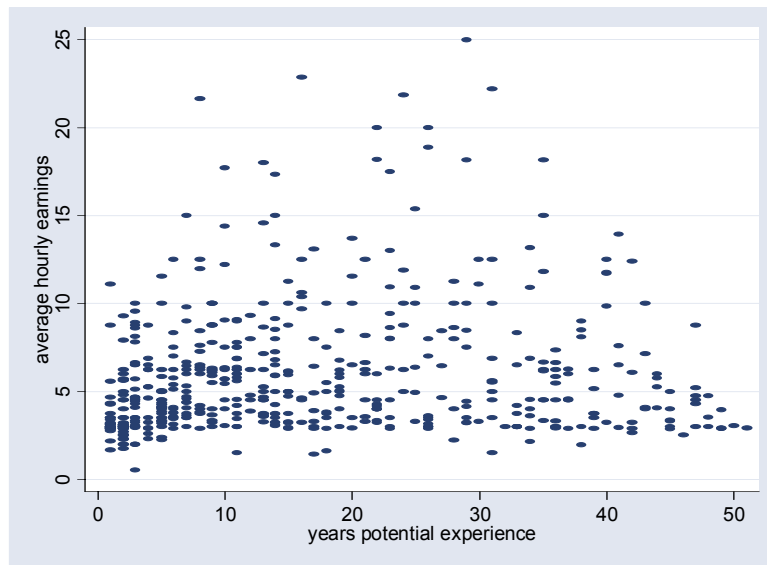
### GRAFICAS EN STATA

Consideremos el siguiente ejemplo:

$$WAGE = \beta_1 + \beta_2 \text{ exper} + u$$

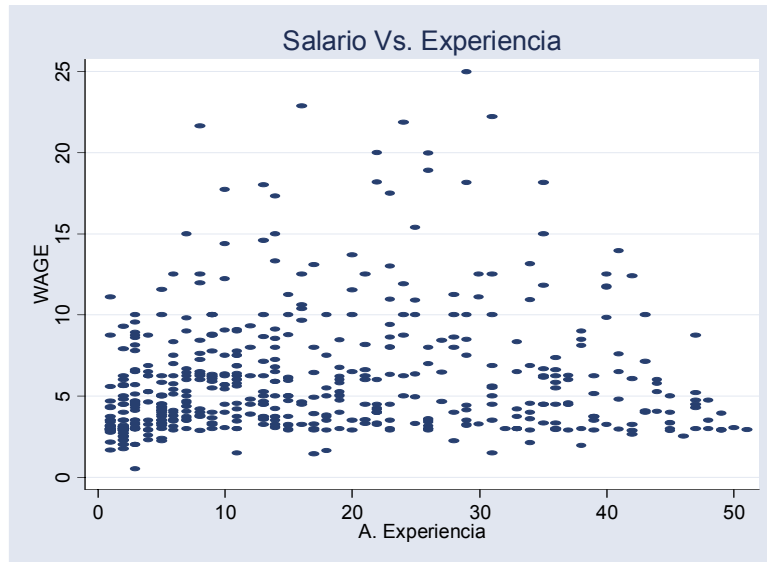
Si queremos graficar los puntos de esta regresión en el plano cartesiano, donde *exper* esta representado en el eje X y *WAGE* en el eje Y, se utiliza el siguiente comando:

**scatter wage exper**



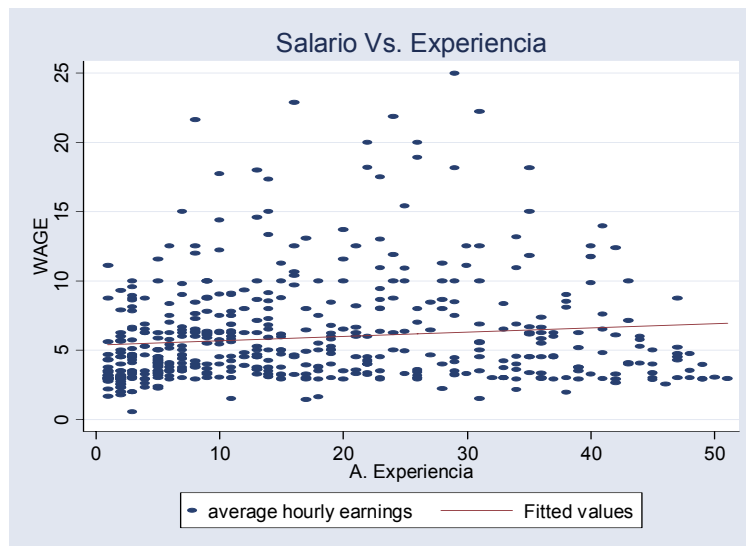
Ahora, si se quiere adicionar a la gráfica título a la gráfica y los nombres a cada uno de los ejes, se utiliza el comando **title**:

**scatter wage expe, title(Salario Vs. Experiencia) xtitle(A. Experiencia) ytitle(WAGE)**



Por último, si se quiere adicionar la línea de tendencia, se utiliza el comando:

**twoway (scatter wage exper) (lfit wage exper), title(Salario Vs. Experiencia) xtitle(A. Experiencia) ytitle(WAGE)**



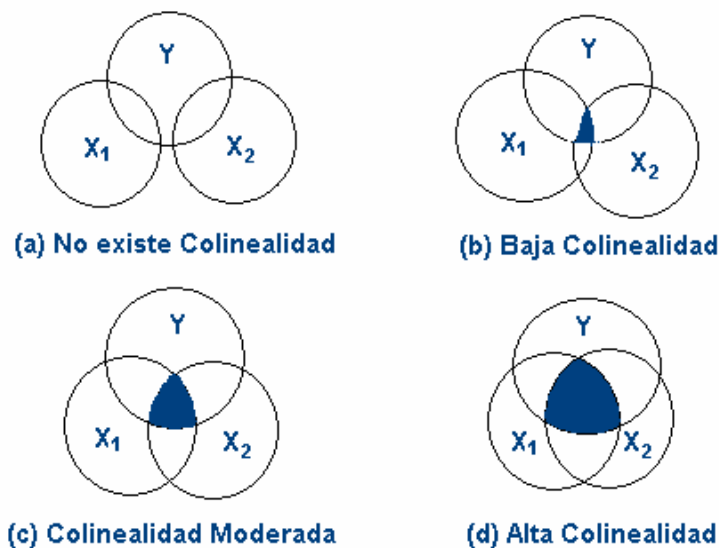


## PROBLEMAS DE ESPECIFICACIÓN DEL MODELO DE REGRESIÓN MÚLTIPLE.

### MULTICOLINEALIDAD

El décimo supuesto del modelo clásico de regresión lineal (MCRL), plantea que no existe multicolinealidad entre las variables explicativas del modelo, es decir, que no debe existir relación perfecta o exacta entre algunas o todas las variables explicativas de un modelo de regresión<sup>4</sup>. Existen otros supuestos que son complementarios a no la multicolinealidad, ellos son el supuesto 7, que indica que el número de regresores debe ser menor al número de observaciones, y el supuesto 8, que especifica que debe existir variabilidad en los valores de los regresores.

Para entender un poco este concepto, se puede representar gráficamente:



Es importante resaltar que la multicolinealidad, como se ha definido, se refiere solamente a relaciones lineales entre las variables explicativas, no elimina las relaciones no lineales existentes entre ellas. Por lo tanto modelos como:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$$

No violan el supuesto de no multicolinealidad. Sin embargo, se encontraran coeficientes de correlación altos para variables  $X_i, X_i^2, X_i^3$ .

<sup>4</sup> Se habla de multicolinealidad cuando hay existencia de más de una relación lineal exacta, y colinealidad se refiere a la existencia de una sola relación lineal.

**NOTA:** Los estimadores de MCO en presencia de multicolinealidad, satisfacen la condición de ser el Mejor estimador lineal e insesgado (MELI) ó MEI (en el caso de añadir el supuesto de normalidad).

### Consecuencias Prácticas de la Multicolinealidad

1. Varianzas de los coeficientes estimados sobredimensionados.
2. Pruebas de hipótesis no validas.
3. Se podría dar error tipo II (Aceptar la hipótesis nula, dado que es falsa)
4. Intervalos de Confianzas más amplios.
5. No se podrían separar los efectos de una variable explicativa sobre la variable dependiente. Por ejemplo, se tiene  $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + u$ , y  $X_1$  esta relacionado con  $X_2$ , no se puede determinar con certeza cual es el efecto de  $\beta_2$  sobre  $Y$ , ya que existe un efecto también de  $\beta_3$ .
6. Contradicciones en las estimaciones realizadas al modelo, se encuentran pruebas  $t$  bajas y estadísticos  $F$  y  $R^2$  altos.
7. Sensibilidad del modelo, los estimadores y sus errores estándar se tornan muy sensibles. Por ejemplo, estimo un modelo de regresión con 90 observaciones, elimino algunas observaciones y al realizar las estimaciones nuevamente, encuentro signos de los betas distintos.

### Como detectar la Multicolinealidad

Teniendo en cuenta que la multicolinealidad es un fenómeno esencialmente de tipo muestral<sup>5</sup>, no se tiene un método único para detectarla. Lo que se tienen, son ciertas pruebas formales e informales que permiten determinar su presencia. Algunas de ellas son:

1. Observar los estadísticos estimados: Cuando se tiene un  $R^2$  alto, y la prueba  $F$  muestra que el modelo es globalmente significativo, es decir, los coeficientes estimados del modelo son estadísticamente diferentes de cero, pero se encuentran unos  $t$  calculados bajos, que demuestran que los coeficientes no son significativos ( $\beta_i = 0$ ).
2. Observar la matriz de correlación entre parejas de regresores: Si este coeficiente es mayor a 0.8, entonces la multicolinealidad es un problema grave. Sin embargo, esta condición se puede considerar suficiente pero no necesaria, debido a que la multicolinealidad puede existir, a pesar de que las correlaciones sean comparativamente bajas (es decir, inferiores a 0.5). Si el coeficiente de correlación es -1, existe correlación negativa, a medida que una variable aumenta, la otra disminuye. Cuando el coeficiente es 1, hay correlación positiva, cuando aumenta una variable, también aumenta la otra.

---

<sup>5</sup> Aun cuando las variables  $X$  no están linealmente relacionadas en la población, pueden estarlo en la muestra que se ha tomado, en dicho caso, nuestra muestra puede no ser lo suficiente rica para acomodar todas las variables  $X$  en el análisis.

3. Regresiones auxiliares: Dado que la multicolinealidad surge por la relación lineal entre variables explicativas, se pueden estimar regresiones entre las variables explicativas. Posteriormente, se puede adoptar la regla práctica de Klien, quien sugiere que si el  $R^2$  obtenido en la regresión auxiliar es mayor que el  $R^2$  global, es decir, aquel obtenido en la regresión de  $Y$  sobre todos los regresores, hay un serio problema de multicolinealidad.
4. Estimar el Factor de Inflación de Varianza (FIV). El cual está dado por;  
$$FIV = \frac{1}{1 - R_j^2}$$
 , donde  $R_j^2$  es el  $R^2$  de la regresión auxiliar.

Si el  $FIV$  es superior a 10, se dice que esa variable es altamente colineal.

Algunos autores, hacen referencia a la medida de tolerancia para determinar la multicolinealidad. Esta se define como:

$$Tolerancia = (1 - R_j^2)$$

Si la tolerancia tiende a 1, entonces no hay multicolinealidad.

Si la tolerancia tiende a 0, entonces hay multicolinealidad.

### Ejemplo Práctico:

Dadas las observaciones de la base de datos Gastoconsumo.dta, estímesese el siguiente modelo:

$$gastoconsumo = \beta_1 + \beta_2 ingreso + \beta_3 riqueza + u$$

Donde:

$gastoconsumo$  , es la variable dependiente que representa el gasto de consumo familiar semanal.

$ingreso$  , es la variable independiente que representa el ingreso semanal familiar.

$riqueza$  , es la variable independiente que representa la riqueza de la familia.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Salida en Stata: `reg gastoconsumo ingreso riqueza`

Source		SS		df		MS		Number of obs =	10
-----+-----								F( 2, 7) =	92.40
Model		8565.55407		2		4282.77704		Prob > F	= 0.0000
Residual		324.445926		7		46.349418		R-squared	= 0.9635
-----+-----								Adj R-squared =	0.9531
Total		8890		9		987.777778		Root MSE	= 6.808
-----									
gastoconsumo		Coef.		Std. Err.		t	P> t	[95% Conf. Interval]	
-----+-----									
ingreso		.9415373		.8228983		1.14	0.290	-1.004308	2.887383
riqueza		-.0424345		.0806645		-0.53	0.615	-.2331757	.1483067
_cons		24.77473		6.7525		3.67	0.008	8.807609	40.74186
-----									

Teniendo en cuenta las estimaciones anteriores, procedemos a identificar si existe en este caso colinealidad entre las variables *ingreso* y *riqueza*. Para ello, tendremos en cuenta cada uno de los puntos mencionados para detectar la multicolinealidad.

1. De la regresión anterior se observa que las variables *ingreso* y *riqueza* explican en un 96% los gastos en consumo semanal familiar (puesto que el  $R^2 = 0.9635$ ). También se evidencia que el valor del estadístico  $F = 92.40$ , es alto, lo cual indica que cae en la zona de rechazo, indicando que el modelo es globalmente significativo.

Sin embargo, ninguno de los coeficientes de las pendientes es estadísticamente significativo, lo cual indica que dichas variables están altamente correlacionadas y es imposible aislar el impacto individual del ingreso o la riqueza sobre el consumo. Asimismo, el signo de la variable *riqueza* no es el esperado (se espera que la relación sea positiva).

2. Al obtener la matriz de correlación de las variables:  
Matriz de correlación: `cor gastoconsumo ingreso riqueza`

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

```

               | gastoc~o ingreso riqueza
-----+-----
gastoconsumo |    1.0000
            ingreso |    0.9808    1.0000
            riqueza |    0.9781    0.9990    1.0000

```

Se muestra que el coeficiente de correlación entre la variable *ingreso* y *riqueza* es bastante alto (0.9990) cercano al 1. Como se había mencionado antes, si el coeficiente era mayor a 0.8, se evidencia un problema de colinealidad entre dichas variables. La correlación es positiva, a medida que aumenta el ingreso, aumenta la riqueza.

3. Realicemos la siguiente regresión auxiliar:

$$\text{ingreso} = \beta_1 + \beta_2 \text{riqueza} + u$$

**Salida en Stata:** `reg ingreso riqueza`

```

Source |      SS      df      MS                Number of obs =      10
-----+-----
Model | 32931.5534      1 32931.5534            F( 1,      8) = 3849.02
Residual | 68.4466181      8  8.55582726          Prob > F      =  0.0000
-----+-----
Total |    33000      9 3666.66667          R-squared      =  0.9979
                                           Adj R-squared =  0.9977
                                           Root MSE      =   2.925

-----+-----
ingreso |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
riqueza |   .0979231   .0015784    62.04   0.000    .0942834   .1015629
   _cons |  -.3862708   2.897956   -0.13   0.897   -7.068968   6.296427
-----+-----

```

La salida muestra que existe colinealidad casi perfecta entre las variables *ingreso* y *riqueza*. Asimismo, si realizamos la regla práctica de Klien, al comparar el  $R^2$  obtenido en la regresión auxiliar y el  $R^2$  global, se tiene que el primero (0.9979) es mayor que el segundo (0.9635); por tanto, hay un serio problema de colinealidad.

4. Calculemos el *FIV* y la *tolerancia* :

$$FIV = \frac{1}{1 - R_j^2} = \frac{1}{1 - 0.9970} = 333,33$$

Como el *FIV* > 10 , se demuestra colinealidad nuevamente.

Ahora estimamos la *tolerancia* :

$$Tolerancia = (1 - R_j^2) = 1 - .9970 = 0.003$$

Como la *tolerancia* esta cercana a cero, se puede decir que hay colinealidad casi perfecta entre el *ingreso* y la *riqueza* .

### Medidas Remédiales

Teniendo en cuenta nuevamente que la multicolinealidad es esencialmente un problema muestral, no hay un método específico. Sin embargo, existen algunas reglas prácticas, que son:

1. Eliminación de una(s) variable(s) y el sesgo de especificación: Una de las soluciones más simples para solucionar el problema de la multicolinealidad es la eliminación de una de las variables que causa este tipo de violación de los supuestos.

Sin embargo, se debe tener en cuenta no caer en el sesgo de especificación del modelo, con el fin de no violar otro de los supuestos del modelo de regresión clásico (sesgo de especificación). Para esto, es importante tener en cuenta que dice la teoría económica con respecto a la explicación de la variable dependiente.

Por ejemplo, retomando el modelo de regresión:

$$gastoconsumo = \beta_1 + \beta_2 \text{ingreso} + \beta_3 \text{riqueza} + u$$

En donde se evidenció la presencia de la colinealidad entre variables *ingreso* y *riqueza* , se procede a eliminar en primera instancia la variable *riqueza* .

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

**Salida en Stata:** `reg gastoconsumo ingreso`

Source	SS	df	MS	Number of obs = 10		
-----+-----				F( 1, 8)	=	202.87
Model	8552.72727	1	8552.72727	Prob > F	=	0.0000
Residual	337.272727	8	42.1590909	R-squared	=	0.9621
-----+-----				Adj R-squared	=	0.9573
Total	8890	9	987.777778	Root MSE	=	6.493
-----						
gastoconsumo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
ingreso	.5090909	.0357428	14.24	0.000	.4266678	.591514
_cons	24.45455	6.413817	3.81	0.005	9.664256	39.24483

Se observa que ahora la variable ingreso es estadísticamente significativa.

Ahora, estimaremos el gasto del consumo familiar en función de la *riqueza*, es decir, eliminamos la variable *ingreso*.

**Salida en Stata:** `reg gastoconsumo riqueza`

Source	SS	df	MS	Number of obs = 10		
-----+-----				F( 1, 8)	=	176.67
Model	8504.87666	1	8504.87666	Prob > F	=	0.0000
Residual	385.123344	8	48.1404181	R-squared	=	0.9567
-----+-----				Adj R-squared	=	0.9513
Total	8890	9	987.777778	Root MSE	=	6.9383
-----						
gastoconsumo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
riqueza	.0497638	.003744	13.29	0.000	.0411301	.0583974
_cons	24.41104	6.874097	3.55	0.007	8.559349	40.26274

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Se evidencia que la variable *riqueza* ahora es altamente significativa.

Sin embargo, como se menciono anteriormente, es importante tener claro lo que dice la teoría económica, con el fin de no caer en el sesgo de especificación del modelo.

2. Información a priori: La información a priori puede provenir de trabajo empírico que se haya realizado con anterioridad. Continuando con el ejemplo del gasto en el consumo familiar, se podría tener información a priori que  $\beta_3 = 0.1\beta_2$ , es decir, que la tasa de consumo con respecto a la riqueza es una décima parte de la correspondiente con respecto al ingreso.

Teniendo en cuenta lo anterior se tiene que:

$$\text{gastoconsumo} = \beta_1 + \beta_2 \text{ingreso} + \beta_3 \text{riqueza} + u$$

$$\text{gastoconsumo} = \beta_1 + \beta_2 \text{ingreso} + 0.1\beta_2 \text{riqueza} + u$$

$$\text{gastoconsumo} = \beta_1 + \beta_2 X + u$$

Donde:  $X = (\text{ingreso} + 0.1\text{riqueza})$

Generamos la variable X: **gen X= ingreso+0.1\*riqueza**

Ahora estimamos el modelo.

**Salida es Stata:** `reg gastoconsumo X`

Source	SS	df	MS	Number of obs =	10
-----+-----				F( 1, 8) =	191.20
Model	8532.97312	1	8532.97312	Prob > F =	0.0000
Residual	357.026877	8	44.6283596	R-squared =	0.9598
-----+-----				Adj R-squared =	0.9548
Total	8890	9	987.777778	Root MSE =	6.6804
-----					
gastoconsumo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
X	.2517807	.0182086	13.83	0.000	.2097915 .2937699
_cons	24.38745	6.610424	3.69	0.006	9.14378 39.63111
-----					



Se observa que el beta que acompaña la variable  $X$ , es altamente significativo. Como sabemos el valor de  $\beta_2$ , se puede estimar  $\beta_3$ , a partir de la relación postulada a priori.

Por tanto,  $\beta_3 = 0.1(0.25) = 0,02517$

3. Transformación de las variables: En ocasiones, se pueden realizar transformación de las variables, con el fin de evitar la colinealidad. Por ejemplo, se podría estimar un modelo donde una de las variables que causan la colinealidad este en logaritmo.

4. Datos nuevos o adicionales: Teniendo en cuenta que la multicolinealidad es de la muestra, se puede pensar que tomando una nueva muestra de las mismas variables, o ampliando el tamaño de la misma, se puede atenuar el problema de la colinealidad.

## HETEROCEDASTICIDAD

El modelo de regresión lineal múltiple exige que la varianza condicional de las perturbaciones aleatorias a los valores de la variable explicativas “X” sea constante:

Homocedasticidad:  $E(U_i^2) = \sigma^2$

Heterocedasticidad:  $E(U_i^2) \neq \sigma^2$

Las fuentes de la Heterocedasticidad se puede atribuir a:

- Factores exclusivos de la regresión.
- Errores de explicación del modelo.
- Irregularidad en la distribución de las variables.
- Errónea transformación de la forma funcional del modelo

Las propiedades que tienen los estimadores se enumeran a continuación.

1. Los estimadores siguen siendo INSESGADOSE  $(\beta) = \beta \rightarrow$  Condición de Insesgamiento.
2. Los estimadores siguen siendo CONSISTENTES.
3. La propiedad de Consistencia es de las muestras grandes y consiste en que la Varianza de  $\beta$  tiende a cero cuando  $n$  tiende a  $\infty$ . Bajo el supuesto de heterocedasticidad se sigue cumpliendo.
4. Los estimadores dejan de ser EFICIENTES ya que no son los de mínima varianza.
5. Las varianzas y covarianzas de los estimadores de MCO son SESGADAS e INCONSISTENTES. Por este motivo los test de hipótesis ya no son válidos.

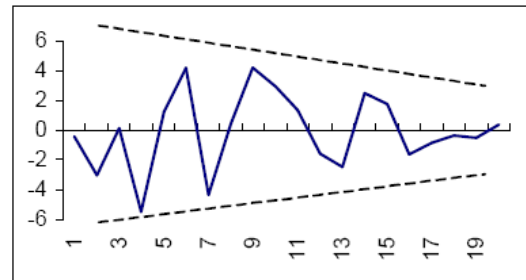
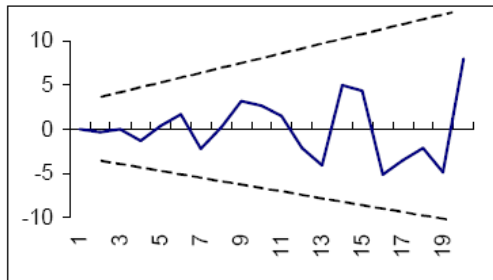
Como detectar la Heterocedasticidad:

Es imposible encontrar la presencia de heterocedasticidad ya que, en la mayoría de los análisis regresiones múltiples, sólo dispondremos de un valor de “Y” para cada valor de “X” por lo que se obtiene que resulta imposible observar si la varianza de las “U” para cada valor de “X” es la misma.

- Existe un comando en el paquete **STATA**, en cual colocamos el comando “`hettest`”, por medio de este comando se pretende aceptar la hipótesis nula.

Heterocedasticidad PRETENDE ACEPTAR $H_0$	
$H_0 = \sigma^2 = \sigma^2$	<p><math>P - \text{valor} &gt; (\alpha)</math> significancia.</p> <p>No rechazo o acepto la hipótesis nula y rechazo la hipótesis alternativa.</p> <p>Lo que quiero es que pase.</p>
$H_A = \sigma^2 \neq \sigma^2$	$P - \text{valor} < (\alpha)$ significancia.

- Existe una metodología a través de los gráficos, la cual consiste en hacer una regresión bajo el supuesto de la homocedasticidad y luego examinar los errores “u” con la variables regresoras y observamos comportamiento de los gráficos.



- TEST DE GOLDFELD-QUANT: Se basa en la idea que si la varianza de los errores es igual a través de todas las observaciones, entonces la varianza para una parte de la muestra será la misma que la calculada con otra parte de la misma.
  - Se identifica una variable Z relacionada con la varianza de los errores. Si suponemos que la relación es POSITIVA, ordenamos de manera creciente los datos de la muestra.
  - Dividimos la muestra en 2 partes omitiendo los valores centrales.
  - Estimamos las regresiones por separado.
  - Obtenemos SEC de cada una de las regresiones y calculamos las estimaciones de la varianza como  $SEC1/n1-k$  y  $SEC2/n2-k$ .
  - Calculamos  $F_{calc} = \frac{SEC1/n-k}{SEC2/n-k}$
  - Comparamos  $F_{calc}$  con el valor F tabla con  $(n1-K)$  GL numerador y  $(n2-K)$  GL denominador.
  - Si  $F_{calc} > F_{tabla}$  rechazo  $H_0$  de Homocedasticidad.

El éxito depende de este Test es seleccionando correctamente la “X”.

- TEST DE WHITE: También es un test para muestras grandes no necesita ningún supuesto previo acerca de las causas de la heterocedasticidad.

1. Estimamos el modelo por MCO.
2. Calculamos  $U_i^2$  (estimado).
3. Estimamos un modelo de regresión utilizando  $U_i^2$  (estimado) como variable dependiente sobre las X originales, las X y los productos cruzados.
4. Calculamos  $R^2$  para la regresión y  $n.R^2$ .

$$5. \quad H_0: \alpha_2 = \alpha_3 = \dots = 0$$

$$H_1: \text{al menos una } \alpha \neq 0$$

$$\text{Si } n R^2 > \chi^2(k-1), \alpha$$

Rechazo  $H_0$  y tengo Heterocedasticidad.

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \dots + \alpha_k x_{ki} + \alpha_{k+1} x_{1i}^2 + \dots + \alpha_{k+k} x_{ki}^2 + \alpha_{k+k+1} x_{1i} x_{2i} + \dots + \alpha_{k+k+2} x_{1i} x_{3i} + \dots + \alpha_{k+k+1} x_{2i} x_{3i} + \dots + \varepsilon_i$$

## Soluciones a la Heterocedasticidad

Mínimos Cuadrados Generalizados : Consiste en dividir cada término por  $\sigma_i$ .

Modelo transformado

$$Y_1/\sigma_i = \beta_1 X_1/\sigma_i + \beta_2 X_2/\sigma_i + \dots$$

Este modelo satisface los supuestos de MCO, pero se puede presentar el inconveniente de no conocer  $\sigma_i$ .

Mínimos Cuadrados Ponderados: es una extensión del MCG.

Definimos  $w_i = 1/\sigma_i$ . Y transformando el modelo nos queda

$$Y_1 W_1 = \beta_1 (X_1 W_1) + \beta_2 (X_2 W_1) + \dots + (U_i W_1)$$

En este modelo transformado cada observación de la variable está ponderada por  $W_1$  (inversamente proporcional a  $\sigma_i$ )

Conocemos la estructura de la Heterocedasticidad.

Suponemos  $\text{Var}(U_i) = \sigma_2 \cdot Z_2$  ( se denomina Heterocedasticidad Multiplicativa)  $W = 1/Z \dots$   
Nos queda el modelo transformado.

La Var ( $U_i$  transformado) =  $\sigma_2$ .(porque se nos elimina  $Z_2$ ) , de esta manera nos queda un modelo Homocedastico.

## AUTOCORRELACION

La autocorrelación se puede definir como la correlación entre miembros de series de observaciones ordenadas en el tiempo (información de series de tiempo) o en el espacio (información de corte transversal). El modelo de regresión lineal supone que no debe existir autocorrelación en los errores ( $u_i$ ), es decir, el término de perturbación relacionado con una observación cualquiera no debería estar influenciado por el término de perturbación relacionado con cualquier otra observación.

$$E(u_i u_j) = 0 \text{ para todo } i \neq j$$

### Causas de la Autocorrelación

Algunas de las causas son las siguientes<sup>6</sup>:

**Trabajo con datos de serie temporal:** cuando se trabaja con datos de corte longitudinal (p.e.: una variable explicativa cuyas observaciones correspondan a valores

obtenidos en instantes temporales sucesivos), resulta bastante frecuente que el término de perturbación en un instante dado siga una tendencia marcada por los términos de perturbación asociados a instantes anteriores. Este hecho da lugar a la aparición de autocorrelación en el modelo.

**Especificación errónea en la parte determinista del modelo (autocorrelación espuria):**

1. Omisión de variables relevantes: en tal caso, las variables omitidas pasan a formar parte del término de error y, por tanto, si hay correlación entre distintas observaciones de las variables omitidas, también la habrá entre distintos valores de los términos de perturbación.
2. Especificación incorrecta de la forma funcional del modelo: si usamos un modelo inadecuado para describir las observaciones (p.e.: un modelo lineal cuando en realidad se debería usar un modelo cuadrático), notaremos que los residuos muestran comportamientos no aleatorios (i.e.: están correlacionados).

**Transformaciones de los datos:** determinadas transformaciones del modelo original podrían causar la aparición de autocorrelación en el término de perturbación del modelo transformado (incluso cuando el modelo original no presentase problemas de autocorrelación).

**Trabajo con modelos dinámicos:** cuando se trabaja con series temporales suele ser habitual considerar modelos de regresión que incluyan no sólo los valores actuales sino también los valores retardados (pasados) de las variables explicativas. Es el caso de un **modelo de retardos distribuidos de orden s** o **RD(s)**:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + \beta_s X_{t-s} + u_t$$

---

<sup>6</sup> Novales, A. (1993): "Econometría". McGraw-Hill. ISBN 84-481-0128-6

Otro tipo de modelo dinámico que presentaría problemas de autocorrelación sería aquel que incluyese entre sus variables explicativas uno o más valores retardados de la variable dependiente. Este otro tipo de modelo dinámico se conoce como **modelo autorregresivo de orden s o AR(s)**:

$$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \dots + \gamma_s Y_{t-s} + u_t$$

Otra causa común de la autocorrelación es la existencia de tendencias y ciclos en los datos. Es decir, la mayoría de las variables económicas no son estacionarias en media. Esto significa que si la variable endógena del modelo tiene una tendencia creciente o presenta un comportamiento cíclico que no es explicado por las exógenas, el término de error recogerá ese ciclo o tendencia.

### Consecuencias de la Autocorrelación:

La consecuencia más grave de la autocorrelación de las perturbaciones es que la estimación MCO deja de ser eficiente y la inferencia estadística también se verá afectada. Las consecuencias dependen del tipo de autocorrelación (positiva o negativa):

1. Cuando se tiene autocorrelación positiva, la matriz de varianza y covarianza de los residuos esta subestimada, si el tipo de autocorrelación es negativa, se tiene una sobrestimación de la misma.
2. Cuando se tiene autocorrelación positiva, la matriz de varianza y covarianza de los coeficientes (betas) esta subestimada, si el tipo de autocorrelación es negativa, se tiene una sobrestimación de la misma.
3. Cuando se tiene autocorrelación positiva, los intervalos de confianza son angostos, si el tipo de autocorrelación es negativa, se tienen intervalos de confianza más amplios.
4. Cuando se tiene autocorrelación positiva, se tiende a cometer error tipo I (rechazar la hipótesis nula cuando es verdadera), si el tipo de autocorrelación es negativa, se tiende a cometer error tipo II (no rechazar la hipótesis nula cuando es falsa).
5. Los  $\beta_{MCO}$  son lineales, insesgados, pero ineficientes (no tienen varianza mínima).
6. Las pruebas  $t$  y  $F$  pierden validez.

### Detección de la Autocorrelación:

Para analizar la posible presencia de autocorrelación 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 Durbin-Watson, test h de Durbin, test de Breusch-Godfrey, test Q de Box-Pierce, etc.).

### Análisis Gráfico:

Al realizar la regresión por MCO, se pueden graficar los residuos (o, alternatively, los residuos estandarizados, es simplemente dividir  $u_t$  por el error estandar de la estimación  $\hat{\sigma}$ ) frente al tiempo. Dado que los residuos MCO son estimadores consistentes de los términos de perturbación, si se aprecian en el gráfico anterior patrones de comportamiento sistemático (no aleatorio) podremos afirmar que los términos de perturbación presentan algún tipo de autocorrelación.

### Contrastes:

#### Test de Durbin-Watson

Es la prueba mas conocida para detectar correlación serial; permite contrastar si el término de perturbación está autocorrelacionado. Dicha prueba presenta algunos supuestos:

Es válido para autocorrelación serial de 1° orden en los residuos, no aplica para modelos con variable dependiente rezagada como variable explicativa, las variables explicativas son no estocásticas (son fijas en muestreo repetido), el modelo de regresión lineal debe incluir el intercepto, y no hay observaciones faltantes en los datos.

$$d = \frac{\sum (u_t - u_{t-1})^2}{\sum u_t^2} \cong 2(1 - \hat{\rho})$$

Una vez hallado DW, es posible usar su valor para estimar el coeficiente de autocorrelación simple  $\rho$  mediante la expresión:

$$\hat{\rho} \cong 1 - \frac{d}{2}$$

El estadístico DW es un valor comprendido entre 0 y 4. Como se observa en el siguiente gráfico, para valores de DW cercanos a 2 no rechazaremos la hipótesis nula, por el contrario, para valores de DW alejados de 2, sí rechazaremos la hipótesis nula

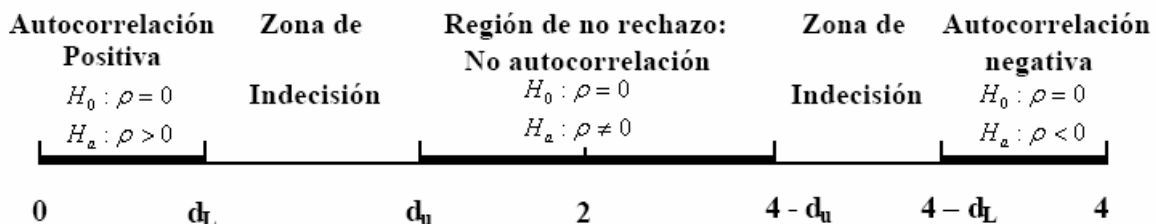


Tabla de decisión:

$0 < d < d_L$ , se rechaza  $H_0$ , existe autocorrelación positiva.

$4 - d_L < d < 4$ , se rechaza  $H_0$ , existe autocorrelación negativa.

$d_u < d < 4 - d_u$ , no se rechaza  $H_0$ , no existe autocorrelación.

$d_L < d < d_u$  o  $4 - d_u < d < 4 - d_L$ , el contraste no es concluyente.

Los pasos a seguir de este contraste son:

1. Estimación por mínimos cuadrados ordinarios (MCO) del modelo de regresión.
2. Cálculo de los residuos MCO.
3. Obtención del estadístico  $d$  (experimental) de Durbin-Watson.
4. Búsqueda de los *niveles críticos* del contraste.
5. Aplicación de la regla de decisión.

Un inconveniente que presenta este contraste es que a veces puede no ser concluyente, por lo que hay que considerar, utilizando otros criterios, si existe o no autocorrelación.

### Ejemplo en Stata:

Se trabajara con la base de datos PHILLIPS.DTA, la cual contiene las siguientes variables:

*year*, indica el año.

*inf*, es la tasa de inflación.

*unem*, es la tasa de desempleo.

Con el fin de realizar estimaciones de series de tiempo en Stata, es importante escribir el siguiente comando:

***tsset year***

Donde *year* es la variable que contiene los años.

Automáticamente el sistema reconoce la serie de tiempo, y muestra:



## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

*time variable: year, 1948 to 1996*

**Salida en Stata:** `reg inf unem`

Source	SS	df	MS	Number of obs = 49		
-----+-----				F( 1, 47)	=	2.62
Model	25.6369575	1	25.6369575	Prob > F	=	0.1125
Residual	460.61979	47	9.80042107	R-squared	=	0.0527
-----+-----				Adj R-squared	=	0.0326
Total	486.256748	48	10.1303489	Root MSE	=	3.1306
-----						
inf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
unem	.4676257	.2891262	1.62	0.112	-.1140213	1.049273
_cons	1.42361	1.719015	0.83	0.412	-2.034602	4.881822
-----						

Una vez estimada la regresión, se procede a ejecutar el siguiente comando con el cual se obtiene el estadístico Durbin-Watson:

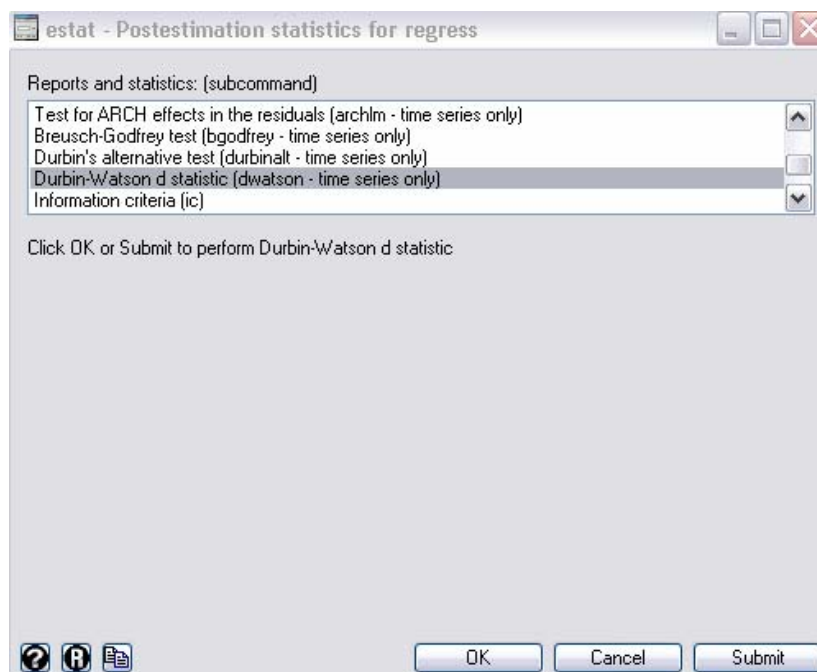
***estat dwatson o dwstat***

***Durbin-Watson d-statistic( 2, 49) = .8027005***

Si se quiere estimar el Durbin-Watson por las ventanas en Stata 9, la ruta a seguir es:

**Statistics/time-series/tests/time series epecification tests after regress**

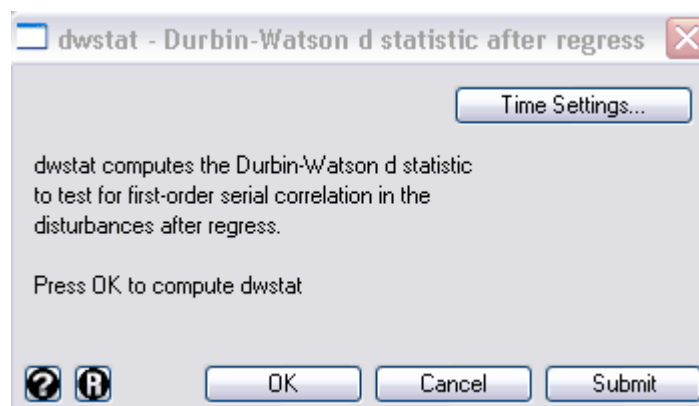
Automáticamente se despliega el siguiente recuadro, en donde se muestra la opción a seleccionar, y le damos **OK**.



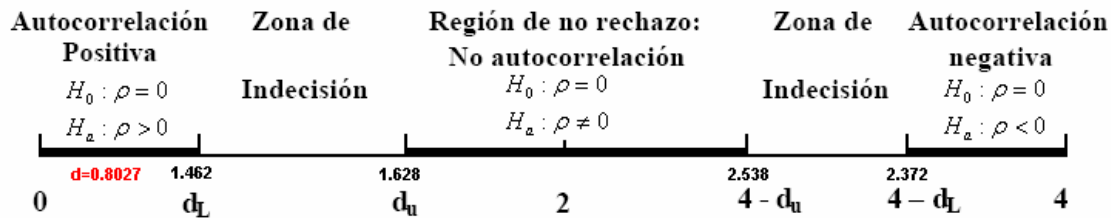
La ruta a seguir en Stata 8.2 es:

### **Statistics/time-series/tests/Durbin-Watson d statistics after regress**

Automáticamente se despliega el siguiente recuadro, en donde se muestra la opción a seleccionar, y le damos **OK**.



Teniendo en cuenta que DW es 0.8027, gráficamente se tiene:



Por tanto se rechaza la hipótesis nula, hay autocorrelación.

### Prueba de Breusch – Godfrey (BG) sobre autocorrelación de orden superior

Este estadístico es muy sencillo de calcular y resuelve los problemas del contraste de Durbin-Watson; por ejemplo, los regresores incluidos en el modelo pueden contener valores rezagados de la variable dependiente, es decir,  $Y_{t-1}, Y_{t-2}$ , etc. Pueden aparecer como variables explicativas.

Supóngase que el termino de perturbación  $u_t$  es generado por el siguiente esquema autorregresivo de orden  $p$  :

$$u_t = \theta_1 u_{t-1} + \theta_2 u_{t-2} + \dots + \theta_p u_{t-p} + \varepsilon_t$$

Donde  $\varepsilon_t$  es un término de perturbación puramente aleatorio con media cero y varianza constante.

Dado el modelo anterior, la hipótesis será:

$H_0 : \theta_1 = \theta_2 = \dots = \theta_p = 0$  No hay autocorrelación de ningún orden.

$H_a : \text{ Hay autocorrelacion}$

Dicha hipótesis puede ser probada de la siguiente manera:

1. Estimación por MCO del modelo de regresión y obtención de los residuos MCO ( $\hat{u}_t$ ).
2. Estimación de una regresión auxiliar de los residuos  $\hat{u}_t$  sobre  $p$  retardos de los mismos,  $\hat{u}_{t-1}, \hat{u}_{t-2}, \dots, \hat{u}_{t-p}$ .
3. Obtención del coeficiente de determinación ( $R^2$ ) de la regresión auxiliar ( $R_{aux}^2$ ).
4. Si el tamaño de la muestra es grande, Breusch y Golfrey han demostrado que:  $(n - p) * R_{aux}^2$  se distribuye con  $\chi^2$  con  $p$  g.l.

5. Si el valor calculado excede el valor crítico de  $\chi^2$  al nivel de significancia seleccionado, se puede rechazar la hipótesis nula, en cuyo caso, por lo menos un  $\theta$  es significativamente diferente de cero (se admite que hay autocorrelación), en caso contrario no habría autocorrelación.

## Ejemplo en Stata:

El comando a ejecutar es:

***estat bgodfrey* o *bgodfrey***

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	18.472	1	0.0000

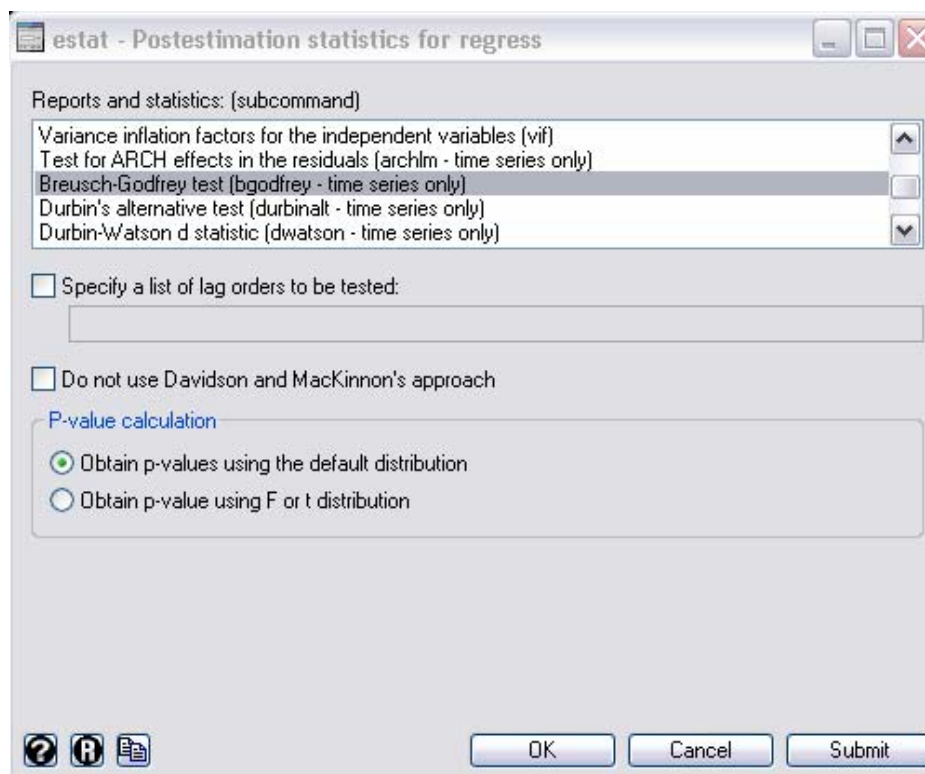
H0: no serial correlation

De acuerdo a la salida anterior, se puede observar que el p-valor asociado al  $\chi^2$  es 0.000, lo cual confirma la presencia de autocorrelación.

Si se quiere estimar la prueba Breusch – Godfrey por las ventanas en Stata 9, la ruta a seguir es:

**Statistics/time-series/tests/time series epecification tests after regress**

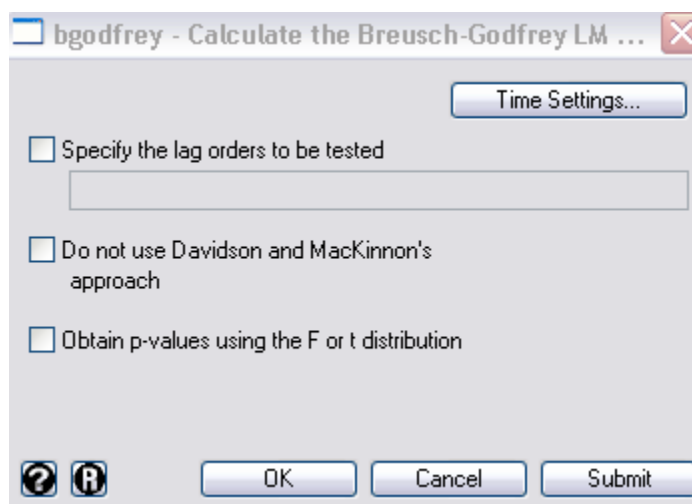
automáticamente se despliega el siguiente recuadro, en donde se muestra la opción a seleccionar, y le damos **OK**.



La ruta a seguir en Stata 8.2 es:

## Statistics/time-series/tests/Breusch-Godfrey LM test for autocorrelation

Automáticamente se despliega el siguiente recuadro, en donde se muestra la opción a seleccionar, y le damos **OK**.



**Como solucionar la autocorrelación**

**Cuando  $\hat{\rho} \cong 1 - \frac{d}{2}$  es conocido:**

1. se tiene:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (\text{a})$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \quad (\text{b})$$

2. Multiplico **(b)** por  $\rho$ , y se tiene:

$$\rho Y_{t-1} = \rho \beta_1 + \rho \beta_2 X_{t-1} + \rho u_{t-1} \quad (\text{c})$$

4. Se resta **(a)-(c)**:

$$Y_t - \rho Y_{t-1} = \beta_1 - \rho \beta_1 + \beta_2 X_t - \rho \beta_2 X_{t-1} + u_t - \rho u_{t-1}$$

$$Y_t - \rho Y_{t-1} = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + (u_t - \rho u_{t-1})$$

$$Y_t - \rho Y_{t-1} = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \quad (\text{d})$$

Donde  $\varepsilon_t = (u_t - \rho u_{t-1})$

6. Estimo **(d)** por MCO.

**Cuando  $\hat{\rho}$  desconocida:**

Se utiliza en algoritmo de Cochrane Orcutt: Considérese el siguiente modelo:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (\text{e})$$

Y supóngase que  $u_t$ , es generado por el esquema AR(1):

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Cochrane Orcutt recomienda realizar los siguientes pasos:

1. Estimar **(e)** por MCO y se obtener  $\hat{u}_t$ .

2. Utilizando los residuos estimados  $\hat{u}_t$ , realizo las siguiente regresión:

$$\hat{u}_t = \hat{\rho} \hat{u}_{t-1} + v_t \quad (\text{f})$$

3. Utilizando  $\hat{\rho}$  obtenido en la regresión anterior, efectúese la ecuación en diferencia planteada en **(d)** por MCO.

4. Obtengo los  $\hat{\beta}_i$  y los sustituyo en **(a)**.

5. Se estima nuevamente:

$$\hat{u}_t = \hat{\rho} \hat{u}_{t-1} + w_t; \text{ donde } \hat{\rho} \text{ es la estimación de } \rho \text{ de (f).}$$

6. Se continúan haciendo estimaciones, y se suspenden las iteraciones cuando las estimaciones consecutivas de  $\rho$  difieren en una cantidad muy pequeña, es decir, en menos de 0.01 o 0.05.

### Ejemplo en Stata:

Para ejecutar el algoritmo de Cochrane Orcutt en Stat por comando, se escribe:

***prais inf unem, corc***

```
Iteration 0: rho = 0.0000
Iteration 1: rho = 0.5727
Iteration 2: rho = 0.7160
Iteration 3: rho = 0.7611
Iteration 4: rho = 0.7715
Iteration 5: rho = 0.7735
Iteration 6: rho = 0.7740
Iteration 7: rho = 0.7740
Iteration 8: rho = 0.7740
Iteration 9: rho = 0.7741
Iteration 10: rho = 0.7741
```

Cochrane-Orcutt AR(1) regression -- iterated estimates

Source	SS	df	MS	Number of obs =	48
-----+-----				F( 1, 46) =	4.33
Model	22.4790685	1	22.4790685	Prob > F =	0.0430
Residual	238.604008	46	5.18704365	R-squared =	0.0861
-----+-----				Adj R-squared =	0.0662
Total	261.083076	47	5.55495907	Root MSE =	2.2775
-----					
inf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

```

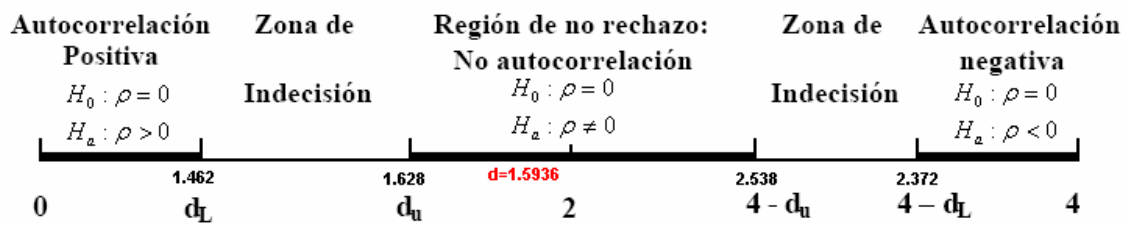
unem | -.6653356   .3196035   -2.08   0.043   -1.308664   -.0220071
_cons |  7.583458    2.38053    3.19   0.003    2.7917    12.37522
-----+-----
rho |   .7740512
-----+-----

Durbin-Watson statistic (original)    0.802700
Durbin-Watson statistic (transformed) 1.593634

```

En la salida anterior, se puede observar el numero de iteraciones que realizó el algoritmo (en este caso fueron 10), la regresión transformada, y el DW del modelo original y el DW del modelo corregido. Se puede concluir, con el nuevo DW=1.59, que ya no existe autocorrelación, pues dicho valor se encuentra muy cerca de 2.

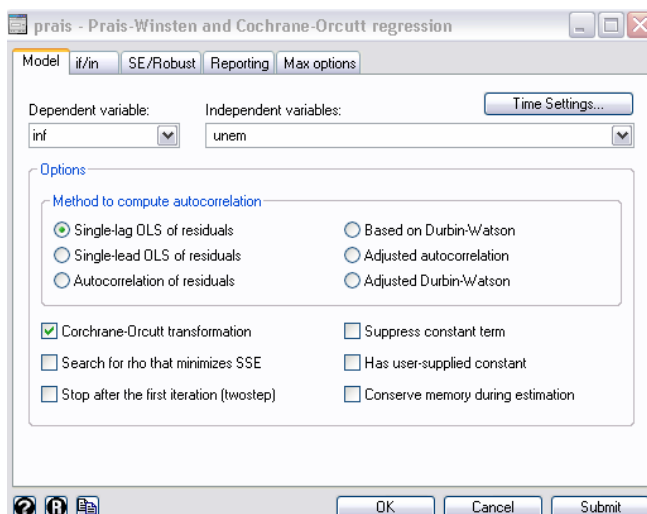
Gráficamente se tiene:



Si se quiere ejecutar el algoritmo por las ventanas en Stata, la ruta a seguir es:

### Statistics/time-series/tests/prais-winsten regression

Automáticamente se despliega el siguiente recuadro, en donde se selecciona la variable dependiente y las independientes, seleccionamos **Corchrane-Orcutt transformation**, y le damos **OK**.

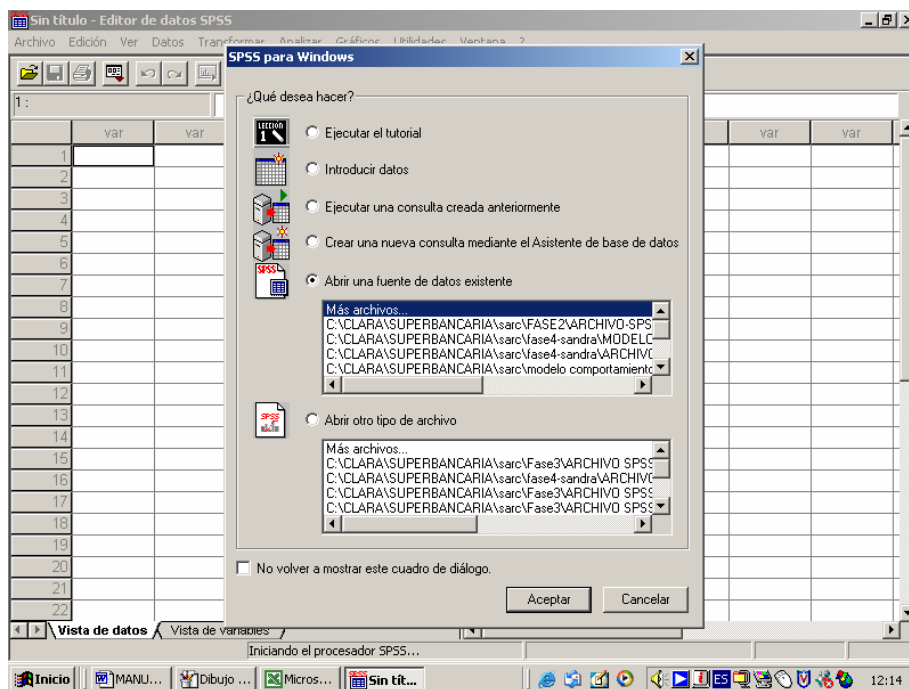




## APLICACIÓN DE MODELO LOGISTICOS EN SPSS 11.0

### a. Entrar al paquete estadístico SPSS

Automáticamente el sistema muestra una pantalla de opciones de entrada. Se debe seleccionar “Abrir una fuente de datos existente” y “aceptar”. El sistema abre la pantalla de búsqueda de archivos. En el menú opciones se debe seleccionar “Todos los archivos”. Se elige el archivo de Excel (Base de datos de los clientes) y se pulsa aceptar. (El archivo debe estar cerrado en Excel)



Automáticamente, el sistema importa el archivo a la plantilla de SPSS, como se muestra a continuación:

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Sin título - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1 : pacid 0,200824459586711

	pacid	rotcart	rotcp	parcart	cefect	racttot	mgneto	mgoper	roa	levtot
1	,2008245	39,32740	21,73045	,8138297	17,59694	3,966384	,0467275	,1109097	,3659717	,9746026
2	,3662763	86,91815	21,84156	,5822640	65,07659	1,639094	,0509922	,0517172	,3717169	3,447387
3	24,57783	50,24988	18,19095	,1474204	32,05893	,9141951	,0370163	,0569472	,1335753	2,947250
4	6,959167	18,53925	47,85788	,0788829	-29,3186	,9574105	,0333581	,0927101	,0348178	,0901914
5	,5531825	37,92979	2,522164	,5850129	35,40763	4,830206	,0078301	,0277292	,1184314	2,131377
6	2,091693	,0000000	,0000000	,0000000	,0000000	,9704997	,0084450	,0219464	,0107666	,3136551
7	1,053311	101,3320	43,31412	,5397440	58,01789	1,813841	,0419620	,1084599	,1297365	,7045405
8	,8553404	72,01667	,0161858	,3099434	72,00048	1,411112	,0026033	,0000619	,0625188	16,01855
9	6,528533	1065,600	6522,486	,1848029	-5456,89	,0510431	,6800000	1,000000	,1169474	2,369340
10	1,606258	33,25923	4,188725	,4777279	29,07050	1,841593	,2180284	,2180284	,4541053	,1309646
11	1,496460	19,89455	17,54673	,4782821	2,347816	5,822069	,1757061	,0030057	1,424214	,3922298
12	,4051902	49,59275	53,42197	,4875071	-3,82922	2,345248	,0296564	,0157371	,4778465	5,870388
13	,7470413	105,8593	116,1888	,5480540	-10,3295	1,741729	-,020440	-,008779	-,311878	7,760449
14	,6116593	95,55199	33,67644	,5584652	61,87554	1,023747	,1192590	,3961119	,1881879	,5413736
15	4,473301	1,239212	8,475423	,0212922	-7,23621	4,980731	,0492077	,0966287	,4743087	,9352420
16	,1707692	171,6023	60,50138	,8638321	111,1009	,1645696	,1445738	,1063586	,0666640	1,801897
17	,0495091	59,70362	550,8785	,6500000	-491,175	,3829053	,0665947	,1020370	,0824297	2,232603
18	1,130885	,1164154	20,62355	,0005357	-20,5071	,8028754	,0478657	,0608751	,1963015	4,108002
19	1,418790	192,5857	38,16286	,7370135	154,4228	,6887660	,3334319	,3900424	,2966682	,2917910
20	3,162439	54,51143	9,587801	,3420851	44,92363	2,116180	,0943326	,0116098	,4205861	1,106883
21	,1349490	192,4880	122,7839	,8727214	69,70410	,3127423	,4420046	,4266901	,1687249	,2205786
22	3,081476	,0000000	1,392600	,0000000	-1,39260	1,049428	-,310712	-,236892	-,383439	,1759403

Vista de datos Vista de variables

SPSS El procesador está preparado

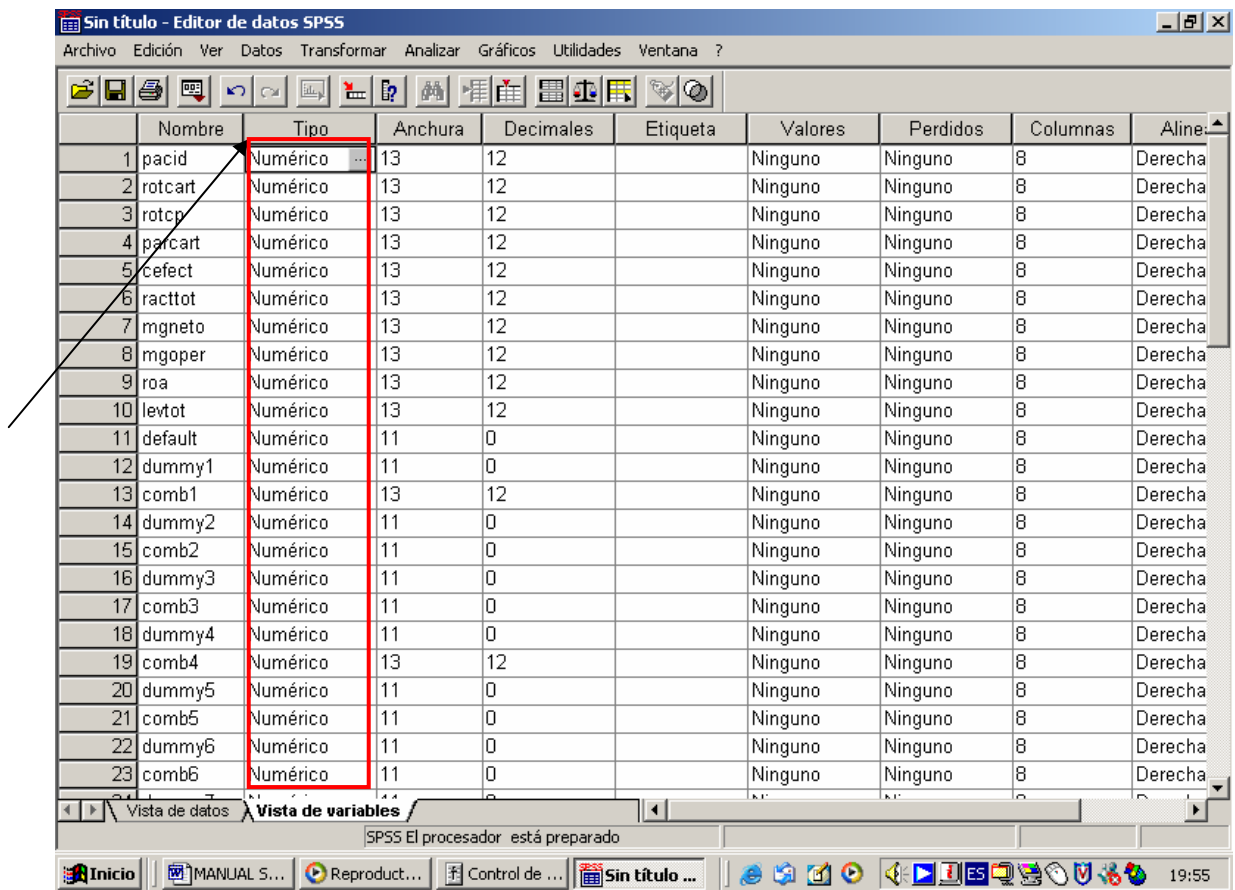
Inicio MANU... Dibujo... Micros... Sin tit...

12:08

### b. Verificación del cargue de la información del archivo plano a SPSS.

El usuario debe verificar que la información cargada esta completa; para esto se debe entrar a la hoja de SPSS “vista de variables” y revisar que en la columna “tipo”, todos los datos aparezcan como “numérico”.

Si algún dato aparece con otro nombre, es porque hay errores en el archivo de Excel, por tanto hay que corregirlos y cargar el archivo nuevamente.



## c. Generación del STEWISE o proceso Paso a Paso

El siguiente paso al cargue de información, es entrar a validar cuales variables, de todas las seleccionadas en la base de datos, son significativas. Para esto se va a utilizar la metodología STEPWISE, bajo la cual el sistema toma variable por variable y evalúa si es significativa, de no serlo la saca del modelo.

Para generar el Stepwise en SPSS el usuario debe entrar por el menú de la parte superior "Analizar" a la opción "Regresión logística binaria", la cual corresponde a los modelos tipo "LOGIT"

# MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

## Generación STEPWISE

**Sin título - Editor de datos SPSS**

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1 : pacid 0,2008244

	pacid	rotcart	ro	racttot	mgneto	mgoper	roa	levtot
1	,2008245	39,32740	21					
2	,3662763	86,91815	21					
3	24,57783	50,24988	18					
4	6,959167	18,53925	47					
5	,5531825	37,92979	2,5					
6	2,091693	,0000000	,00					
7	1,053311	101,3320	43					
8	,8553404	72,01667	,0161858	,3099434	72,00			
9	6,528533	1065,600	6522,486	,1848029	-545			
10	1,606258	33,25923	4,188725	,4777279	29,07658	,1,041353	,2166284	,2166284
11	1,496460	19,89455	17,54673	,4782821	2,347816	5,822069	,1757061	,0030057
12	,4051902	49,59275	53,42197	,4875071	-3,82922	2,345248	,0296564	,0157371
13	,7470413	105,8593	116,1888	,5480540	-10,3295	1,741729	-,020440	-,008779
14	,6116593	95,55199	33,67644	,5584652	61,87554	1,023747	,1192590	,3961119
15	4,473301	1,239212	8,475423	,0212922	-7,23621	4,980731	,0492077	,0966287
16	,1707692	171,6023	60,50138	,8638321	111,1009	,1645696	,1445738	,1063586
17	,0495091	59,70362	550,8785	,6500000	-491,175	,3829053	,0665947	,1020370
18	1,130885	,1164154	20,62355	,0005357	-20,5071	,8028754	,0478657	,0608751
19	1,418790	192,5857	38,16286	,7370135	154,4228	,6887660	,3334319	,3900424
20	3,162439	54,51143	9,587801	,3420851	44,92363	2,116180	,0943326	,0116098
21	,1349490	192,4880	122,7839	,8727214	69,70410	,3127423	,4420046	,4266901
22	3,081476	,0000000	1,392600	,0000000	-1,39260	1,049428	-,310712	-,236892

Regresión logística

SPSS El procesador está preparado

Inicio MANU... Reprod... Control... Sin tit... Dibujo ... 19:59

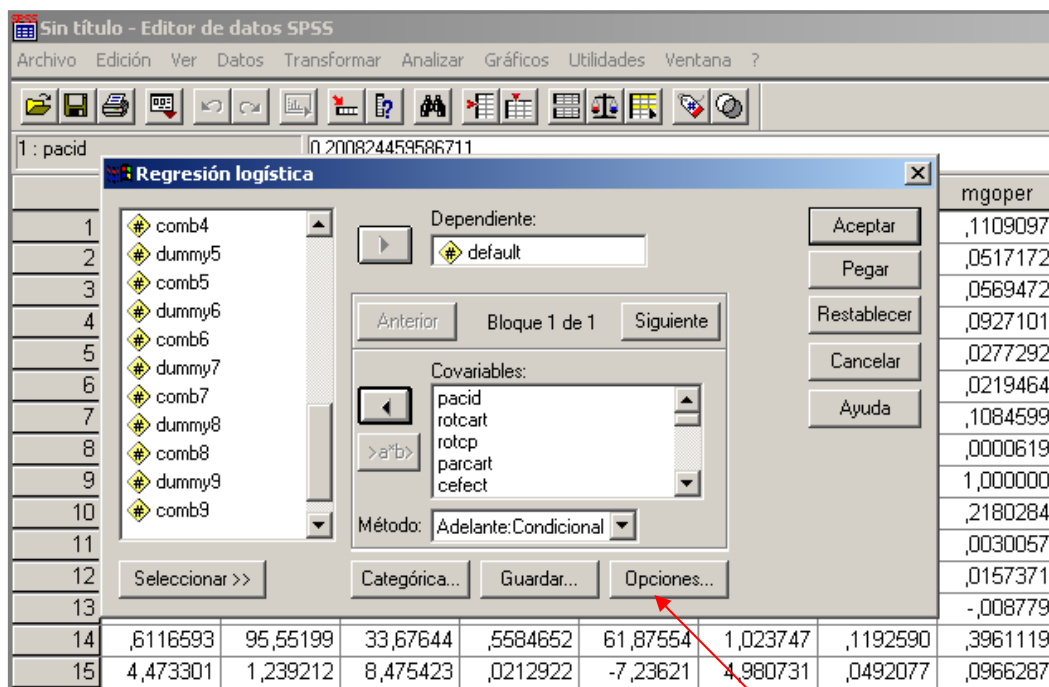
## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Automáticamente el sistema muestra una pantalla donde aparecen al lado derecho el listado de las variables del modelo definidas en el archivo de excel que se cargo en el sistema. Al lado izquierdo aparecen unos campos en blanco, en donde el usuario debe definirle al sistema cuales variables son independientes (covariables) y cual es la variable dependiente.

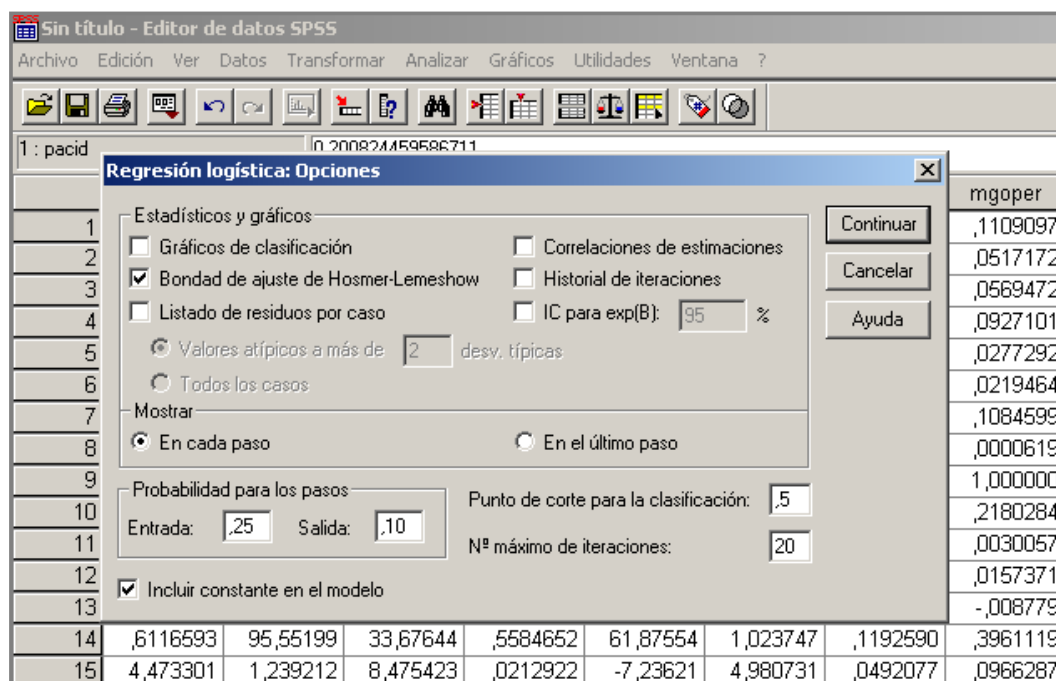
Las variables independientes son: Los indicadores financieros, las variables dummies y las variables combinadas. La variable dependiente es el indicador de Default. Se seleccionan las variables independientes y con la tecla con le signo flecha a la derecha las paso al campo de las variables covariables y luego se selecciona la variable Default y se pasa al otro lado al ítem dependiente.

El método que se debe utilizar para la selección de variables es “Adelante condicional”. Stepwise

A continuación se muestra la pantalla de registro de información, la cual incluye las opciones anteriormente descritas.



El indicador de desempeño de los modelos que se va a utilizar es la “Prueba de Hosmer-Lemeshow.” El usuario debe entrar por el menú “opciones” y seleccionar la prueba de Hosmer Lemeshow. Adicionalmente, en la sección “probabilidad para pasos” se debe digitar en el campo entrada: 0.25 y en el campo salida 0.10, y pulsar continuar. Posteriormente, se debe pulsar aceptar.



Automáticamente el SPSS genera el reporte de Stepwise, el cual incluye los mejores modelos calculados por el sistema, los cuales están basados en los parámetros predefinidos por el usuario.

El parámetro utilizado para la selección del modelo de medición de riesgo de crédito es el indicador de HOSMER LEMESHOW.

Se debe elegir el modelo (Paso) con mayor Hosmer Lemeshow. La siguiente tabla muestra la prueba de Hosmer Lemeshow para cada paso o modelo generado por el sistema.

## Prueba de Hosmer y Lemeshow

PASO	Chi-cuadrado	gl	Sig
2	0,1	1	0,87
3	5,45	6	0,456
4	10,67	7	0,67

De acuerdo con la tabla, el mejor modelo corresponde al paso 2, donde el Hosmer-Lemeshow es de 0,87.

## MANUAL DE APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

Ahora que se tiene el paso o modelo se debe ver que variables lo definieron Se debe buscar la siguiente tabla:

Variables en la ecuación							
		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1	D5ROACT	1,416	,375	14,222	1	,000	4,120
	Constante	-1,194	,361	10,939	1	,001	,303
Paso 2	D5ROACT	1,551	,383	16,409	1	,000	4,717
	D11MGNET	,659	,209	9,952	1	,002	1,934
	Constante	-1,576	,387	16,586	1	,000	,207
Paso 3	C1PACID	,779	,323	5,818	1	,016	2,180
	D5ROACT	1,540	,386	15,945	1	,000	4,665
	D11MGNET	,630	,211	8,928	1	,003	1,877
	Constante	-1,750	,398	19,346	1	,000	,174
Paso 4	VTAS_ACT	,003	,004	,471	1	,492	1,003
	C1PACID	,802	,324	6,122	1	,013	2,229
	D5ROACT	1,652	,401	17,001	1	,000	5,217
	D11MGNET	,616	,211	8,513	1	,004	1,852
	Constante	-1,866	,413	20,373	1	,000	,155

Para elegir el mejor modelo, es necesario SELECCIONAR LOS DOS MEJORES MODELOS ARROJADOS POR EL STEPWISE Y correr cada modelo en SPSS bajo la metodología de regresión logística binaria por el **método de introducir**.

El paso 2 o modelo 2 esta dado por las variables: D5ROACT y D11MGNET.

### Selección del modelo y calculo de las Probabilidades de Incumplimiento

Ahora que se conoce el modelo, se debe correr únicamente este, bajo la metodología Logit por el método Introducir como se muestra a continuación.

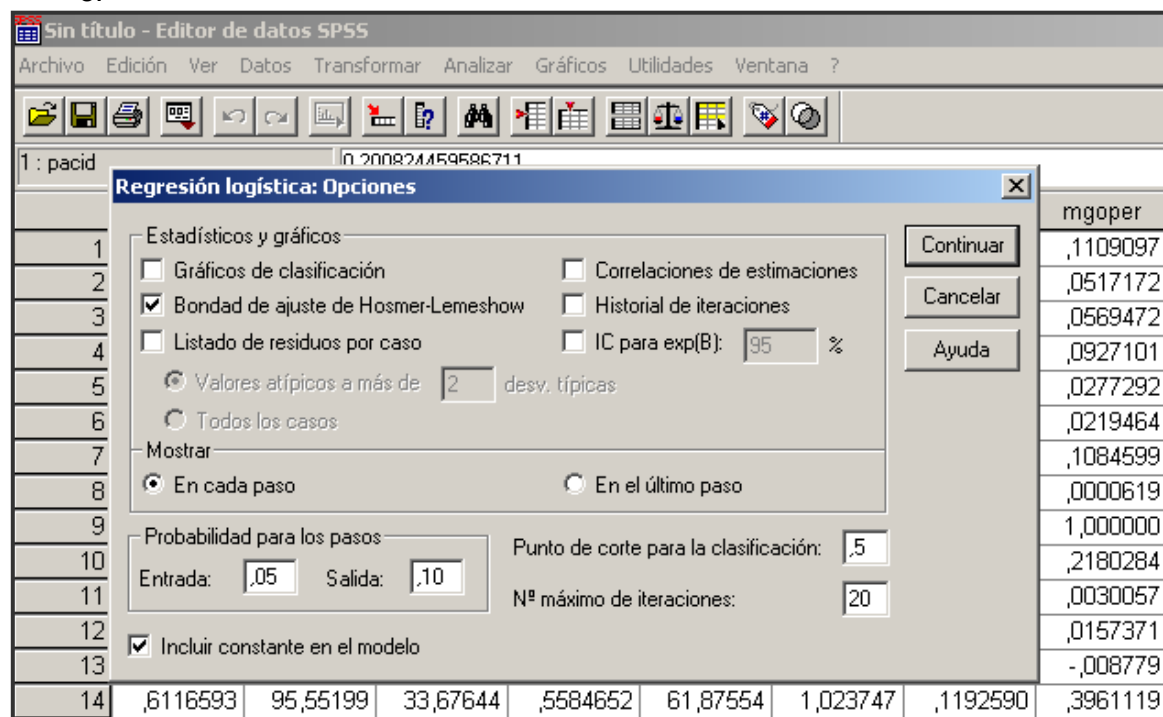
Se deben seguir los siguientes pasos:

- Entrar por el menú “analizar” a la opción “Regresión logística”, que corresponde a los modelos tipo “LOGIT”
- Seleccionar el método “**introducir**”
- En el menú “Guardar” se debe seleccionar “Probabilidades” y luego pulsar “Continuar”



- d. En el menú “opciones” se debe seleccionar “Prueba de Hosmer y Lemeshow”. Adicionalmente, en la sección “Probabilidades en datos” en entrada se debe digitar 0.05 y en salida 1.00. Las demás opciones se deben dejar como aparecen inicialmente en el sistema.

e.





- f. Pulsar “continuar” y luego “aceptar”

De esta forma el sistema procesa la información, genera el reporte del modelo. Con los resultados se debe evaluar la significancia individual de las variables, es decir, si las variables son significativas.

### **Calculo de la Probabilidad de incumplimiento**

El calculo de la probabilidad de incumplimiento de un modelo de tipo “LOGIT” esta dada por la siguiente ecuación:

$$F(Z) = \frac{e^Z}{1 + e^Z}$$

Donde: F(z) es la variable dependiente en función de la Z. F(z) es la Probabilidad de incumplimiento.

Z: Es la ecuación que representa la combinación de variables independientes que permiten explicar la probabilidad de incumplimiento.

Z esta dado por  $Z = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n$ .

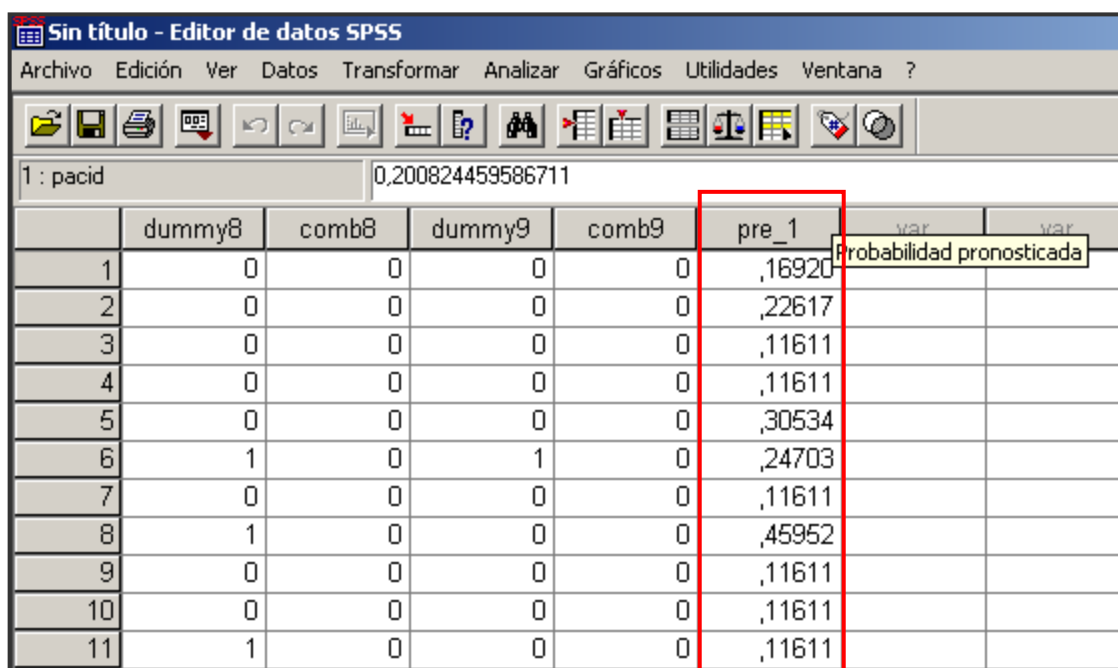
Los betas  $B_0, B_1, \dots, B_n$ , representan la ponderación de cada variable dentro del modelo.

SPSS calcula las probabilidades de incumplimiento (PI) de los clientes que pertenecen a la muestra del modelo. Para los demás clientes , que no pertenecen a la base de datos, es necesario aplicar la formula de F(Z) o probabilidad de incumplimiento de forma manual o a través de un archivo de excel.

## Calculo de la probabilidad de incumplimiento en SPSS.

En el momento en que se corre la regresión logística bajo la metodología de introducir, el sistema calcula las probabilidades de incumplimiento y las reporta en la columna final de la plantilla de captura de información. En la ventana Editor de datos SPSS aparecen las variables, la ultima columna “pre\_1” hace referencia a las probabilidades de incumplimiento calculadas con base en el modelo seleccionado.

A continuación se muestra un ejemplo:



Sin título - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1 : pacid 0,200824459586711

	dummy8	comb8	dummy9	comb9	pre_1	var	var
1	0	0	0	0	,16920		
2	0	0	0	0	,22617		
3	0	0	0	0	,11611		
4	0	0	0	0	,11611		
5	0	0	0	0	,30534		
6	1	0	1	0	,24703		
7	0	0	0	0	,11611		
8	1	0	0	0	,45952		
9	0	0	0	0	,11611		
10	0	0	0	0	,11611		
11	1	0	0	0	,11611		

Hoja “ Datos” del SPSS. Columna final.

## BIBLIOGRAFIA

- **MADDALA, G. S. (1996)** Introducción a la econometría. Ed. Prentice-Hall Hispano Americana S.A. Mexico.
- **NOVALES, A. (1993)** Econometría, 2ª Edición. Ed. McGraw-Hill. Madrid.
- **JOHNSTON, J. (1987)** Métodos de econometría. Barcelona: Vicens Vices.
- **PINDYCK, R. S. y D. L. RUBINFELD (2001)** Econometría. Modelos y pronósticos. México: McGraw—Hill.
- **JUDGE, G. G., R. C. HILL, W. E. GRIFFITHS, H. LÜTKEPOHL y T. C. LEE (1988)** Introduction to the Theory and Practice of Econometrics. New York: John Wiley & Sons.
- **GREENE, W. H. (1999)** Análisis econométrico. Madrid: Prentice Hall.
- **PENA, B., J. ESTAVILLO, M. E. GALINDO, M. J. LECETA y M. M. ZAMORA (1999)** Cien ejercicios de econometría. Madrid: Pirámide.