Unidad 2. El modelo básico de regresión lineal (MBRL)

Prof. Dra. Eva Romero Ramos

Universidad Europea
LAUREATE INTERNATIONAL UNIVERSITIES

Contenidos

- 1. El modelo de regresión lineal simple.
- 2. El modelo de regresión lineal múltiple.
- 3. Estimación del modelo de regresión lineal múltiple con Gretl.
- 4. Normalidad, inferencia y bondad de ajuste.

1. El modelo de regresión lineal simple.

El modelo de regresión lineal Simple

El Modelo de Regresión Lineal Simple busca encontrar la recta de regresión que relaciona 2 variables X e Y de la siguiente forma:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

donde:

Y es la variable dependiente.

X es la variable independiente.

β_0 y β_1 son los parámetros del modelo que debemos estimar.

Ɛ es el termino error.

El modelo de regresión lineal Simple - Hipótesis

Para poder aplicar el Modelo de Regresión Lineal simple se deben cumplir las siguientes hipótesis:

Hipótesis 1 (H1): Las variables independiente y dependiente (X e Y) son cuantitativas y aleatorias y presentan por tanto una relación aleatoria.

Hipótesis 2 (H2): La variable independiente explica la dependiente, es decir, X explica a Y y no al revés.

Hipótesis 3 (H3): La variable independiente se relaciona linealmente con la variable dependiente, X se relaciona linealmente con Y.

El modelo de regresión lineal Simple - Hipótesis

Hipótesis 4 (H4): El modelo esta correctamente especificado y la relación entre las variables es de causalidad o causa-efecto.

Hipótesis 5 (H5): β_1 es constante, lo que implica que las variaciones de Y ante cambios de X presentan un valor estable para las distintas muestras.

Hipótesis 6 (H6): El tamaño muestral es suficientemente grande para afrontar la estimación de los parámetros del modelo.

El modelo de regresión lineal Simple - Hipótesis

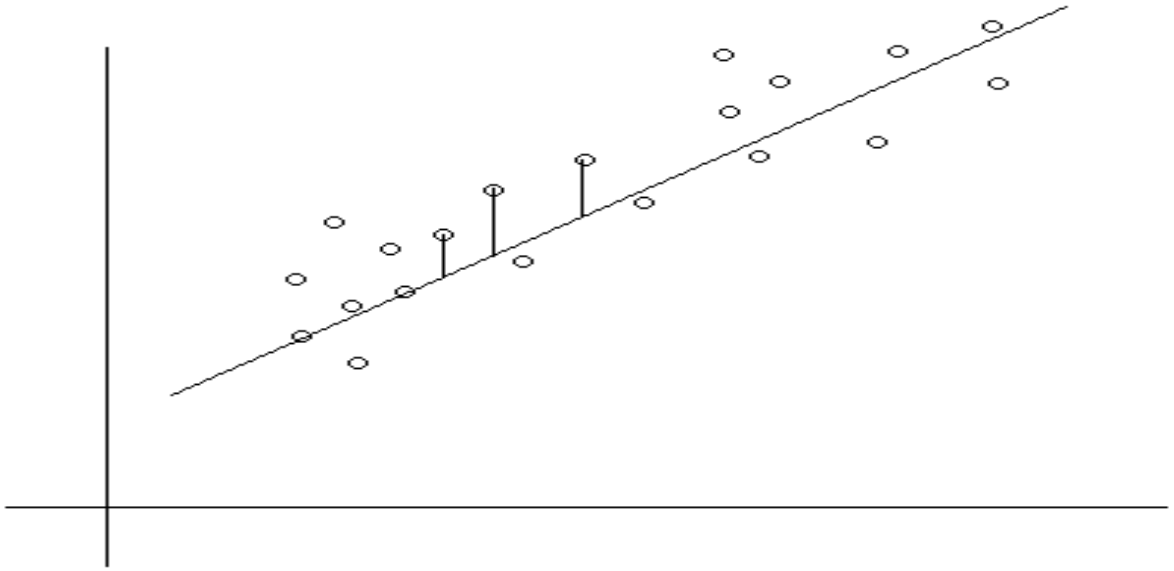
Hipótesis 7 (H7): El término error es un término completamente aleatorio que sigue una distribución normal, de esperanza 0 ($E[\varepsilon_i]=0$).

Hipótesis 8 (H8): La varianza del error es constante a lo largo de las observaciones del modelo ($Var[\varepsilon_i]=\sigma^2$).

Hipótesis 9 (H9): El término error no esta correlacionado entre los elementos del modelo y es independiente también de la variable explicativa.

El modelo de regresión lineal Simple

Para estimar los parámetros del modelo buscaremos los valores de β_0 y β_1 que construyan una recta, de modo que la distancia de los puntos a ella sea la mínima posible.



El modelo de regresión lineal Simple

- Los estimadores por el método de mínimos cuadrados ordinarios para el modelo de regresión lineal simple son:

$$\hat{\beta}_1 = \frac{COV(X,Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

El modelo de regresión lineal Simple – Ejemplo 1

Se plantea el estudio del salario de un individuo, encontrando el modelo que lo relaciona con su educación, medida a través de los años dedicados a su formación. Los datos son los siguientes:

Salario Bruto anual	Años de formación
22.000	20
19.000	15
25.000	20
30.000	25
35.000	27
24.000	21
26.000	23
41.000	26
45.000	25
18.000	13
19.000	14
21.000	15

Calcular los estimadores por el método de mínimos cuadrados.

Coeficiente de correlación lineal

Llamamos correlación al grado de dependencia mutua entre las variables. La correlación trata de medir la intensidad con que dos variables pueden estar relacionadas.

Coeficiente de correlación lineal

$$r = \frac{COV(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}$$

El valor del coeficiente de correlación lineal siempre estará entre -1 y 1.

Coeficiente de correlación lineal

- Si **r=1**: **correlación lineal perfecta positiva** y los valores teóricos coinciden con los observados, ya que todos los puntos de la nube están en la recta. Es decir, existe dependencia funcional que viene reflejada por una recta creciente.
- Si **r=-1**, la **correlación lineal es perfecta negativa** y, aquí también, los valores teóricos coinciden con los observados, pero la recta es decreciente. De nuevo es un caso de dependencia funcional.
- Si **r=0**, la **correlación lineal es nula**. Es decir, no hay asociación lineal y por mucho que varíe X, la variable Y no se verá afectada (de forma lineal).

Coeficiente de correlación lineal

- Si **-1<r<0**, la **correlación lineal será negativa** y la recta será decreciente puesto que el signo de su pendiente coincide con el de la covarianza que es la que da el signo a r, luego al ser r negativo también lo será la pendiente.

  Si r es cercano a 0 diremos que la relación es débil, y cuanto más se acerque a -1 consideráremos que la relación es más fuerte.

- Si **0<r<1**, la **correlación lineal es positiva**. Esto indica que la recta es creciente y cuando los valores de una variable crecen lo de la otra también crecerán.

  Consideraremos también que cuanto más se acerque a 0 más débil es la relación entre las variables y si el valor es próximo a 1 la relación podrá considerarse fuerte.

Coeficiente de correlación lineal

- Cuando dos variables son estadísticamente independientes su covarianza es cero. Por consiguiente, si las variables son independientes, están también incorrelacionadas linealmente, al ser r=0.

- Sin embargo: Dos variables pueden estar incorrelacionadas linealmente y ser dependientes, puesto que cuando r=0 lo único que podemos decir es que la dependencia estadística lineal es nula, pero esas variables pueden depender según otro tipo de función (parabólica, exponencial, etc.)

- Además se puede demostrar la invarianza de r ante transformaciones lineales.

Coeficiente de correlación lineal – Ejemplo 1

Calcule el coeficiente de correlación lineal para el modelo de regresión lineal simple planteado en el ejemplo 1.

Coeficiente de determinación

Definición.- El coeficiente de determinación se interpreta como el porcentaje de variación de la variable dependiente explicado por el modelo.

En modelos de regresión lineal simple, se calcula simplemente como el cuadrado de coeficiente de correlación lineal:

$$R^2 = \frac{Cov(X,Y)^2}{Var(X) \cdot Var(Y)}$$

Coeficiente de determinación – Ejemplo 1

Calcule el coeficiente de determinación para el modelo de regresión lineal simple planteado en el ejemplo 1.

Los residuos

Definición.- El residuo para la observación i es la diferencia que hay entre el verdadero valor de y_i y su valor estimado $\hat{y}_i$.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i$$

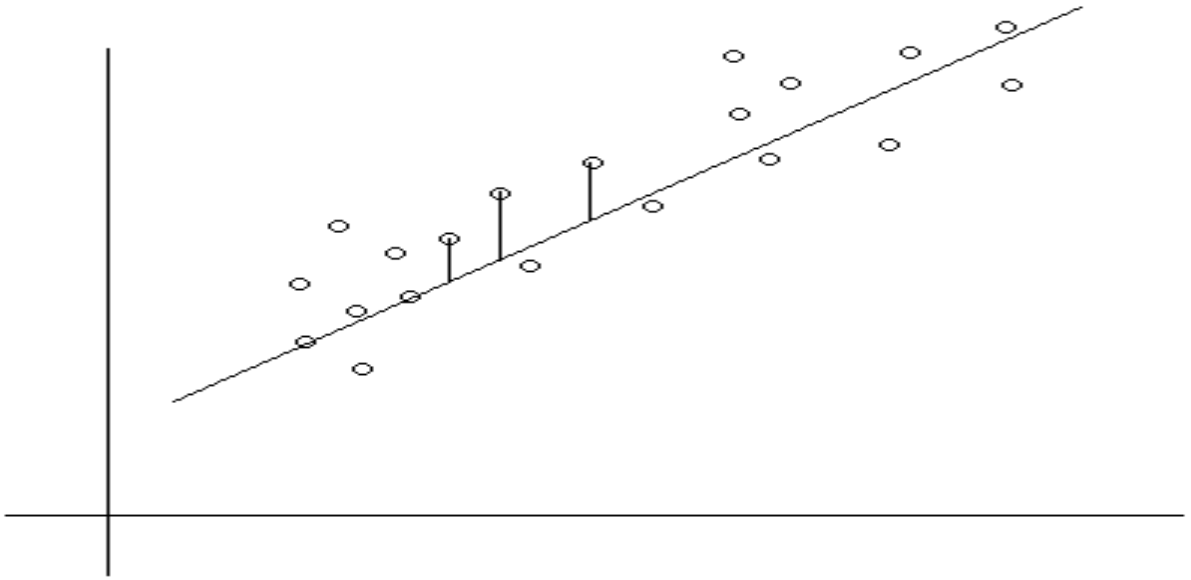
De este modo, cuando estimamos un modelo tendremos tantos residuos como observaciones incluya la muestra (n).

Los residuos no deben confundirse con el error.

Podremos escribir y_i como la suma entre su valor estimado $\hat{y}_i$ y el residuo $\hat{\varepsilon}_i$:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i.$$

Los residuos



- Sabemos que el método de mínimos cuadrados calcula los estimadores de β_0 y β_1 de forma que la suma de los cuadrados de los residuos sea mínima.
- La expresión “mínimos cuadrados ordinarios” viene de este hecho.

Los residuos – Ejemplo 1

Calcule los residuos para el modelo de regresión lineal simple planteado en el ejemplo 1.

Suma total de cuadrados (STC)

La variabilidad total de la variable y, se medirá por su suma total de cuadrados definida por:

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Suma explicada de cuadrados (SEC)

La variabilidad de la variable dependiente que se consigue explicar con el modelo se define por:

$$SEC = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Suma de cuadrados de los residuos (SCR)

La variabilidad de la variable dependiente que no se consigue explicar con el modelo se define por:

$$SCR = \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2$$

STC, SEC y SCR

- La variabilidad total de la variable dependiente se puede descomponer entonces en la variabilidad que explica el modelo y la variabilidad que no explica el modelo.
- Se cumple por tanto que:

$$STC = SEC + SCR$$

- El *coeficiente de determinación* se puede calcular en base a estas sumas, como el cociente entre la suma explicada y la suma total, es decir:

$$R^2 = \frac{SEC}{STC} = 1 - \frac{SCR}{STC}$$

STC, SEC y SCR – Ejemplo 1

Calcule la suma total de cuadrados (STC), la suma explicada de los cuadrados (SEC) y la suma de los cuadrados de los residuos (SCR) para el modelo de regresión lineal simple planteado en el ejemplo 1.

Compruebe que se cumple que STC = SEC + SCR y calcule el coeficiente de determinación a partir de estas sumas.

Propiedades de los estimadores de MCO

Los estimadores de mínimos cuadrados ordinarios obtenidos para el modelo de regresión lineal simple son **insesgados** y **eficientes**.

2. El modelo de regresión lineal múltiple.

El modelo de regresión lineal Múltiple

El modelo de regresión múltiple tiene como objetivo explicar el comportamiento de una variable dependiente utilizando la información proporcionada por los valores de un conjunto de variables explicativas.

La ecuación del modelo de regresión múltiple es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \varepsilon_i$$

- Los *coeficientes (parámetros)* $\beta_1, \beta_2, \ldots, \beta_K$ denotan la magnitud del efecto que las variables explicativas $X_1, X_2, \ldots, X_K$ tienen sobre la variable dependiente o explicada, considerando que el resto de variables permanece constante.

- El coeficiente β_0 se denomina *término constante* o *independiente* del modelo.

- El término ε se denomina término de error del modelo.

El modelo de regresión lineal Múltiple - Hipótesis

Para poder aplicar el Modelo de Regresión Lineal simple se deben cumplir las siguientes hipótesis:

Hipótesis 1 (H1): La variable dependiente y las variables independientes (X_1, X_2,… X_K) presentan una relación aleatoria y la variable dependiente es cuantitativa.

Hipótesis 2 (H2): Las variables independientes explican a la dependiente, es decir, (X_1, X_2,… X_K) explican a Y y no al revés.

Hipótesis 3 (H3): La variable dependiente se relaciona linealmente con las variables independientes, es decir Y se relaciona linealmente con (X_1, X_2,… X_K).

El modelo de regresión lineal Múltiple - Hipótesis

Hipótesis 4 (H4): El modelo esta correctamente especificado, es decir:

- No se omiten variables explicativas relevantes para explicar la variabilidad de Y.
- No se incluyen variables explicativas superfluas para explicar la variabilidad de Y.
- La muestra de datos se adecua a los requerimientos del modelo.

Hipótesis 5 (H5): Los parámetros β_i son constantes, lo que implica que las variaciones de Y ante cambios cada una de las X_i presentan un valor estable para las distintas muestras.

Hipótesis 6 (H6): El tamaño muestral es suficientemente grande para afrontar la estimación de los parámetros del modelo.

$$n >> k+1$$

El modelo de regresión lineal Múltiple - Hipótesis

Hipótesis 7 (H7): Las variables explicativas son deterministas, es decir, no son aleatorias y no existe incertidumbre sobre ellas. Esta hipótesis no siempre se cumple.

Un ejemplo de su incumplimiento son series temporales en las que en ocasiones es necesario incluir como variable explicativa la variable dependiente retardada.

Hipótesis 8 (H8):

Las variables explicativas son linealmente independientes entre sí. Esto garantiza que no haya variables redundantes.

El modelo de regresión lineal Múltiple - Hipótesis

Hipótesis 9 (H9):

El término error es un término completamente aleatorio que sigue una distribución normal, de esperanza 0 ($E[\varepsilon_i]=0$).

Hipótesis 10 (H10): La varianza del error es constante a lo largo de las observaciones del modelo ($Var[\varepsilon_i]=\sigma^2$).

Hipótesis 11 (H11): El término error no esta correlacionado entre los elementos del modelo y es independiente también de las variables explicativas.

Estimación del Modelo Lineal por mínimos cuadrados Ordinarios MCO

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \mathrm{X} = \begin{pmatrix} x_{11} \dots x_{j1} \dots x_{K1} \\ \vdots \ \ddots \ \vdots \ \ddots \ \vdots \\ x_{1i} \cdots x_{ji} \cdots x_{Ki} \\ \vdots \ \ddots \ \vdots \ \ddots \ \vdots \\ x_{1n} \cdots x_{jn} \cdots x_{Kn} \end{pmatrix};$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_K \end{pmatrix}; \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Estimación del Modelo Lineal por mínimos cuadrados Ordinarios MCO

Los estimadores de mínimos cuadrados ordinarios (MCO) para los parámetros del modelo (β_i) se calculan mediante:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Teniendo en cuenta que este vector de parámetros no incluye el término independiente, β_0. Para calcularlo debemos usar:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 - \ldots - \hat{\beta}_K \bar{X}_K$$

Teorema de Gauss-Markov

Supuestos:

1. Puede establecerse la relación lineal entre la variable dependiente y las independientes que plantea el modelo.
2. Las observaciones han sido obtenidas mediante un muestreo aleatorio.
3. La esperanza del termino error es nula.
4. Ninguna de las variables independientes es constante y no existen relaciones lineales exactas entre ellas.
5. Homoscedasticidad: La varianza del error es constante.

El teorema de Gauss-Markov dice que bajo los supuestos 1 a 5, los estimadores obtenidos por el método de mínimos cuadrados ordinarios son estimadores *lineales*, *insesgados* y *óptimos* de los parámetros del modelo de regresión lineal múltiple.

Ejemplo 2.- Prima de un Warrant

Se pretende estudiar la prima de un determinado Warrant en función de los tipos de interés y el tiempo.

La teoría nos dice que la prima se mueve en la misma dirección que los tipos de interés y sin embargo disminuye a medida que transcurre el tiempo y nos acercamos al vencimiento de la opción.

La matriz de varianzas-covarianzas de una muestra de tamaño 100 es la siguiente:

	Prima	Interés	Tiempo
Prima	526,5	0,095	-655,47
Interés	0,095	1,003	1,05
Tiempo	-655,47	1,05	833,25

Y las medias son:

Prima	Interés	Tiempo
36,28	3,93	49,5

Plantear y estimar bajo MCO un modelo que describa la prima del warrant en función de los tipos de interés y del tiempo.

Propiedades algebraicas de los estimadores de mínimos cuadrados

- Hacen que la línea de regresión muestral pase por el centro de gravedad de las variables que intervienen, es decir, por sus valores medios Y, X_1,…, X_k .

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \ldots + \hat{\beta}_K \bar{X}_K$$

- Hacen que los residuos tenga media 0, supuesto clave cuando componemos el modelo de regresión lineal en desviaciones a las medias.
- Conforman residuos de regresión no correlacionados con los regresores.
- Conforman residuos de regresión no correlacionados con la estimación de la variable que queremos explicar.
- El valor medio de las n estimaciones de Y coinciden con su valor medio observado.
- Son **insesgados** y **eficientes** (de varianza mínima).

Análisis de la Varianza

	Variabilidad	Grados de Libertad	Varianza
Variación Explicada por el modelo (VE)	$\hat{\beta}^t X^t Y$	K	$\frac{VE}{g.l.} = \frac{\hat{\beta}^t X^t Y}{k}$
Variación No Explicada por el modelo (VNE)	$Y^t Y - \hat{\beta}^t X^t Y$	n-k -1	$\frac{VNE}{g.l.} = \frac{Y^t Y - \hat{\beta}^t X^t Y}{\text{n-k -1}}$
Variación Total	$Y^t Y$	n-1	$\frac{VT}{g.l.} = \frac{Y^t Y}{n-1}$

Ejemplo 2.- Prima de un Warrant

Para el ejemplo anterior, calcule la variación Explicada, la variación no explicada y la variación total y compruebe que se cumple:

$$VT = VE + VNE$$

Coeficiente de determinación

Al igual que en el modelo de regresión lineal simple, el coeficiente de determinación se interpreta como el porcentaje de variación de la variable dependiente explicado por el modelo.

Se calcula como:

$$R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

Problema: El valor del coeficiente de determinación *siempre aumenta cuando incluimos nuevas variables en el modelo*, incluso cuando están son poco significativas o tienen poca correlación con la variable dependiente.

Coeficiente de determinación – Ejemplo 2

Calcular e interpretar el coeficiente de determinación para el ejemplo de la Prima de un Warrant.

Coeficiente de Determinación Corregido

El coeficiente de Determinación es un coeficiente que corrige el problema del coeficiente de determinación.

Se define como:

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-K-1}$$

Como vemos este coeficiente tiene en cuenta el número de variables incluidas en el modelo (K).

Este coeficiente permite comparar modelizaciones alternativas que, manteniendo las mismas observaciones, incluyen distinto número de variables.

Coeficiente de Determinación Corregido – Ejemplo 2

Calcular e interpretar el coeficiente de determinación corregido para el ejemplo 2.

Esperanza y Varianza de $\widehat{\beta}$

❑ Esperanza de $\hat{\beta}$

Por ser $\hat{\beta}$ un estimador insesgado de β, su esperanza coincide con el verdadero valor de β:

$$E[\hat{\beta}] = \beta$$

❑ Matriz de Varianzas-Covarianzas de $\hat{\beta}$

$$\text{Var}[\hat{\beta}] = \sigma_{\varepsilon}^{2}\,(X^{t}X)^{-1}$$

Se demuestra que $\hat{\beta}$ es de varianza mínima, es decir eficiente.

Esperanza y Varianza de $\widehat{\beta}$ - Ejemplo 2

Calcular la matriz de varianzas covarianzas de β.

3. Estimación del modelo de regresión lineal múltiple con Gretl.

Ejemplo 3.- Gasto en mantenimiento

Supongamos que una empresa está interesada en encontrar los factores que afectan al *gasto anual en reparaciones de maquinaria parar la producción (GR)*. Esta será por tanto nuestra variable dependiente (Y).

Inicialmente se plantea como variables explicativas la **antigüedad de la maquinaria medida en años *(ANT)***, el **gasto anual en revisiones *(REV)*** y las **horas de funcionamiento anuales *(FUN)***.

Se dispone para el estudio de una muestra de datos que incluye 50 maquinas en el archivo "***gasto_maquinaria.gdt***".

Hipótesis previas

- Se espera que la relación entre el gasto anual en reparaciones y la **antigüedad** sea positiva, es decir, que cuantos más años tiene una máquina, mayor será el gasto en reparaciones.
- Con respecto al **gasto anual en revisiones**, se espera encontrar una relación negativa con el gasto en reparaciones, de forma que cuanto mayor sea el gasto revisiones, menor sea el gasto en reparaciones.
- Finalmente se espera que las **horas de funcionamiento** tengan una relación positiva con el gasto en reparaciones, ya que cuanto mayor sea el tiempo de funcionamiento de una maquina, previsiblemente mayor será el número de reparaciones que precise.

Hipótesis previas

En resumen, las relaciones que esperamos encontrar son:

Antigüedad	+
Gasto en revisiones	-
Tiempo de funcionamiento	+

Análisis descriptivo de las variables del modelo

- Todo estudio estadístico debe comenzar con un *análisis descriptivo de las variables que intervienen*, para conocerlas mejor y centrar adecuadamente el análisis.

- Es importante explicar bien con que variables estamos trabajando y como se miden.

- Podemos obtener una tabla con los principales estadísticos descriptivos de las variables a través del menú:

Ver –> estadísticos principales

Análisis descriptivo de las variables del modelo

gretl: estadísticos principales

	Media	Mediana	Mínimo	Máximo
GR	3200.2	3177.8	1600.3	4765.2
ANT	12.340	13.000	1.0000	24.000
REV	393.90	390.19	334.49	471.27
FUN	3253.0	3263.0	2510.0	4267.0

	Desv. Típica.	C.V.	Asimetría	Exc. de curtosis
GR	715.43	0.22356	-0.048305	-0.61835
ANT	7.1074	0.57596	0.072597	-1.2308
REV	33.888	0.086034	0.31705	-0.39355
FUN	450.60	0.13852	0.21006	-0.81481

	Perc. 5%	Perc. 95%	Rango IQ	Observaciones ausentes
GR	2040.8	4324.1	1103.5	0
ANT	2.5500	24.000	12.500	0
REV	340.93	461.29	43.840	0
FUN	2510.0	4016.0	564.75	0

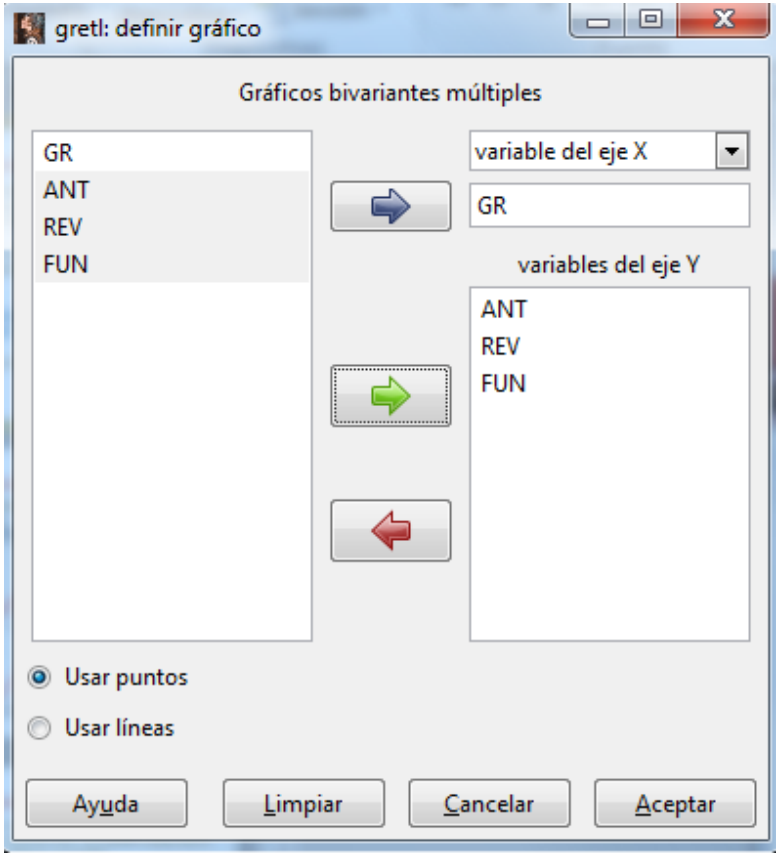
Análisis de relaciones con la variable dependiente

Una vez planteado el modelo y las hipótesis previas, podemos analizar las relaciones existentes entre las variables explicativas y la variable explicada, para confirmar si existe entre ellas una relación lineal.

Para obtener los gráficos de dispersión en Gretl usaremos la secuencia:

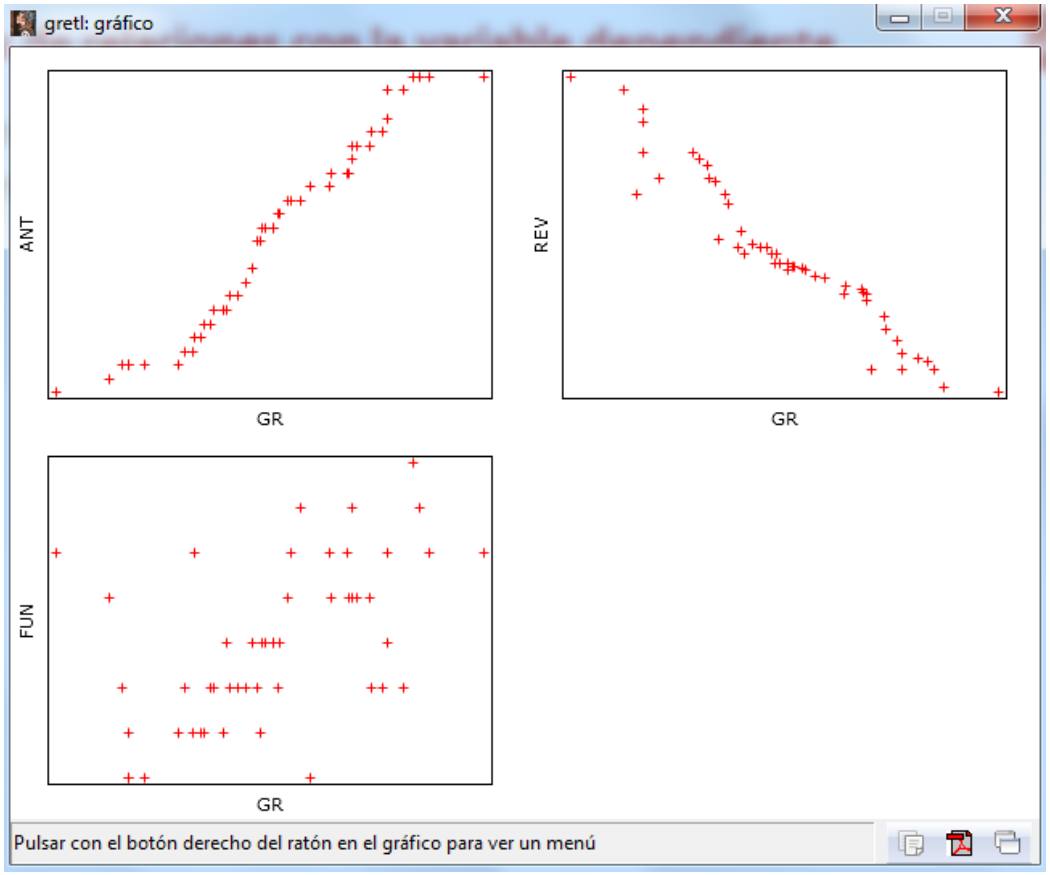
Ver -> Gráficos múltiples -> Gráficos X Y (Scatter)

Que nos llevará al siguiente cuadro de diálogo:



Análisis de relaciones con la variable dependiente

Las variables antigüedad y gasto en revisiones tienen una clara relación lineal con la variable dependiente. Como esperábamos, esta relación es positiva en el caso de la variable antigüedad y negativa en el caso del gasto en revisiones.



Con la variable horas de funcionamiento anuales, el gasto en reparaciones presenta una relación positiva pero no tan claramente lineal con el caso de las anteriores.

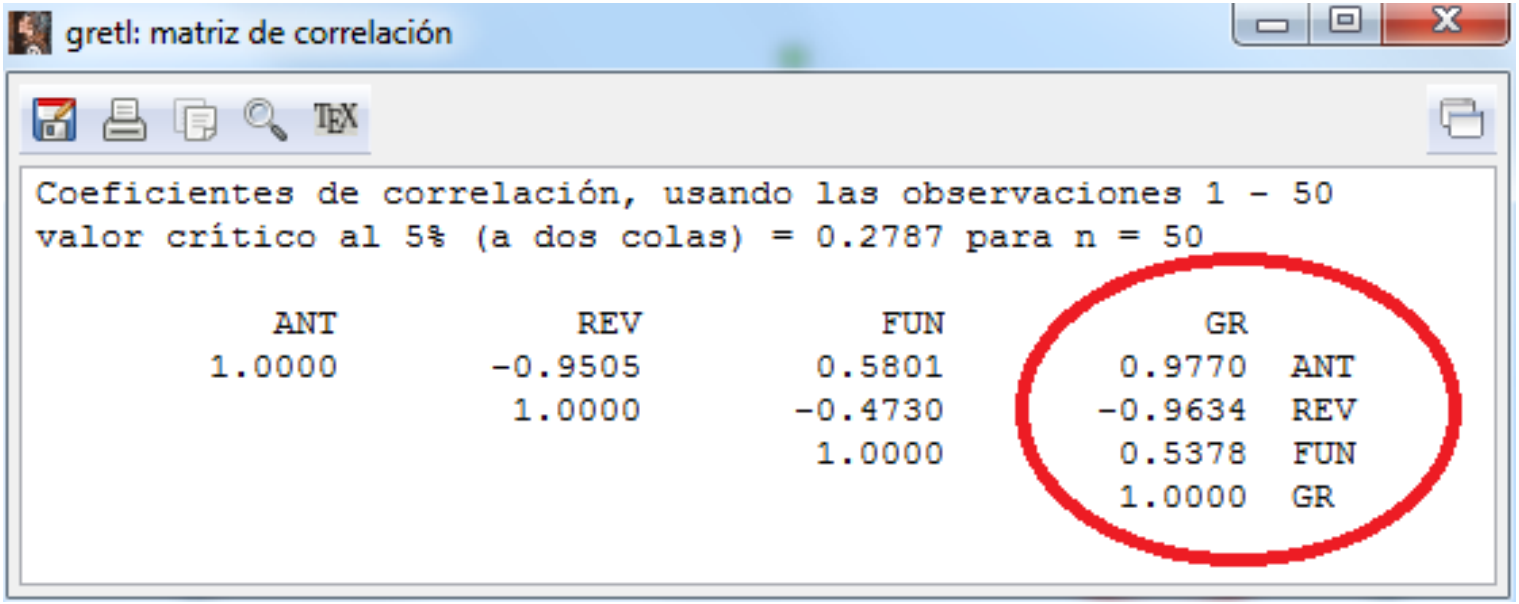
Análisis de relaciones con la variable dependiente

- En este caso los gráficos muestran relaciones lineales bastante claras, no obstante se deben corroborar también mediante la matriz de correlaciones. Para ello usaremos la secuencia:

  Ver -> matriz de correlación

- Si colocamos la variable dependiente la última de la lista aparecerá al final de matriz y tendremos toda la información que nos interesa junta.

Análisis de relaciones con la variable dependiente



- La matriz de correlaciones muestra el coeficiente de correlación entre cada par de variables.
- **Nos interesa que las variables independientes estén muy relacionadas con la variable dependiente, para que puedan explicar adecuadamente su variabilidad.**

Análisis de relaciones con la variable dependiente

- Comprobamos que la relación es fuerte y lineal entre el gasto en reparaciones y la antigüedad y el gasto en reparaciones y el gasto en revisiones. Positiva en el caso de la primera y negativa con la última.
- La relación entre el gasto en reparaciones y las horas de funcionamiento de la maquinaria no es tan fuerte, es moderada y positiva.
- Nos interesará también que las variables independientes no presenten fuertes correlaciones entre sí. Si así fuera tendríamos un problema de multicolinealidad en el modelo.
- En este caso observando la matriz de correlaciones vemos hay una fuerte correlación entre el gasto en revisiones y la antigüedad, lo podría significar un problema de multicolinealidad.
- Obviaremos este problema que estudiaremos más adelante y continuaremos con nuestro análisis.

Estimación del modelo

El modelo que nos planteamos estimar es el siguiente:

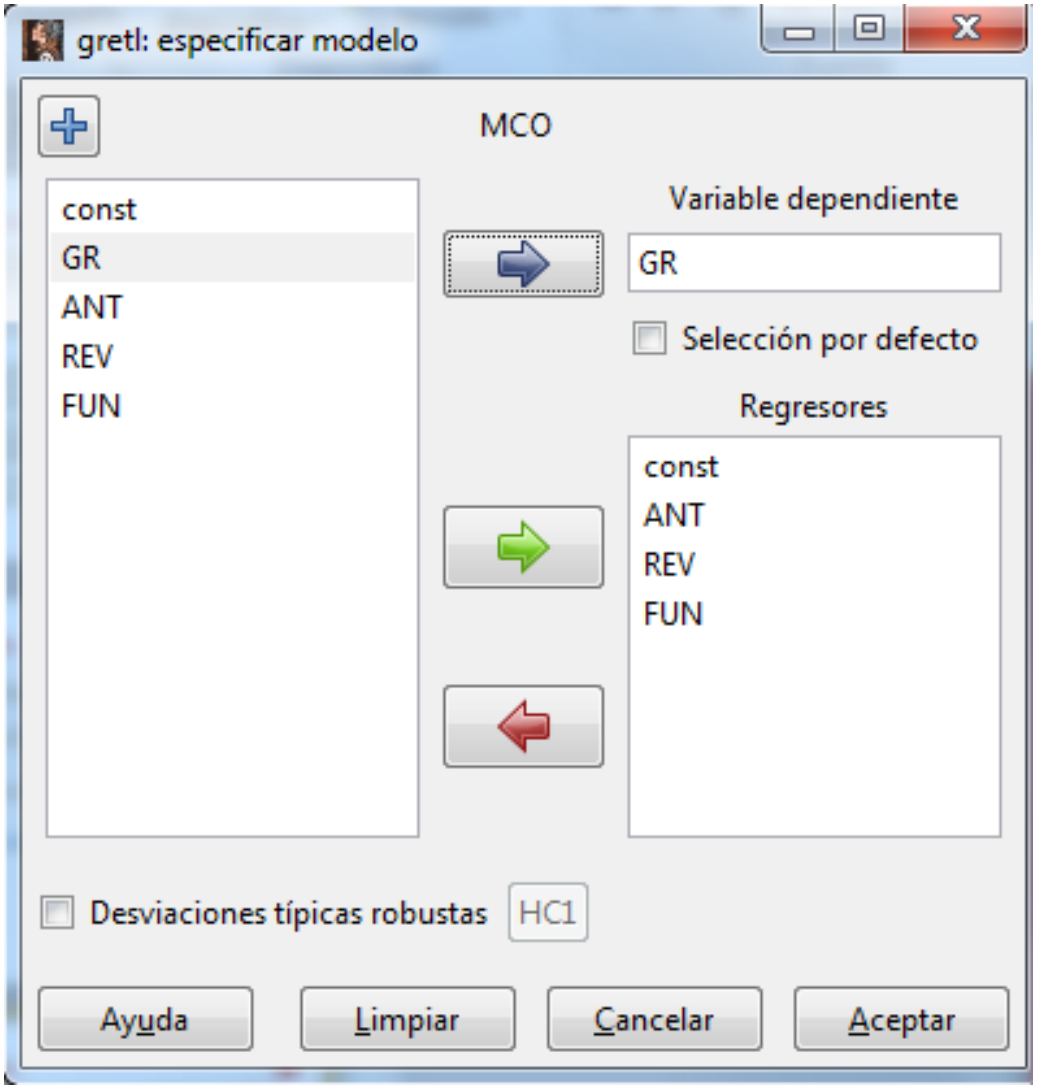
$$GR_i = \beta_0 + \beta_1 \cdot ANT_i + \beta_2 \cdot REV_i + \beta_3 \cdot FUN_i + \varepsilon_i$$

Para estimarlo usaremos la siguiente secuencia de Gretl:

Modelo -> Mínimos cuadrados ordinarios…

Estimación del modelo

- Debemos seleccionar la variable dependiente en el recuadro superior y las variables independientes en el inferior.
- “const” se refiere al término constante del modelo.



Estimación del modelo

El resultado de la estimación es el siguiente:

gretl: modelo 5

Archivo Editar Contrastes Guardar Gráficos Análisis LaTeX

```
Modelo 5: MCO, usando las observaciones 1-50
Variable dependiente: GR

             Coeficiente   Desv. Típica  Estadístico t  valor p
  ------------------------------------------------------------
  const      5399.31        825.951          6.537       4.52e-08 ***
  ANT          64.0932        9.85538        6.503       5.08e-08 ***
  REV          -7.57448       1.91090       -3.964       0.0003   ***
  FUN          -0.00198486    0.0548434     -0.03619     0.9713

Media de la vble. dep.   3200.207   D.T. de la vble. dep.    715.4294
Suma de cuad. residuos   825838.4   D.T. de la regresión     133.9888
R-cuadrado               0.967072   R-cuadrado corregido     0.964925
F(3, 46)                 450.3290   Valor p (de F)           4.34e-34
Log-verosimilitud       -313.7502   Criterio de Akaike       635.5004
Criterio de Schwarz      643.1485   Crit. de Hannan-Quinn    638.4129

Sin considerar la constante, el valor p más alto fue el de la variable 4 (FUN)
```

Estimación del modelo

- En la parte superior el software nos informa de las observaciones empleadas en la estimación del modelo y el nombre de la variable dependiente.

- A continuación aparece una tabla con los resultados de la estimación. La tabla incluye por columnas: los nombres de las variables, las estimaciones de cada uno de los coeficientes, las desviaciones típicas de los estimadores y el contraste de significatividad individual con su p-valor asociado.

- La zona inferior incluye un conjunto de estadísticos de bondad de ajuste, que nos permitirán realizar una evaluación parcial de la estimación realizada.

Estimación del modelo

- La columna *coeficiente* incluye el valor de los estimadores de los parámetros asociados a cada una de las variables explicativas. Estos valores se han obtenido mediante una estimación por el método de mínimos cuadrados ordinarios, a partir de la expresión matricial:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

- Sabemos por el teorema de Gauss-Markov que, si se cumplen las hipótesis clásicas del modelo de regresión múltiple, éstos estimadores son lineales, insesgados y óptimos (ELIO).

- Estos estimadores miden la magnitud de influencia de cada variable sobre la variable dependiente, entendiendo que las demás permanecen constantes.

Estimación del modelo

La ecuación estimada para este modelo es por tanto:

$$\widehat{GR} = 5399{,}31 + 64{,}09 \cdot ANT - 7{,}57 \cdot REV - 0{,}0019 \cdot FUN$$

- Con esta ecuación podríamos predecir el gasto anual en reparaciones que tendrá una máquina, conociendo su antigüedad, su gasto en revisiones y sus horas de funcionamiento anual, asumiendo que el modelo cumple con todas las hipótesis básicas.

Interpretación de los parámetros estimados

- Como ya adelantábamos, los parámetros del modelo miden la magnitud de la influencia de cada variable explicativa, sobre la variable dependiente.

- Tal y como está especificado nuestro modelo los coeficientes están midiendo ***la variación que experimenta la variable endógena ante un cambio de una unidad en la variable explicativa correspondiente***, suponiendo que el resto de variables permanecen constantes.

Interpretación de los parámetros estimados

En nuestro ejemplo diremos que:

- ✓ Cuando se aumenta la antigüedad de la máquina en una unidad, el gasto en reparaciones aumenta en 64,09 unidades, siempre que el resto de variables permanezcan constantes.
- ✓ Si el gasto en revisiones aumenta en una unidad, el gasto en reparaciones disminuye en 7,57 unidades , siempre que el resto de variables permanezcan constantes.
- ✓ Y finalmente, si las horas de funcionamiento anual aumentan en una unidad, el gasto en reparaciones disminuye en 0,0019 unidades , siempre que el resto de variables permanezcan constantes.

Por supuesto, esta interpretación esta siempre supeditada a la validez del modelo, que aún no hemos analizado.

Interpretación de los parámetros estimados

Es frecuente que los modelo requieran incluir transformaciones logarítmicas en algunas variables para resolver algunos problemas que pueden surgir.

Si utilizamos este tipo de transformaciones, la interpretación de los parámetros es diferente y se resume en la siguiente tabla:

Modelo	Variable dependiente	Variable independiente	Interpretación del parámetro
Nivel-nivel	y	x	$\Delta y = \beta_i \Delta x$
Nivel-log	y	Log(x)	$\Delta y = (\beta_i/100)\%\Delta x$
Log-nivel	Log(y)	x	$\Delta\%y = (100\beta_i)\Delta x$
Log-log	Log(y)	Log(x)	$\Delta\%y = \beta_i\%\Delta x$

(Wooldridge, J. Introducción a la econometría: un enfoque moderno. 2006. Ed. Thomson. Pg. 49)

4. Normalidad, inferencia y bondad de ajuste.

Contraste de normalidad

- Antes de comenzar a analizar los contrastes de hipótesis asociados al modelo de regresión lineal múltiple es importante comprobar si se cumple la hipótesis de normalidad del error (H7).

- Recordemos que la hipótesis 7 decía: “El término error es un término completamente aleatorio que sigue una distribución normal, de esperanza 0 (E[εi]=0).”

❑ **Los contrastes** de hipótesis asociados al modelo de regresión múltiple son todos contrastes paramétricos, por lo que **no serán aplicables si no se cumple esta hipótesis de normalidad del error**.

❑ Para comprobar la normalidad del error se puede utilizar cualquiera de los contrastes de normalidad que incorpora Gretl sobre la serie de los residuos del modelo.

Contraste de normalidad

Para realizar el contraste de normalidad en Gretl podemos guardar los residuos a través del menú del modelo:

Guardar -> Residuos

Y después aplicar a esta serie cualquiera de los contrastes de normalidad (desde el menú general).

Otra opción es realizar el contraste de normalidad para los residuos directamente a través del menú del modelo:

Contrastes -> Normalidad de los residuos

En este caso nos dará la distribución de frecuencias y el contraste de Doornik-Hansen.

Contraste de normalidad

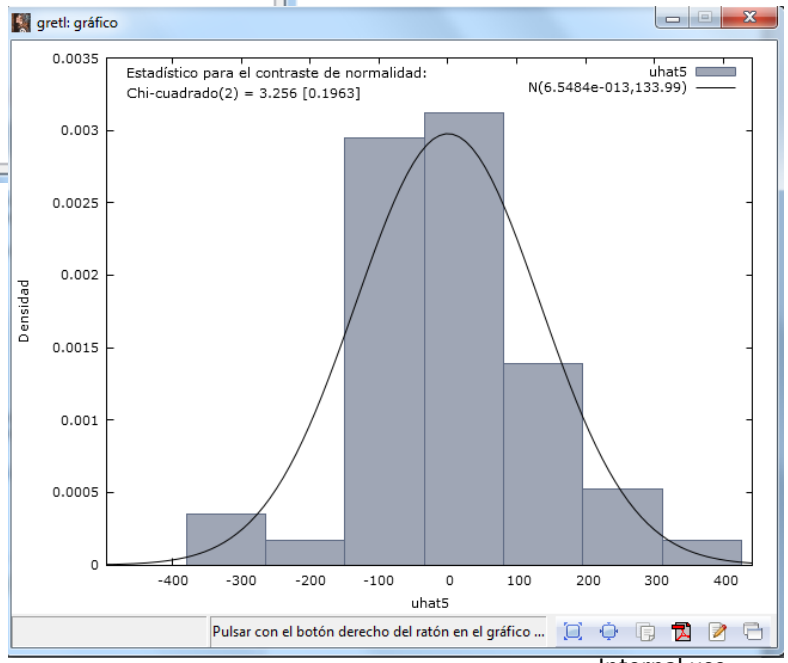
gretl: dist.. de los residuos.

```
Distribución de frecuencias para uhat5, observaciones 1-50
número de cajas = 7, media = 6.54836e-013, desv.tip.=133.989

      intervalo     punto medio   frecuencia  rel     acum.

          < -264.94   -322.55         2      4.00%    4.00% *
-264.94 - -149.72     -207.33         1      2.00%    6.00%
-149.72 - -34.507     -92.115        17     34.00%   40.00% ************
-34.507 -  80.710      23.102        18     36.00%   76.00% ************
 80.710 -  195.93      138.32         8     16.00%   92.00% *****
 195.93 -  311.14      253.54         3      6.00%   98.00% **
       >=  311.14      368.75         1      2.00%  100.00%

Contraste de la hipótesis nula de distribución normal:
Chi-cuadrado(2) = 3.256 con valor p  0.19632
```



Comprobamos mediante el contraste de Doornik-Hansen que los residuos provienen de una distribución normal, por lo que podemos considerar que el error en normal y se cumple la hipótesis 7.

Contraste de normalidad

- El cumplimiento de esta hipótesis nos permite seguir adelante con el estudio.

- *Si no se cumpliese* la hipótesis de normalidad, *los contrastes* que utilizaremos a partir de ahora para comprobar el grado de validez del modelo *no serían válidos*, ya que todos están basados en esta hipótesis.

- Tendríamos por tanto que reespecificar el modelo, porque no sería válido.

- Es importante, por tanto realizar este contraste antes que el resto, para asegurar su validez.

Diagnosis – Error estándar

- Volviendo a la información que proporciona Gretl sobre la estimación del modelo, junto a la columna de coeficientes tenemos la columna **Desv. Típica** que muestra la varianza de los estimadores de los parámetros del modelo o error estándar de estimación.

- La desviación típica de los estimadores, mide la precisión con la que dichos estimadores estiman los parámetros del modelo. Es, por tanto un indicador del grado de confianza que podemos tener en la estimación.

- Sabemos por el teorema de Gauss- Markov que, siempre que se cumplan las hipótesis básicas del modelo, los estimadores de mínimos cuadrados ordinarios son **eficientes**, es decir, que tienen la menor varianza (y en consecuencia desviación típica) que pueden tener.

Diagnosis – Contraste de significatividad individual

- Las dos últimas columnas contienen el estadístico t y su p-valor asociado. Ambas hacen referencia al contraste de significatividad individual.
- Este contraste contrasta la hipótesis nula de que la variable considerada no es individualmente significativa para explicar el comportamiento de la variable dependiente, es decir:

  $H_0: \beta_i=0$

  $H_1: \beta_i\neq0$

- El estadístico de contraste de este test es: $$t-statistic=\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

- Sabemos que bajo la hipótesis nula, el estadístico t sigue una distribución t de student con n-k-1 grados de libertad.

- La región crítica será por tanto: $|t|\geq t_{n-k-1,\alpha/2}$

Diagnosis – Contraste de significatividad individual

- La forma más sencilla de trabajar con el contraste de significatividad individual es tomar la decisión sobre la hipótesis nula en base al p-valor. Recordemos que el p-valor nos indica la probabilidad de cometer el error de rechazar la hipótesis nula siendo cierta (error de tipo I).

- En el modelo planteado no tenemos evidencias suficientes para rechazar la hipótesis nula de no significatividad de las variables, únicamente en el caso de las horas de funcionamiento anuales de la máquina.

- Para el resto de variables, *la hipótesis nula de no significatividad individual de la variable se debe rechazar*, por lo que aceptaremos que **sí son significativas**.

Diagnosis

- En la parte inferior, lo primero que aparece es **la media aritmética de la variable dependiente**, que como sabemos coincidirá con la de su estimación. Al lado está la **desviación típica de la variable dependiente**.

- A continuación tenemos la **suma de cuadrados de los residuos (SCE)** o variabilidad no explicada por el modelo (VNE). Se obtiene mediante:

$$Y^tY - \hat{\beta}^tX^tY$$

- El dato **D.T. de la regresión** es la estimación de la desviación típica del error, de modo que su cuadrado es la varianza del error. Se obtiene mediante la siguiente expresión:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{Y^tY - \hat{\beta}^tX^tY}{n-k}}$$

Diagnosis – Coeficiente de determinación

- En la tabla inferior aparece el coeficiente de determinación (R-cuadrado (R^2)), que como ya sabemos es una medida para valorar la capacidad explicativa de la regresión. Se define como el cociente entre la variación explicada o suma explicada de cuadrados (SEC) y la variación total:

$$R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

- *Nos informa por tanto del porcentaje de variación de la variable dependiente que conseguimos explicar con el modelo.*

- En nuestro ejemplo vemos que la capacidad explicativa de las variables es elevada, pues explican el 96,70% de la variabilidad de la variable endógena.

Diagnosis – Coeficiente de determinación corregido

- El coeficiente de determinación corregido, R-cuadrado corregido ($\bar{R}^2$), se obtiene a partir del R^2, ponderándolo en base al número de variables que incluye el modelo al tamaño muestral:

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

- Este coeficiente permite comparar la capacidad explicativa de modelos referidos a una misma muestra de la misma variable dependiente con distinto número de variables independientes.

- En nuestro caso obtenemos un valor de 0,9649. Si lo comparamos con el valor que obtenemos al estimar un modelo con las mismas variables explicativas menos las horas de funcionamiento (0,9656), podemos concluir que este último modelo es mejor que el anterior por ser el R cuadrado ajustado mayor.

- Si comparásemos directamente el coeficiente de determinación observaríamos que este siempre sube al incluir nuevas variables, aunque no sean significativas.

Diagnosis – Contraste de significatividad global

- El contraste de significatividad global (F(3,46)), permite contrastar:

$$H_0: \beta_1=\beta_2=\ldots=\beta_k=0$$

$$H_1: \exists\ \beta_i \neq 0$$

- Contrasta por tanto si todos los parámetros asociados a cada una de las variables explicativas del modelo son iguales a cero. Es por tanto, una forma de comprobar si el modelo es válido o debemos especificarlo de nuevo.
- La alternativa indica que al menos uno de los parámetros es significativo o distinto de cero.

$$F = \frac{VNE/K}{VT/n-K-1}$$

- Bajo la hipótesis nula, el estadístico F sigue una distribución F-Snedecor (n-1;n-k-1), de modo que la región crítica será: F>F-Snedecor (n-1;n-k);α

Diagnosis

- El contraste de significatividad global se puede entender como la forma de comprobar si el coeficiente de determinación del modelo es suficientemente grande como para considerar que el modelo tiene una capacidad explicativa adecuada.

- En nuestro caso, con un p-valor de 4.34e-34 se debe rechazar la hipótesis nula, por lo que se acepta que el modelo es significativo.

- **Log verosimilitud** nos da el valor máximo de la función de verosimilitud. Este es el valor de la función de verosimilitud para los parámetros, ya que estos son los más verosímiles o los que maximizan la probabilidad de la muestra.

Diagnosis - Criterios de información de Akaike y Schwarz

- Los criterios de información de Akaike y Schwarz proporcionan un método para la elección del mejor modelo a partir de una misma muestra.

- Ambos son medidas de la perdida de información que tenemos al modelizar la variable con el modelo planteado.

- Se calculan a partir de las siguientes expresiones:

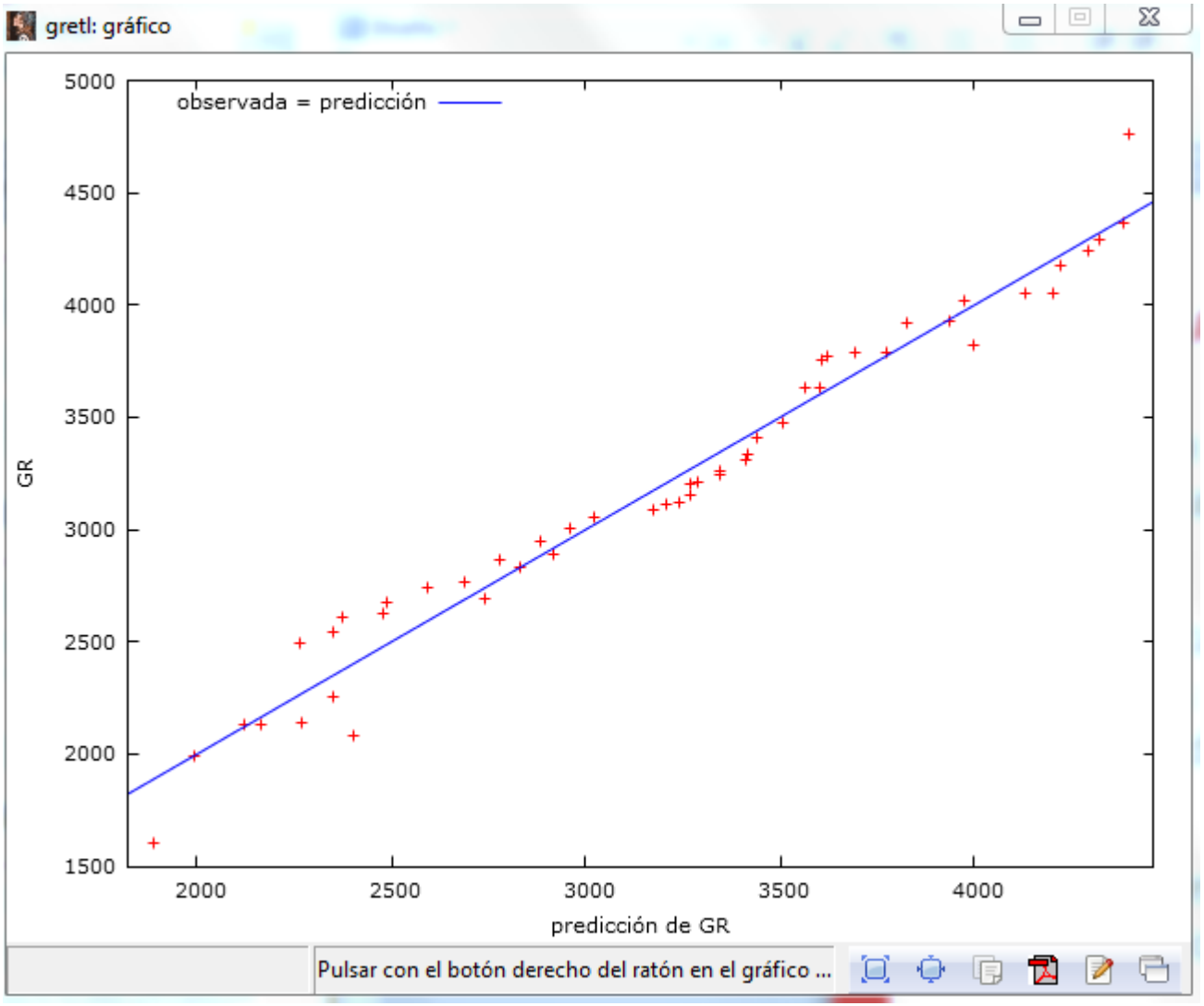
  - AIC = -2*(L/N)+2*((k+1)/N)
  - BIC = -2*(L/N) + (k+1)*Ln(N)/N

- Siendo L el valor máximo de la función de verosimilitud, k el número de variables explicativas del modelo y N el tamaño muestral.

Diagnosis – Algunas herramientas de Gretl

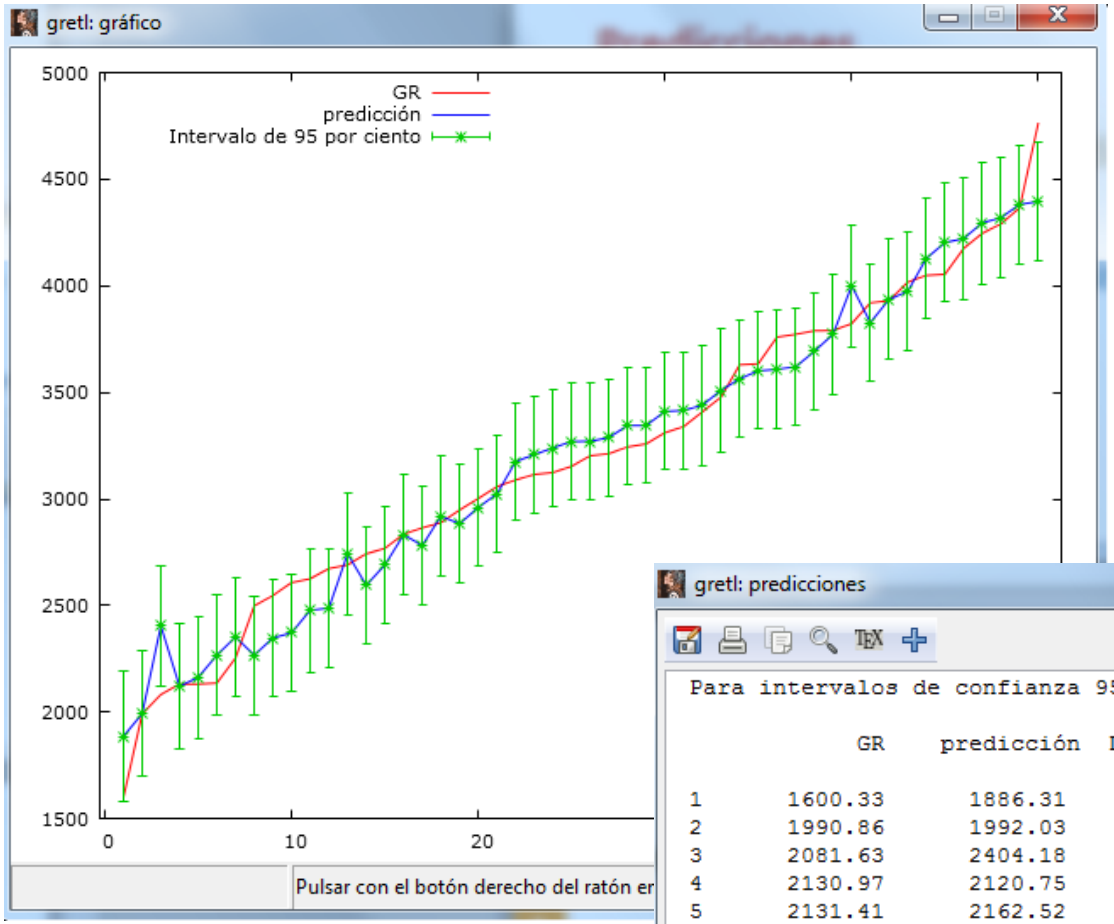
- La ventana del modelo en Gretl tiene su propio menú que incorpora acciones que podemos realizar con el modelo.
- En el apartado de contrastes se incluyen importantes contrastes sobre el modelo que serán de interés, como el contraste de normalidad del error que ya hemos utilizado.
- A través del menú guardar, podemos guardar como escalares los valores obtenidos en la estimación.
- Y el menú de gráficos obtiene interesantes gráficos sobre los residuos y sobre las predicciones.

Gráfico de la variable observada frente a la estimada



Predicciones

Podemos obtener las predicciones del modelo mediante el menú: Análisis -> Predicciones



gretl: predicciones

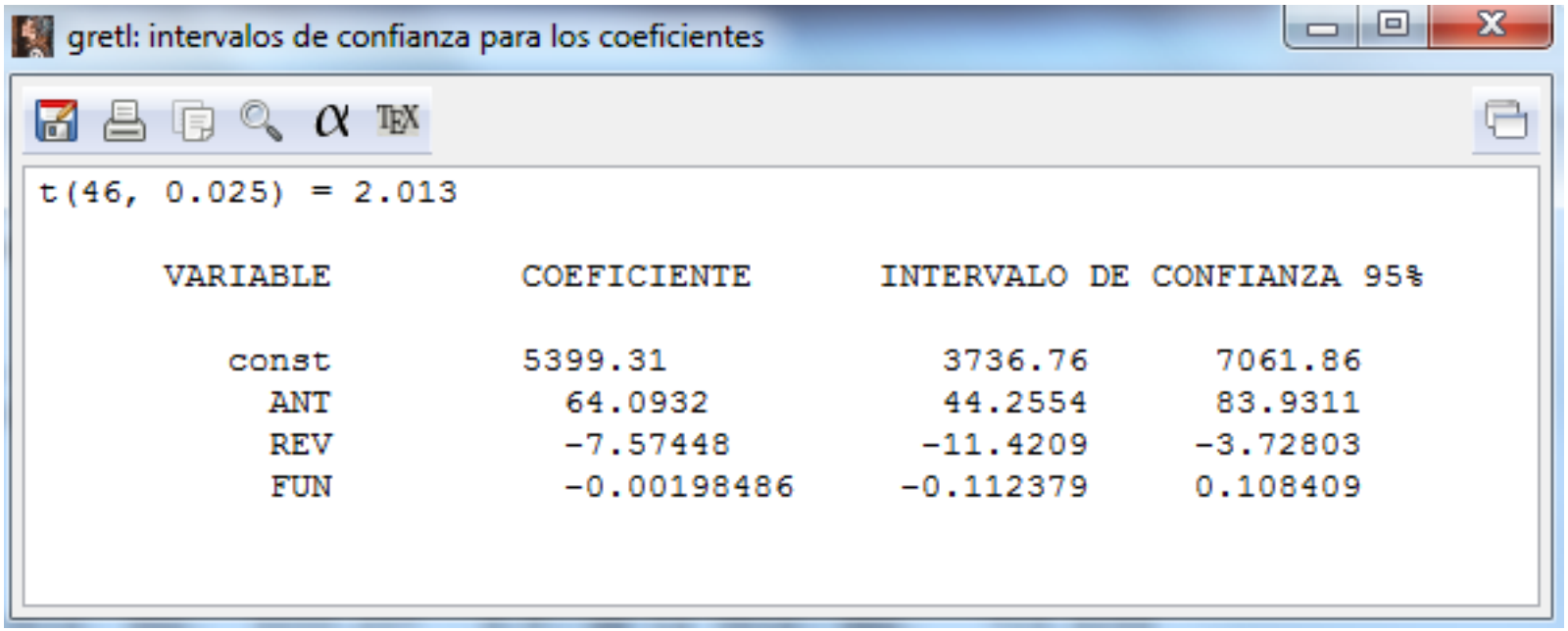
Para intervalos de confianza 95%, t(46, .0.025) = 2.013

	GR	predicción	Desv. Típica	Intervalo de confianza 95%
1	1600.33	1886.31	150.799	1582.76 - 2189.85
2	1990.86	1992.03	146.897	1696.34 - 2287.72
3	2081.63	2404.18	141.573	2119.21 - 2689.15
4	2130.97	2120.75	147.111	1824.63 - 2416.87
5	2131.41	2162.52	141.671	1877.35 - 2447.69
6	2135.92	2266.79	139.620	1985.75 - 2547.83
7	2253.15	2351.78	139.619	2070.74 - 2632.81
8	2497.79	2266.29	138.009	1988.49 - 2544.09
9	2546.45	2347.61	137.396	2071.05 - 2624.17

Intervalos de confianza para los coeficientes

Además de la predicción puntual, podemos obtener intervalos de confianza para los coeficientes a través del menú:

Análisis -> Intervalos de confianza para los coeficientes



Matriz de varianzas-covarianzas de los estimadores

La matriz de varianzas-covarianzas de los estimadores se obtiene mediante:

Análisis -> Matriz de covarianzas de los coeficientes