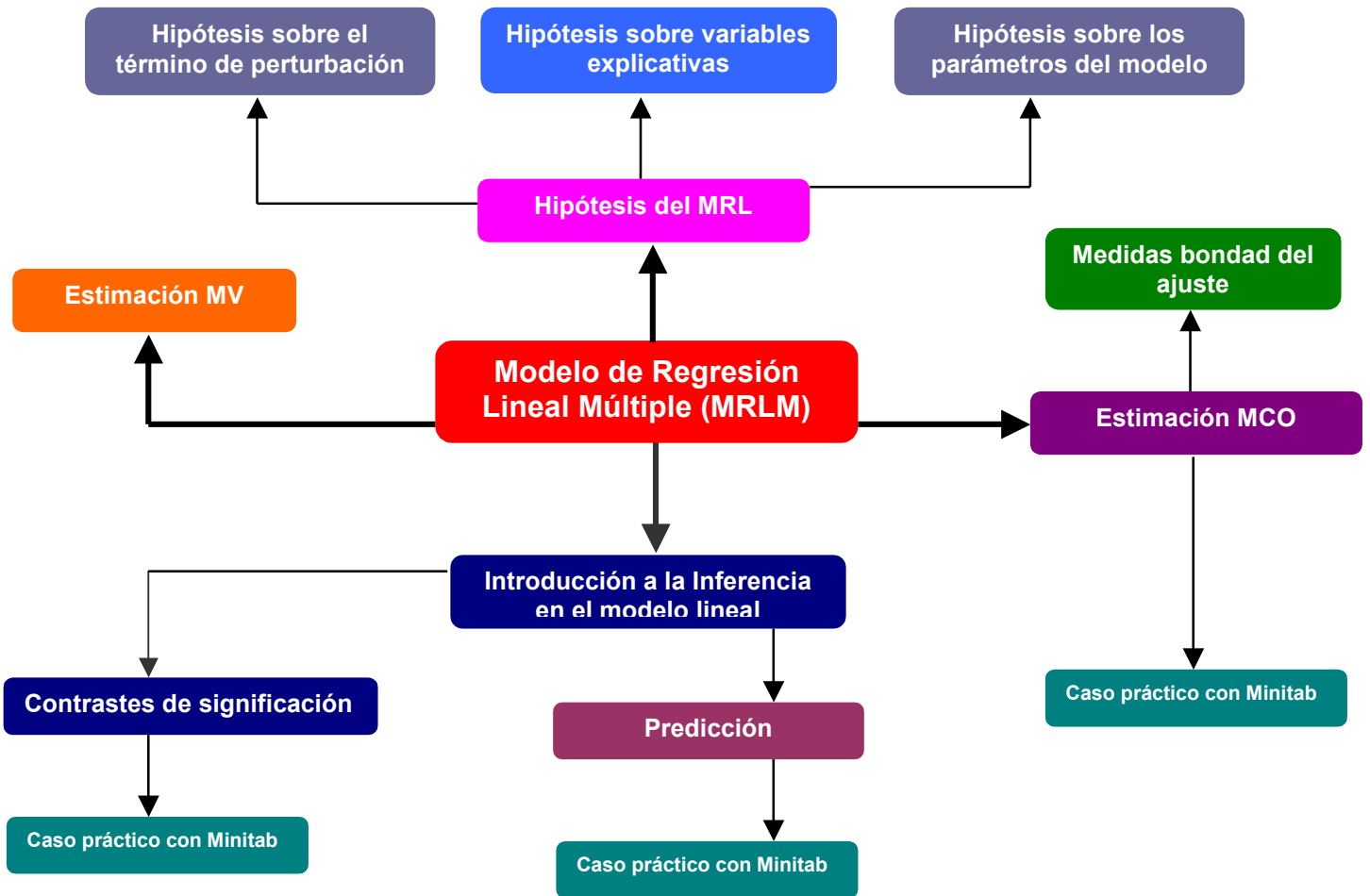


MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Autores: Renatas Kizys (rkizys@uoc.edu), Ángel A. Juan (ajuanp@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

Todo estudio econométrico se centra en dos pilares básicos: la teoría y los hechos. La teoría permite derivar un modelo (el modelo económico) que sintetiza la incógnita relevante sobre el fenómeno (la variable endógena) objeto del análisis y del cual deriva el modelo econométrico que permite medirlo y contrastarlo empíricamente. Los hechos se concretan en una serie de datos que denominaremos información muestral. La muestra, a su vez, consiste en una lista ordenada de valores numéricos de las variables objeto de estudio. En una muestra de corte transversal, diversos agentes económicos de una naturaleza similar proporcionan información solicitada en un mismo instante de tiempo. Alternativamente, el investigador económico trabaja en ocasiones con datos de series temporales, en las que se dispone de información acerca de

unidad económica, como puede ser un país, una empresa, a lo largo de tiempo; estas muestras pueden tener frecuencia diaria, mensual, anual, según frecuencia de observación de los datos. Una vez que se especifica el modelo y se dispone de la información estadística convenientemente tratada, se llega a la etapa siguiente del trabajo econométrico: la etapa de estimación. Los resultados de esta etapa de estimación permiten medir y contrastar las relaciones sugeridas por la teoría económica.

En este *math-block* postularemos una serie de hipótesis básicas de un modelo de regresión múltiple (MRLM) y consideremos los principales métodos de estimación bajo dichas hipótesis. Veremos que los estimadores obtenidos mediante el método de mínimos cuadrados ordinarios (MCO) son insesgados, eficientes y consistentes. Además, utilizaremos la inferencia basada en los contrastes de hipótesis para apreciar estadísticamente una cierta evidencia empírica. Finalmente, conoceremos la importancia del MRLM en la predicción y pronóstico de un cierto fenómeno comprendido por la variable endógena.

OBJETIVOS

- Conocer la estructura del MRLM.
- Familiarizarse con las hipótesis básicas del MRLM y entender su importancia.
- Conocer los métodos de estimación del MRLM, el método de mínimos cuadrados ordinarios (MCO) y el de máxima verosimilitud (MV).
- Introducirse en el uso de Minitab para estimar el MRLM mediante el MCO.
- Saber cuantificar e interpretar bondad del ajuste del modelo.
- Evaluar la contribución de cada variable exógena en explicar el comportamiento de la variable endógena; contrastar la significación individual de un parámetro y la global del modelo.
- En base de la estimación de MRLM, realizar predicciones puntuales y por intervalo de la variable endógena.

CONOCIMIENTOS PREVIOS

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

- Intervalos de confianza y contraste de hipótesis para 1 y 2 poblaciones
- Análisis de regresión y correlación lineal
- Correlación y regresión lineal múltiple

CONCEPTOS FUNDAMENTALES

□ Hipótesis del modelo de regresión lineal múltiple (MRLM)

Mediante un modelo de regresión lineal múltiple (MRLM) tratamos de explicar el comportamiento de una determinada variable que denominaremos variable a explicar, variable endógena o variable dependiente, (y representaremos con la letra Y) en función de un conjunto de k variables explicativas X_1, X_2, \dots, X_k mediante una relación de dependencia lineal (suponiendo $X_1 = 1$):

$$Y = \beta_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + U \quad \text{siendo U el término de perturbación o error}$$

Para determinar el modelo anterior, es necesario hallar (estimar) el valor de los coeficientes $\beta_1, \beta_2, \dots, \beta_k$. La linealidad en parámetros posibilita la interpretación correcta de los parámetros del modelo. Los parámetros miden la intensidad media de los efectos de las variables explicativas sobre la variable a explicar y se obtienen al tomar las derivadas parciales de la variable a explicar respecto a cada una de las variables explicativas:

$$\beta_j = \frac{\partial Y}{\partial X_j}; j = 1, \dots, k.$$

Nuestro objetivo es asignar valores numéricos a los parámetros $\beta_1, \beta_2, \dots, \beta_k$. Es decir, trataremos de estimar el modelo de manera que, los valores ajustados de la variable endógena resulten tan próximos a los valores realmente observados como sea posible.

A fin de poder determinar las propiedades de los estimadores obtenidos al aplicar distintos métodos de estimación y realizar diferentes contrastes, hemos de especificar un conjunto de hipótesis sobre el MRLM que hemos formulado. Existen tres grupos de hipótesis siguientes: *las hipótesis sobre el término de perturbación, las hipótesis sobre las variables explicativas, y las hipótesis sobre los parámetros del modelo.*

Hipótesis sobre el término de perturbación:

Para una muestra de n observaciones (cada observación estará formada por una tupla con los valores de X_2, X_3, \dots, X_k y el valor de Y asociado), tendremos el siguiente sistema de n ecuaciones lineales:

$$\begin{cases} Y_1 = \beta_1 + \beta_2 \cdot X_{21} + \dots + \beta_k \cdot X_{k1} + u_1 \\ Y_2 = \beta_1 + \beta_2 \cdot X_{22} + \dots + \beta_k \cdot X_{k2} + u_2 \\ \dots \\ Y_n = \beta_1 + \beta_2 \cdot X_{2n} + \dots + \beta_k \cdot X_{kn} + u_n \end{cases}$$

o, en forma matricial: $\mathbf{Y} = \mathbf{X} \cdot \mathbf{B} + \mathbf{U}$, donde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{2n} & \dots & X_{kn} \end{bmatrix}, B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}, U = \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix}$$

En estas condiciones, las hipótesis del MRLM se resumen en la esfericidad del término de perturbación, i.e.:

- a) El valor esperado de la perturbación es cero: $E[u_i] = 0 \quad \forall i = 1, \dots, n$
- b) Homoscedasticidad: todos los términos de perturbación tienen la misma varianza (varianza constante):

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

Por tanto, todos los términos de la diagonal principal de la matriz de varianzas y covarianzas serán iguales:

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

- c) No Autocorrelación: los errores son independientes unos de otros, i.e.: la matriz de varianzas y covarianzas es una matriz diagonal (fuera de la diagonal principal todo son ceros):

$$Var[U] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Observar que, bajo las hipótesis de homoscedasticidad y no autocorrelación, la matriz de varianzas y covarianzas tendrá la forma siguiente:

$$Var[U] = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \cdot I_n \quad (I_n \text{ es la matriz identidad de orden } n)$$

- d) El error o perturbación sigue una distribución normal, i.e.:

$$U \approx N(0_n, \sigma^2 \cdot I_n)$$

Hipótesis sobre las variables explicativas:

- a) Las variables explicativas son fijas o deterministas.
- b) La variables explicativas están no correlacionadas con la perturbación aleatoria.

- c) Las variables explicativas no presentan relación lineal exacta entre si.
- d) Además, supondremos que las variables explicativas son medidas sin error.
- e) En el modelo no se excluyen las variables relevantes y que tampoco no se incluyen las variables irrelevantes, a la hora de explicar el comportamiento de la variable endógena.

Hipótesis sobre los parámetros del modelo:

- a) La única hipótesis que haremos acerca de los parámetros del modelo es la hipótesis de permanencia estructural, lo cual quiere decir que los parámetros poblacionales, β_j , se mantienen constantes a lo largo de toda la muestra.

□ **Estimación del MRLM**

Estimar el modelo equivale asignar valores numéricos a los parámetros desconocidos $\beta_1, \beta_2, \dots, \beta_k$, a partir de la información muestral disponible de las variables observables del modelo. Únicamente consideraremos dos métodos de estimación:

- El método de mínimos cuadrados ordinarios (MCO)
- El método de máxima verosimilitud (MV)

Estimación por mínimos cuadrados ordinarios:

Sea un modelo en forma matricial $\mathbf{Y} = \mathbf{X} \cdot \mathbf{B} + \mathbf{U}$. Supongamos que el modelo ha sido estimado, obteniéndose $\hat{\mathbf{Y}}$, vector de valores de la variable dependiente implicado por el modelo. La diferencia entre los valores observados y los valores estimados, $e = Y - \hat{Y} = Y - X \cdot \hat{B}$, la denominaremos vector de residuos. Ahora bien, nuestro problema consiste en minimizar la suma de los cuadrados de residuos, $e'e$ con respecto del vector de parámetros estimados, \mathbf{B} . De este problema de optimización se deduce la siguiente expresión de mínimos cuadrados ordinarios del MRLM [7]:

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

cuya varianza viene dada por: $Var[\hat{B}] = \sigma^2 (X' \cdot X)^{-1}$

Además, el estimador MCO de la varianza del término de perturbación es:

$$\hat{\sigma}_u^2 = \frac{e' \cdot e}{n - k}$$

donde n es el número de observaciones y k es el número de elementos del vector \mathbf{B} .

Bajo la hipótesis de perturbaciones esféricas, el estimador MCO del vector \mathbf{B} cumple una serie de propiedades que le convierten en un insesgado (el valor esperado del estimador coincide con el valor real del parámetro), eficiente (de varianza mínima), y consistente [4].

Además, bajo la hipótesis de esfericidad, el estimador MCO de la varianza del término de error, $\hat{\sigma}_u^2$, es también insesgado.

Estimación por máxima verosimilitud:

El método de estimación por MCO consiste en asignar valores numéricos a los parámetros desconocidos de manera que la suma cuadrática de errores sea mínima y sólo requiere que la matriz $X'X$ sea invertible. A continuación veremos un método de estimación alternativo, el método de máxima verosimilitud.

El método de máxima verosimilitud (MV), en cambio, propone como un estimador el valor que maximiza la probabilidad de obtener la muestra ya disponible.

El método MV se basa, prácticamente, en la distribución que sigue el término de error. A tales efectos, se suele suponer que las perturbaciones aleatorias se distribuyen con una distribución Normal que, además de cumplir las propiedades de una muestra grande, es una aproximación cómoda y fácil de tratar.

El modelo que utilizaremos es $Y = X \cdot B + U$, y supondremos que el término aleatorio sigue la distribución Normal con la siguiente función de densidad:

$$f(u_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\}, i = 1, \dots, N.$$

Maximizar la probabilidad de obtener la muestra ya disponible equivale maximizar la función de densidad conjunta del vector aleatorio, u . Para ello, hemos de suponer homoscedasticidad y ausencia de autocorrelación. Por tanto, la expresión de la función de densidad conjunta es la siguiente:

$$f(U) = \prod_{i=1}^n f(u_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum u_i^2}{2\sigma^2}\right\}$$

Como U sigue una distribución Normal Multivariante de orden k , la variable Y , al ser una combinación lineal de las perturbaciones aleatorias, también se distribuirá con una distribución Normal Multivariante. Así pues, para que la función de densidad conjunta sea una función de verosimilitud, el vector aleatorio U ha de expresarse en función del vector Y , es decir:

$$L(Y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{(Y - X\beta)(Y - X\beta)}{2\sigma^2}\right\}$$

Se trata, por tanto, de maximizar la función de verosimilitud. Como la expresión anterior resulta complicada, aplicaremos una transformación monótona; en concreto, una función logarítmica:

$$\ln L(Y; \beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(2\sigma^2) - \frac{(Y - X\beta)(Y - X\beta)}{2\sigma^2}$$

Derivando la función de verosimilitud con respecto de B y σ^2 , e igualando las derivadas a cero, obtenemos los resultados:

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

cuya varianza es la siguiente: $Var[\hat{B}_{MV}] = \sigma^2 (X' \cdot X)^{-1}$.

Además, el estimador MCO de la varianza del término de perturbación es:

$$\sigma_{MV}^2 = \frac{e' \cdot e}{n},$$

donde n es el número de observaciones y k es el número de elementos del vector B.

Observamos que el estimador de MV de **B** coincide con el MCO, con lo que tendrá las mismas propiedades: será lineal, insesgado, óptimo y consistente.

Es fácil ver que el estimador de MV de σ^2 , en cambio, resulta diferente del MCO y no es insesgado aunque sí es asintóticamente insesgado.

□ Medidas del bondad del ajuste

Las estimaciones por MCO y MV que hemos realizado todavía no nos permite evaluar la calidad de ajuste del modelo. Para ello, de aquí a delante iremos viendo las medidas de bondad de ajuste.

Comenzaremos por la suma de los cuadrados de errores, SCE, que puede expresarse de varias formas:

$$e' \cdot e = \sum_{i=1}^n e_i^2 = Y' \cdot Y - \hat{B}' \cdot X' \cdot Y = Y' \cdot Y - \hat{Y}' \cdot \hat{Y} = \sum_{i=1}^n Y_i^2 - \sum_{i=1}^n \hat{Y}_i^2.$$

Despejando la suma de cuadrados de la variable endógena, queda:

$$Y' \cdot Y = \hat{Y}' \cdot \hat{Y} + e' \cdot e, \text{ o bien, } \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \hat{Y}_i^2 + \sum_{i=1}^n e_i^2.$$

Restando a ambos lados la cantidad $n \cdot \bar{Y}^2$, obtenemos:

$$Y' \cdot Y - n \cdot \bar{Y}^2 = \hat{Y}' \cdot \hat{Y} - n \cdot \bar{Y}^2 + e' \cdot e, \text{ o bien, } \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2.$$

La parte izquierda representa suma de cuadrados totales (SCT) y no es sino la suma de cuadrados de las desviaciones respecto a su media aritmética.

Por otra parte, si el modelo tiene término independiente, a la cantidad $\hat{Y}' \cdot \hat{Y} - n \cdot \bar{Y}^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ se le denomina suma de cuadrados de la regresión (SCR).

En resumen, la variabilidad total de la variable endógena (SCT) puede descomponerse en dos partes: la parte que podemos explicar mediante el modelo especificado (SCR) y la parte que no podemos explicar, la suma de cuadrados de los errores (SCE).

A partir de la descomposición anterior de la SCT, definiremos el coeficiente de determinación, R^2 , el cual será la primera medida de bondad de ajuste:

$$R^2 = 1 - \frac{SCE}{SCT}.$$

Si el modelo tiene término independiente, entonces se cumple la igualdad $SCT = SCR + SCE$, y el coeficiente de determinación podrá expresarse de la siguiente manera alternativa:

$$R^2 = \frac{SCR}{SCT}.$$

El coeficiente de determinación indica que proporción de variabilidad total queda explicada por la regresión. Si el modelo tiene término independiente, entonces R^2 toma valores entre 0 y 1.

En práctica, el uso de R^2 presenta algunas limitaciones a la hora de comparar varios modelos desde la perspectiva de bondad del ajuste. En efecto, cuanto más variables explicativas incorporamos al modelo, mayor será el coeficiente de determinación, pues la SCR disminuye conforme aumenta el número de variables explicativas. Por tanto, cuando queremos llevar a cabo un análisis comparativo entre varios modelos, utilizamos R^2 corregido:

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

Este estadístico es inmune ante la incorporación de extra variables vía interacción de dos efectos: el efecto que permite aumentar R^2 , y el efecto opuesto que surge al descontar un mayor número de las variables explicativas, $\frac{n-1}{n-k}$ [7].

□ Significación de los parámetros del modelo

Distinguiremos entre dos distintas dimensiones de significación: significación económica y significación estadística.

Significación económica

Significación económica nos permite comprobar si las estimaciones obtenidas son coherentes con la teoría económica. Según especificación del modelo, la interpretación y significación de los parámetros puede variar. Si el modelo está especificado en niveles, el parámetro refleja el efecto medio que tiene una variación unitaria de la variable explicativa sobre la variable endógena:

$$\beta_j = \frac{\partial Y}{\partial X_j}.$$

En cambio, si el modelo está especificado en logaritmos neperianos, los parámetros pueden interpretarse como una elasticidad, como es el caso de la función de producción de Cobb-Douglas:

$$\beta_j = \frac{\partial \ln Y}{\partial \ln X_j}.$$

Significación estadística

El análisis econométrico pretende analizar, por medio una serie de contrastes, la significación (o significatividad) estadística individual y conjunta de los parámetros del modelo. En concreto, para contrastar las hipótesis de significatividad individual, tenemos:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_A &: \beta_j \neq 0. \end{aligned}$$

El estadístico **t-Student** que se utiliza para realizar el test es el siguiente:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_u^2 \cdot a_{jj}}} \sim t_{n-k}.$$

donde $\sqrt{\hat{\sigma}_u^2 \cdot a_{jj}}$ es el error estándar estimado de $\hat{\beta}_j$, y a_{jj} es el j-ésimo elemento de la diagonal principal de la matriz $(X'X)^{-1}$.

Dado un nivel de significación α , las tablas de distribuciones nos proporcionan la cantidad $t_{n-k, \alpha/2}$ que es el valor asociado a una t-Student con n-k grados de libertad que deja a su derecha un área de $\alpha/2$ (o, equivalentemente, deja a su izquierda un área de $1 - \alpha/2$). La regla de decisión que utilizaremos para determinar si el parámetro asociado a la variable X_j es individualmente significativo o no es la siguiente:

- Si $|t_j| \geq t_{n-k, \alpha/2}$, el estadístico cae fuera de la región de aceptación, por lo que rechazamos la hipótesis nula. Concluimos, por tanto, que el parámetro es significativamente diferente de cero.
- Si $|t_j| < t_{n-k, \alpha/2}$, el estadístico cae dentro de la región de aceptación, por lo que no podemos rechazar la hipótesis nula. Por tanto, el parámetro no es individualmente significativo.

Nota: si en vez de realizar el contraste bilateral deseamos hacer un contraste unilateral (en el cual la hipótesis alternativa sería $H_1 : \beta_j > 0$ ó $H_1 : \beta_j < 0$), deberemos sustituir en la fórmula anterior $\alpha/2$ por α (ya que ahora trabajaremos con una única cola de la distribución).

En cambio, si queremos contrastar la significación conjunta, las hipótesis especificamos de la manera siguiente:

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = \dots = \beta_k = 0 \\ H_A &: \text{No } H_0. \end{aligned}$$

Nota: el término independiente no contribuye en explicar la variabilidad de la variable endógena, con lo cual no lo incluimos en la restricción.

El estadístico **F de Snedecor** que se utiliza para realizar el test es el siguiente:

$$F_0 = \frac{R^2}{1 - R^2} \cdot \frac{n - k}{n - 1} \sim F_{k-1, n-k}.$$

El estadístico se distribuye bajo la hipótesis nula con una distribución F de Snedecor con k-1 grado de libertad en el numerador y n-k grados de libertad en el denominador. La regla de decisión utilizada para contrastar la significación global del modelo es la siguiente:

- Si $F_0 \geq F_{k-1, N-k; \alpha}$, el estadístico de contraste cae fuera de la región de aceptación, con lo que rechazamos la hipótesis nula. Por tanto, el modelo es globalmente significativo.
- Si $F_0 < F_{k-1, N-k; \alpha}$, el estadístico de contraste cae dentro de la región de aceptación, de modo que ahora la hipótesis nula no la rechazamos. En consecuencia, podemos afirmar que el modelo no es globalmente significativo.

□ Predicción

Una vez hemos especificado, estimado y validado un modelo, podemos utilizarlo con objetivos diferentes.

Cuando trabajamos con una serie temporal, podemos estar interesados en predecir el comportamiento futuro de la variable endógena. Si, por otro lado, trabajamos con un corte transversal (o una sección cruzada), podemos utilizar el modelo ajustado para predecir el comportamiento de un individuo (o una unidad) no incluido en la muestra.

No obstante, para realizar las predicciones, hemos de suponer que todas las hipótesis que hemos formulado sobre **X**, **B** y **U** se mantendrán también para las observaciones fuera de la muestra. En particular, es fundamental suponer que se cumple la hipótesis de permanencia estructural del modelo.

Cuando realizamos predicciones, podemos optar por predecir el valor puntual que tomará la variable endógena, o bien, determinar un intervalo de posibles valores. El primer caso se denomina predicción puntual, y el segundo predicción por intervalo.

Predicción puntual

Supongamos que la variable endógena ajustada para una determinada observación *i* es igual a:

$$\hat{Y}_i = \beta_1 + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki}.$$

Si queremos predecir el valor de la variable endógena, para una observación *n + h*, podemos utilizar la siguiente expresión:

$$\hat{Y}_{n+h} = \beta_1 + \beta_2 \cdot X_{2, n+h} + \dots + \beta_k \cdot X_{k, n+h}.$$

Predicción por intervalo

La fiabilidad de predicción se caracteriza por el intervalo de predicción. Distinguimos entre la predicción por intervalo sobre Y_{n+h} y la predicción por intervalo sobre su valor esperado, $E(Y_{n+h})$.

En primer lugar, para obtener el intervalo del valor esperado de la variable endógena para la observación *n+h*, $E(Y_{n+h})$, utilizaremos la siguiente expresión:

$$\Pr ob \left\{ \left| E(Y_{n+h}) - \hat{Y}_{n+h} \right| < t_{n-k, \frac{\alpha}{2}} \cdot \left[\sigma_u^2 \cdot X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}} \right\} = 1 - \alpha,$$

donde $t_{n-k, \alpha/2}$ es el valor de las tablas de una t de Student de $n-k$ grados de libertad. La expresión sirve para indicar que la probabilidad de que $E(Y_{N+h})$ quede dentro del intervalo de

$$\hat{Y}_{n+h} - t_{n-k, \frac{\alpha}{2}} \cdot \left[\sigma_u^2 \cdot X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}}$$

a

$$\hat{Y}_{n+h} + t_{n-k, \frac{\alpha}{2}} \cdot \left[\sigma_u^2 \cdot X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}}$$

es $(1 - \alpha)$, siendo α el nivel de significación.

Nota: el intervalo de predicción para $E(Y_{N+h})$ coincide con el intervalo de confianza. Es decir, el intervalo de predicción del valor esperado no es sino el intervalo de confianza del parámetro $X_{n+h}'B$.

En segundo lugar, para obtener la predicción por intervalo del valor observado de la variable endógena para la observación $n + h$, Y_{n+h} , utilizaremos la siguiente expresión:

$$\Pr ob \left\{ \left| Y_{n+h} - \hat{Y}_{n+h} \right| < t_{n-k, \frac{\alpha}{2}} \cdot \sigma_u \cdot \left[1 + X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}} \right\} = 1 - \alpha$$

De forma análoga al caso anterior, la expresión indica que la probabilidad de que Y_{n+h} se encuentre dentro del intervalo de

$$\hat{Y}_{n+h} - t_{n-k, \frac{\alpha}{2}} \cdot \sigma_u \cdot \left[1 + X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}}$$

a

$$\hat{Y}_{n+h} + t_{n-k, \frac{\alpha}{2}} \cdot \sigma_u \cdot \left[1 + X_{n+h} \cdot (X' \cdot X)^{-1} \cdot X_{n+h} \right]^{\frac{1}{2}}$$

es $(1 - \alpha)$, siendo α el nivel de significación.

Nota: A la hora de realizar las predicciones, se puede ver que el intervalo de predicción para el valor observado de la variable endógena resulta más grande que el intervalo de predicción para el valor esperado de la variable endógena. El caso es que, al predecir $E(Y_{n+h})$, pretendemos prever sólo componente explicada por X_{n+h} , y la componente puramente aleatoria, u_{n+h} , no forma parte del objetivo de predicción. En cambio, cuando el objetivo es predecir Y_{n+h} , hemos de prever también la perturbación aleatoria u_{n+h} la cual incrementa la varianza del término de error.

CASOS PRÁCTICOS CON SOFTWARE

Estimación MCO del modelo de regresión lineal

Ejemplo 1. Representación gráfica del ajuste de MCO. A efectos de una mejor comprensión del método de estimación de MCO, realizaremos la representación gráfica del ajuste de MCO. Consideremos un modelo de regresión lineal simple:

$$Y_i = \beta_1 + \beta_2 \cdot X_i + u_i \quad i = 1, \dots, n$$

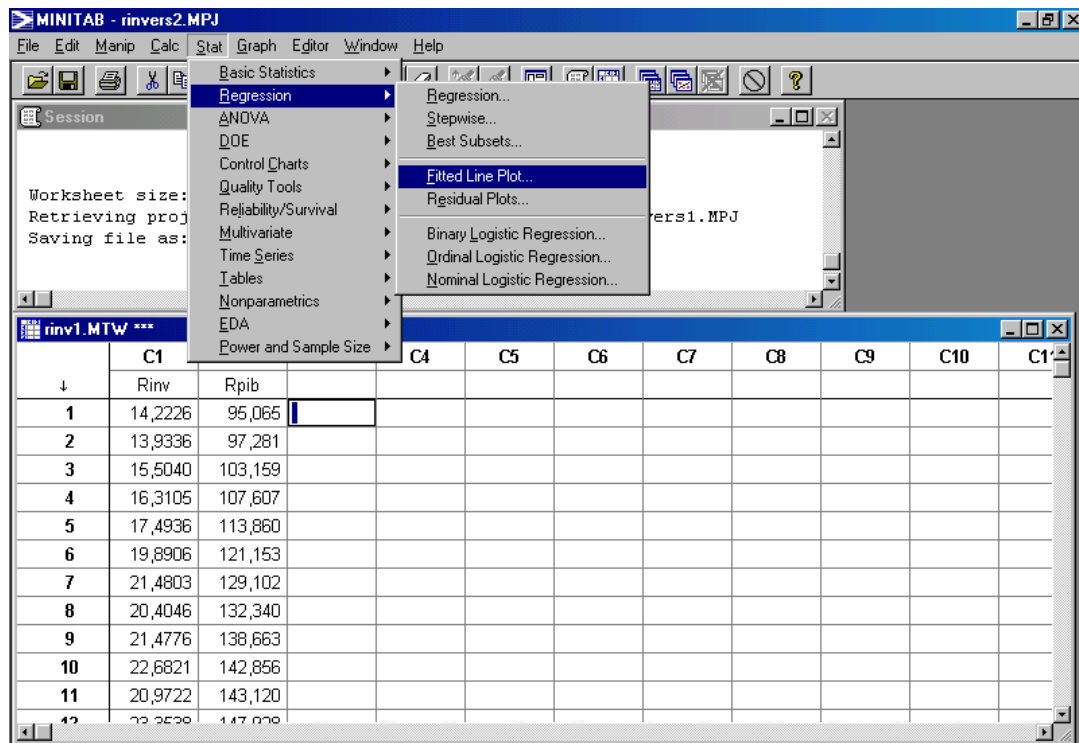
Como ya hemos dicho, nuestro objetivo es asignar valores numéricos a los parámetros desconocidos, en este caso, β_1 y β_2 , y así poder cuantificar la relación de dependencia que hay entre las dos variables. Determinar estos valores equivale a determinar una recta que pasa por la nube de puntos que resultan al representar las observaciones correspondientes a las variables endógena y explicativa.

Consideremos los siguientes datos anuales correspondientes al período 1960-1990 de la economía de los Estados Unidos:

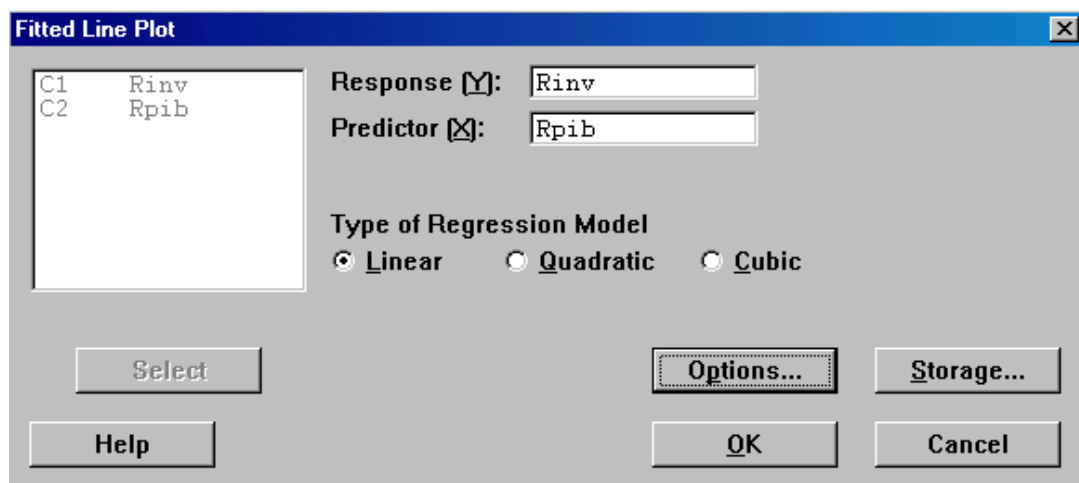
Observación	Año	Y (inversión real)	X (PIB real)
1	1960	14,2226	95,065
2	1961	13,9336	97,281
3	1962	15,5040	103,159
4	1963	16,3105	107,607
5	1964	17,4936	113,860
6	1965	19,8906	121,153
7	1966	21,4803	129,102
8	1967	20,4046	132,340
9	1968	21,4776	138,663
10	1969	22,6821	142,856
11	1970	20,9722	143,120
12	1971	23,3538	147,928
13	1972	26,1040	155,955
14	1973	29,1101	164,946
15	1974	27,2418	163,921
16	1975	23,0096	163,426
17	1976	27,6116	172,485
18	1977	32,1111	180,519
19	1978	36,1788	190,509
20	1979	37,5671	196,497
21	1980	33,5069	196,024
22	1981	36,6088	200,832
23	1982	31,1554	196,769
24	1983	32,7752	205,341
25	1984	41,1886	220,230
26	1985	39,9715	228,703
27	1986	39,6866	236,500
28	1987	40,2991	244,560
29	1988	40,9538	254,771
30	1989	41,9323	263,683
31	1990	39,8393	268,304

Estos datos en el espacio bidimensional constituyen una nube de puntos, para los cuales trazaremos la recta de regresión caracterizada por el mejor ajuste. Para ello, seguiremos los siguientes pasos en el entorno de Minitab:

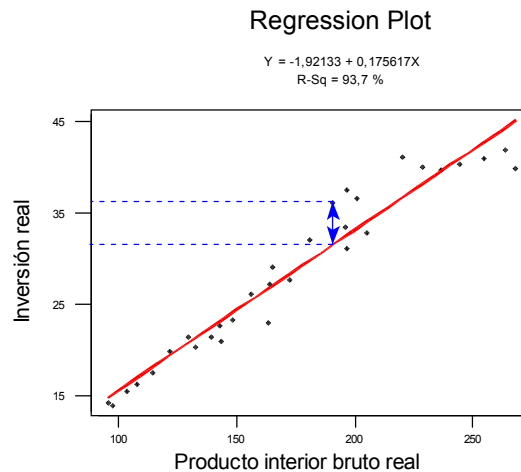
Seleccionamos **Stat > Regression > Fitted Line Plot...** :



A continuación completamos los campos según se indica:



La operación nos proporciona el siguiente gráfico:



La recta en rojo es la que mejor se ajusta, según el criterio de MCO, a la nube de puntos que tenemos. Es decir, es la recta que hace que el error de estimación, definido como la distancia entre el valor observado y el valor estimado de la variable endógena (en el gráfico, es la distancia vertical señalada por la flecha en azul), sea la mínima para cada una de las observaciones. La pendiente de la recta presenta signo positivo, pues es de esperar que el un auge en el PIB genere una mayor cantidad de inversiones y viceversa. Encima de la recta, se aparece la ecuación de MCO con el coeficiente de determinación, R^2 . Podemos apreciar que el modelo se ajusta buenamente a los datos, explicando un 93,7% de la variabilidad de la variable endógena. En consecuencia, el estadístico de significación global del modelo se calcula de la siguiente manera:

$$F_0 = (R^2 / (1 - R^2)) \cdot (n - k) / (n - 1) = (0,937 / 0,063) \cdot 29 / 30 = 14,377$$

Sabemos que en el modelo de regresión lineal simple se cumple que $F_0 = t_2^2$, siendo t_2 el estadístico de contraste de significación individual. De modo que $t_2 = \sqrt{F_0} = 3,792$.

Para contrastar la significación individual de la variable explicativa, a partir de las tablas extraemos $t_{n-k, \alpha/2} = t_{29; 0,025} = 2,0452$. Dado que $t_2 = 3,792 > t_{29; 0,025} = 2,0452$, rechazamos la hipótesis nula. En conclusión, el PIB real es individualmente significativo para explicar la variabilidad de la inversión real en la economía de los Estados Unidos.

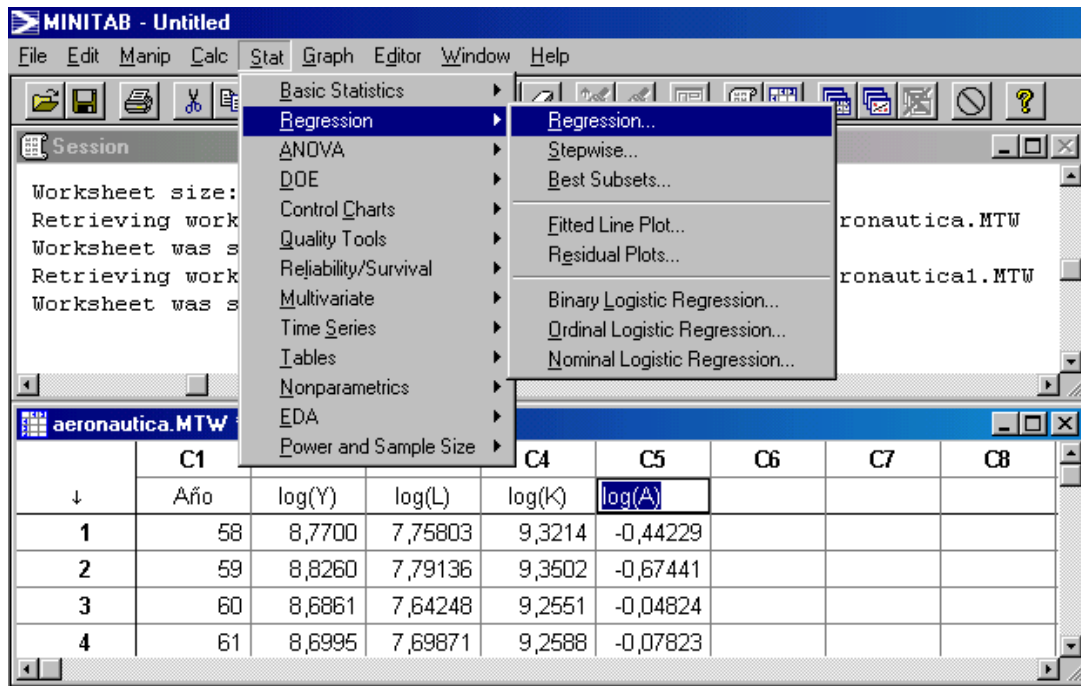
Ejemplo 2: Una empresa de investigación de mercados está interesada en realizar un estudio para el gobierno sobre la industria aeronáutica de los Estados Unidos. Para ello, va a estimar la función de producción Cobb-Douglas estocástica aumentada por la variable el avance tecnológico:

$$\log(Y_t) = \beta_1 \cdot \log(L_t) + \log(K_t) + \log(A_t) + u_t ; t = 1, \dots, T$$

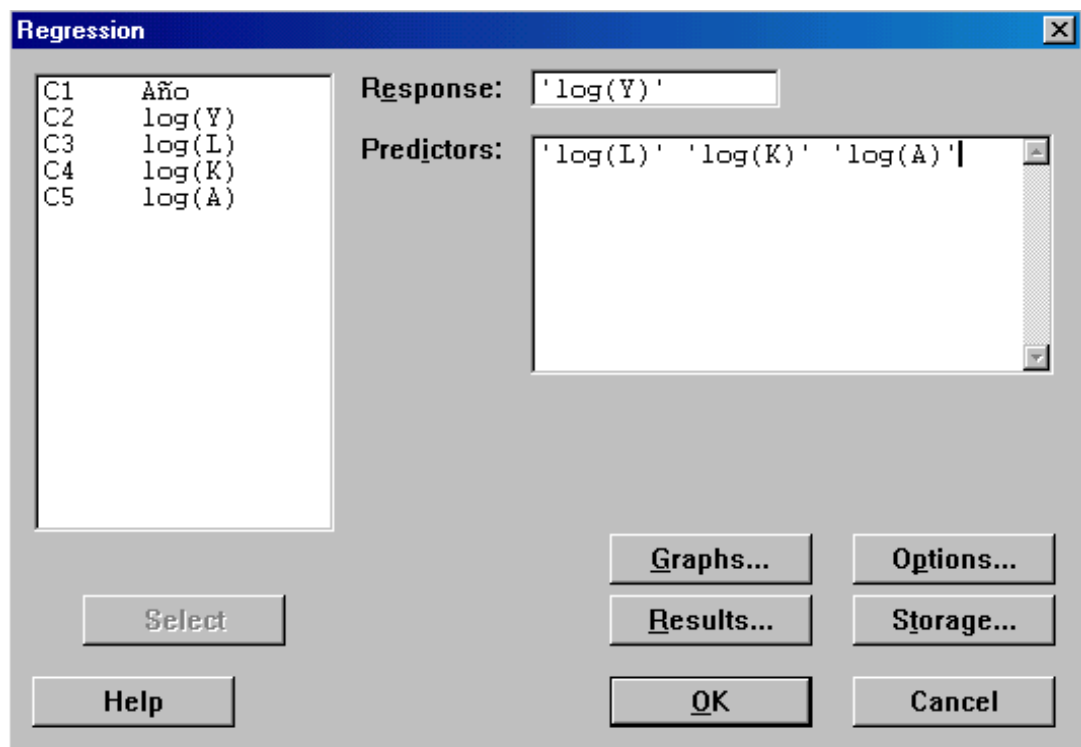
donde Y_t es la producción (en millones de dólares), L_t es el nivel de empleo (que representaremos a través del agregado de las nóminas (en millones de dólares), K_t es el nivel de capital utilizado (en millones de dólares), y A_t es el avance tecnológico, representado por la proporción del PIB de las empresas tecnológicas en el PIB total en la economía americana (en tanto por ciento). Supondremos que esta relación satisface las hipótesis el MRLM con normalidad del término de error. Se dispone de datos anuales correspondientes a 1958-1996 que se muestran en la siguiente tabla:

Observación	Año	log(Y)	Log(L)	Log(K)	Log(A)
1	1958	8,7700	7,75803	9,3214	-0,44229
2	1959	8,8260	7,79136	9,3502	-0,67441
3	1960	8,6861	7,64248	9,2551	-0,04824
4	1961	8,6995	7,69871	9,2588	-0,07823
5	1962	8,7332	7,81145	9,2779	0,02132
6	1963	8,7509	7,77039	9,2977	0,06255
7	1964	8,7924	7,75307	9,3311	0,23289
8	1965	8,8750	7,82740	9,3657	0,43465
9	1966	9,1050	8,07770	9,5809	0,60064
10	1967	9,3129	8,18004	9,8358	0,77948
11	1968	9,4738	8,27055	9,9564	0,84076
12	1969	9,4291	8,31059	10,0004	1,00189
13	1970	9,3468	8,15047	9,9534	1,04609
14	1971	9,2124	7,91517	9,8486	0,95128
15	1972	9,0802	7,96106	9,8342	0,97795
16	1973	9,2748	8,02597	9,8140	1,19855
17	1974	9,3644	8,10119	9,8716	1,37927
18	1975	9,4094	8,14297	9,9271	1,21982
19	1976	9,5044	8,17836	9,9131	1,50437
20	1977	9,6047	8,28801	9,9559	1,71540
21	1978	9,7440	8,46720	10,1037	1,92360
22	1979	10,0222	8,65232	10,3419	2,16460
23	1980	10,1955	8,80499	10,5113	2,26792
24	1981	10,3034	8,98153	10,6039	2,42746
25	1982	10,2417	8,95546	10,7125	2,49750
26	1983	10,3262	8,93089	10,6632	2,47373
27	1984	10,2560	8,91690	10,7302	2,61771
28	1985	10,4624	8,98805	10,7732	2,44101
29	1986	10,5502	9,10319	10,8743	2,53751
30	1987	10,5737	9,17777	10,9206	2,85079
31	1988	10,6333	9,21186	11,0444	2,82018
32	1989	10,6768	9,25614	11,1949	2,82289
33	1990	10,8468	9,32587	11,2812	2,72615
34	1991	10,9698	9,24224	11,3309	2,54905
35	1992	11,0506	9,35001	11,3281	2,55048
36	1993	10,9173	9,28638	11,2780	2,50060
37	1994	10,8390	9,24362	11,1210	2,62398
38	1995	10,7585	9,12033	11,0568	2,77913
39	1996	10,7645	9,19414	11,1375	2,79638

La primera etapa del estudio consiste en estimar el modelo por MCO mediante el Minitab. Para ello, seleccionamos **Stat > Regression > Regression :**



A continuación completamos los campos según se indica:



Los resultados de estimación se muestran en el siguiente cuadro:

Regression Analysis
The regression equation is $\log(Y) = - 1,17 + 0,559 \log(L) + 0,601 \log(K) + 0,0329 \log(A)$

Predictor	Coef	StDev	T	P
Constant	-1,1666	0,4613	-2,53	0,016
log(L)	0,5585	0,1237	4,51	0,000
log(K)	0,6014	0,1018	5,91	0,000
log(A)	0,03291	0,03229	1,02	0,315

S = 0,06750 R-Sq = 99,3% R-Sq(adj) = 99,3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	23,5977	7,8659	1726,58	0,000
Residual Error	35	0,1595	0,0046		
Total	38	23,7572			

A partir de la salida de estimación por MCO, el vector de parámetros estimados, B, resulta ser $B = (-1,17; 0,559; 0,60; 0,033)$. Los signos que presentan los parámetros asociados a las variables explicativas son positivos y, por tanto, eran esperables. Puesto que la función de regresión es una transformación logarítmica de la función de producción de Cobb-Douglas, los parámetros miden las elasticidades de la producción respecto al empleo, al capital y al avance tecnológico respectivamente:

$$e_{Y,L} = \beta_2 = \frac{\partial \log(Y)}{\partial \log(L)};$$

$$e_{Y,K} = \beta_3 = \frac{\partial \log(Y)}{\partial \log(K)};$$

$$e_{Y,A} = \beta_4 = \frac{\partial \log(Y)}{\partial \log(A)}.$$

Una vez estimado el modelo, procedemos a analizar la validez estadística del modelo. Por ejemplo, para contrastar la significación individual de la variable $\log(A)$, especificamos la hipótesis nula $H_0: \beta_4 = 0$ frente a la hipótesis alternativa bilateral $H_A: \beta_4 \neq 0$. El contraste de hipótesis realizaremos en base del estadístico de contraste t y el p-valor asociado. Suponiendo cierta la hipótesis nula, el estadístico de contraste se calcula $t_4 = B_4/SE(B_4)$, siendo $SE(B_4)$ la desviación típica del estimador B_4 . A partir de los resultados de estimación, tenemos que $t_4 = 1,02$ con p-valor = 0,315. Recordemos que p-valor = $\text{Prob}(t > t_4 = 1,02)$. Como p-valor = 0,315 > $\alpha = 0,05$, no podemos rechazar la hipótesis nula para el nivel de significación de 5%. También, haciendo el uso del valor crítico $t_{n-k;\alpha/2} = t_{35;0,025} = 2,0301$ a partir de las tablas de una distribución t-Student, queda $t_4 = 1,02 \notin (-2,0301; 2,0301)$ lo cual nos conduce a la misma conclusión. Por tanto, la variable *el avance tecnológico* resulta estadísticamente no significativa. La evidencia empírica parece indicar que el desarrollo tecnológico no ha sido decisivo para la industria aeronáutica. En cambio, los resultados de los contrastes de significación individual de $\log(L)$ y de $\log(K)$ nos llevan a rechazar la hipótesis nula; concluimos, por tanto, que tanto el capital humano como el capital físico son significativos a la hora de explicar la variación de la producción en el sector aeronáutico.

Una vez analizada la relevancia individual de las variables explicativas, pasamos a contrastar la significación conjunta del modelo. Utilizando el estadístico F_0 a partir del cuadro de estimación y comparándolo con el valor crítico $F_{k-1;n-k;\alpha}$ a partir de las tablas de una distribución F de Snedecor queda:

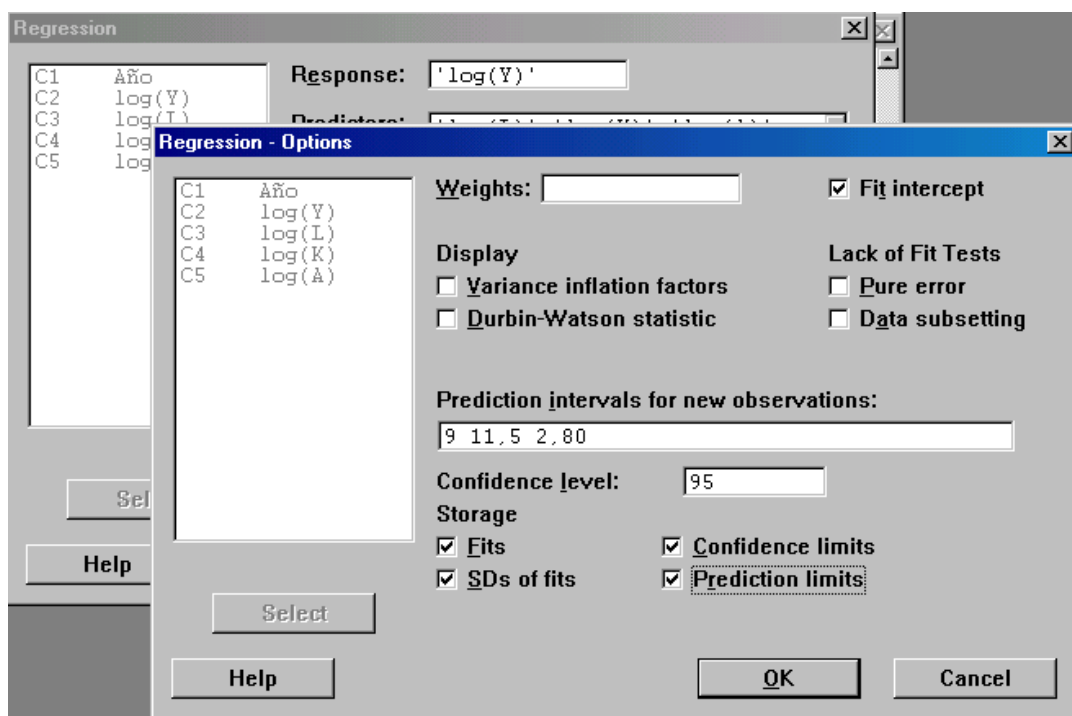
$$F_0 = 1726,58 > F_{3;35;0,05} = 2,8742.$$

Puesto que el estadístico de contraste muestral es muy superior al valor crítico a partir de las tablas, rechazamos la hipótesis nula de no significación global del modelo.

A continuación, a base del modelo estimado, pasaremos a realizar la predicción, tanto del valor esperado como del valor observado, de la variable endógena para el año 1997, teniendo en cuenta la siguiente información sobre las variables explicativas para el año 1997:

$$\log(L_{97}) = 9,00; \log(K_{97}) = 11,50 \text{ y } \log(A_{97}) = 2,80.$$

Volvemos a seleccionar **Stat > Regression > Regresión** y completamos los campos en la ventana "Regresión" tal y como hemos hecho para estimar el modelo de regresión. A continuación, dentro de la misma ventana seleccionamos **Options** e introducimos los valores de predictores, especificando el 95% nivel de confianza. Por último, marcamos las opciones "Fits", "SDs of fits", "Confidence limits" y "Prediction limits" para mostrar el ajuste de predicción, la desviación típica de predicción, los intervalos de confianza y los intervalos de predicción, respectivamente:



Los resultados de predicción aparecen en el siguiente cuadro:

Predicted Values				
Fit	StDev Fit	95,0% CI	95,0% PI	
10,8678	0,0677	(10,7304; 11,0052)	(10,6738; 11,0619) XX	
X denotes a row with X values away from the center				
XX denotes a row with very extreme X values				

Los resultados indican que la predicción de la producción en el sector aeronáutico (predicción puntual) es:

$$\log(Y_{97}) = -1,17 + 0,559 \log(L_{97}) + 0,601 \log(K_{97}) + 0,0329 \log(A_{97}) = -1,17 + 0,559 \cdot 9,0 + 0,601 \cdot 11,5 + 0,0329 \cdot 2,80 = 10,865.$$

Observad que la predicción realizada es de una transformación logarítmica; no obstante, nuestro interés reside en la predicción de la producción en niveles. A tales efectos, calculamos la exponencial del resultado anterior:

$$Y_{97} = \exp(\log(Y_{97})) = \exp(10,865) = 52.293 \text{ millones de dólares.}$$

El intervalo de predicción del valor esperado de la variable endógena en el programa Minitab coincide con el intervalo de confianza para el parámetro $X_{n+h}'B$:

$$IP(E(Y_{n+h})) = IC(X_{n+h}'B) = \{10,7304; 11,0052\}.$$

Finalmente, el intervalo de predicción sobre el valor observado de la variable endógena es:

$$IP(Y_{N+h}) = \{10,6738; 11,0619\}.$$

En efecto, el intervalo de predicción del valor observado de la variable endógena es más grande que el intervalo de predicción para el valor esperado de la variable endógena.

BIBLIOGRAFÍA

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

ENLACES

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