Tema 1: Regresión con variable dependiente continua

Miguel Jerez

Universidad Complutense de Madrid

Septiembre 2017

Índice

- Introducción
- El modelo de regresión
- Estimación
- Inferencia
- Estimación restringida e inferencia

Introducción (I): ¿Por qué modelizar?

- La econometría se centra en la construcción de modelos estadísticos para datos económicos
- La modelización consiste en representar un conjunto de datos ("la muestra") mediante una ecuación matemática ("el modelo")
- Los modelos se construyen porque son más útiles que los datos en bruto para satisfacer una necesidad específica. Por ello, es fundamental que el modelo:
  - …sea mucho más simple que los datos que representa, siempre que
  - …capte las características de datos necesarias para cada aplicación
- Los datos económicos tienen características particulares:
  - Están afectados por errores de observación o muestreo
  - No son el resultado de un experimento, por lo que no pueden muestrearse en condiciones ambientales controladas,
  - A menudo están expresados en unidades monetarias, cuyo valor real cambia con el tiempo

Introducción (II): Objetivos de la modelización

La econometría construye modelos de la actividad económica con distintos objetivos:

- Estimar los parámetros desconocidos de un modelo, como la sensibilidad del valor de una cartera de activos a las fluctuaciones en un índice de referencia
- Estimar valores que no están en la muestra:
  - Previsión: los sistemas de *credit scoring* estiman la probabilidad de que un préstamo resulte impagado en un horizonte temporal dado, en función de las características financieras y personales del solicitante
  - *Nowcasting*: los institutos de estadística de producen estimaciones del PIB en el último trimestre, aunque la cifra anual definitiva tarda varios años en calcularse
  - *Backcasting*: los historiadores estiman series de PIB e inflación en siglos pasados, utilizando datos fragmentarios de muchas fuentes
- Extraer señales enterradas en los datos, como tendencias, ciclos o estacionalidad
- Simular el comportamiento de un sistema: por ejemplo, el comportamiento del balance de un banco en escenarios de estrés, para estimar necesidades de capital
- Controlar el comportamiento de un sistema: los bancos centrales estiman qué tipos de intervención llevará la inflación esperada al objetivo que se quiere conseguir

Introducción (III): Tipología de muestras y variables

Las muestras contienen un conjunto de observaciones de las variables de interés y pueden ser de sección cruzada, series temporales o paneles de datos

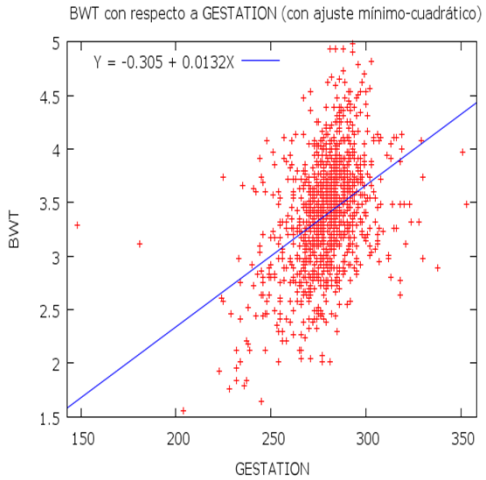
		Dimensión temporal	
		La muestra está referida a un solo período temporal	**Las variables de la muestra evolucionan en el tiempo**
Dimensión espacial	**La muestra cambia para distintas "unidades" (países, empresas, familias, individuos, …)**	"**Sección cruzada**" (o "**Corte transversal**") **Ejemplo**: Gasto en sanidad y PIB de distintos países	**Panel** **Ejemplo**: Gasto en sanidad y PIB de distintos países en varios años
	La muestra está referida a una sola unidad	**No es una muestra**, sino un solo dato	**Serie temporal** **Ejemplo**: Gasto en sanidad y PIB de un país en varios años

- Las variables muestreadas pueden ser continuas en la recta real o discretas
- Las variables discretas pueden clasificarse en dicotómicas ("binarias", "0-1" o "dummys") como, por ejemplo, se fumador o no
- …y policotómicas como, por ejemplo, el número de habitaciones de una casa

Introducción (IV): Ejemplos de distintos tipos de muestras

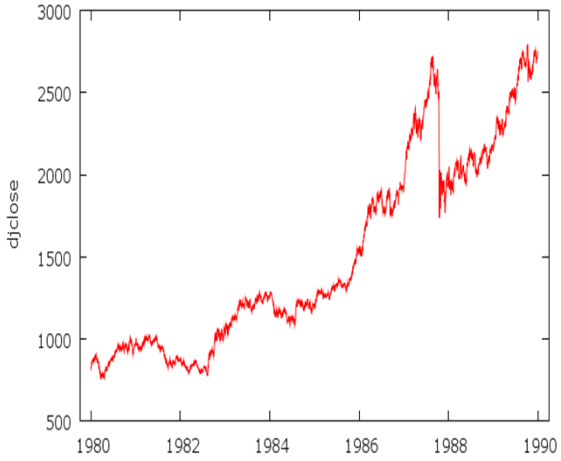
Peso en kilos del neonato frente a número de días de gestación

Sección cruzada de variables continuas



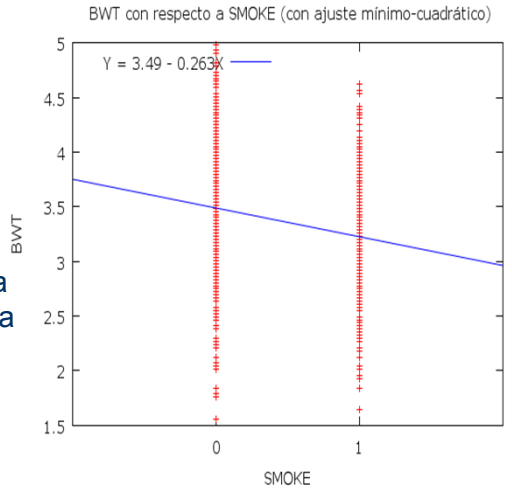
Valor al cierre del índice Dow Jones

Serie temporal continua



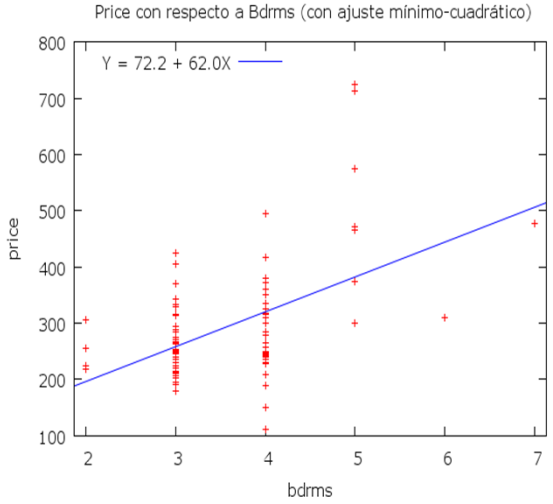
Peso en kilos del neonato frente a madre fumadora (SMOKE=1) y no-fumadora (SMOKE=0)

Sección cruzada, una variable continua, otra dicotómica



Precio en miles de US$ de la vivienda frente a número de dormitorios

Sección cruzada, una variable continua, otra policotómica



Índice

- Introducción
- El modelo de regresión
- Estimación
- Inferencia
- Estimación restringida e inferencia

El modelo de regresión (I): Regresión simple y múltiple

El modelo de regresión simple se define como:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (t = 1, 2, \ldots, n)$$

en donde:

y_t : variable endógena o dependiente,

x_t : variable exógena, explicativa o regresor,

β_i : *i*-ésimo parámetro o coeficiente, β_0 es conocido como "término constante" y β_1 como "pendiente"

ε_t : término de error o perturbación

t : subíndice que denota el caso *t*-ésimo de una muestra de tamaño *n*

Si hay más de una variable explicativa el modelo se denomina de regresión múltiple o Modelo Lineal General (MLG):

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} \ldots + \beta_k x_{tk} + \varepsilon_t \quad (t = 1, 2, \ldots, n) \qquad (1)$$

…en donde resulta inmediato añadir un término constante haciendo $x_{t1} = 1$

El modelo de regresión (II): El principio de mínimos cuadrados

Distinguiremos entre el "modelo verdadero" y el "modelo estimado":

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\varepsilon}_t$$

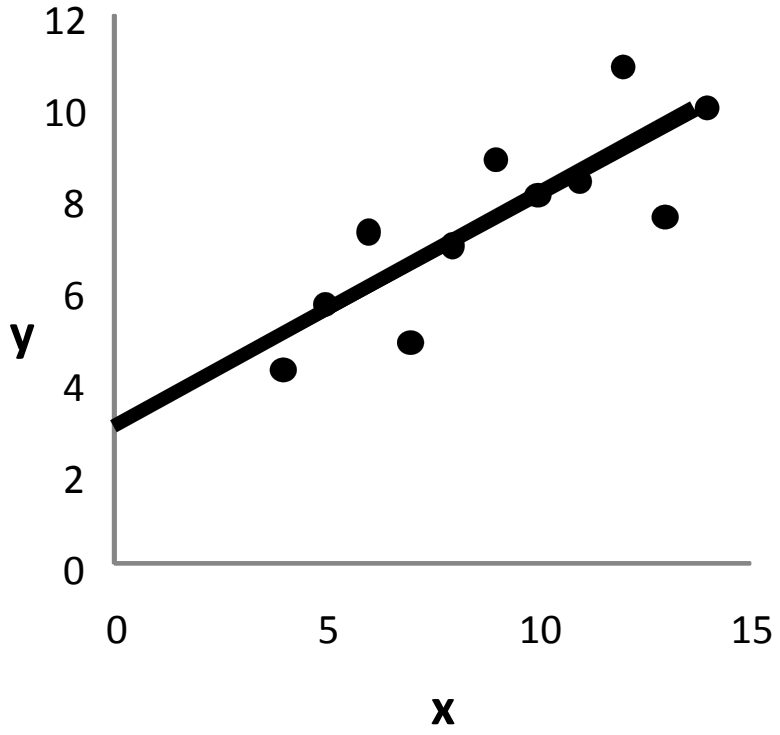
…y consideraremos los conceptos de "valores ajustados" ($\hat{y}_t$) y "residuos" ($\hat{\varepsilon}_t$)

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

$$\hat{\varepsilon}_t = y_t - \hat{y}_t$$

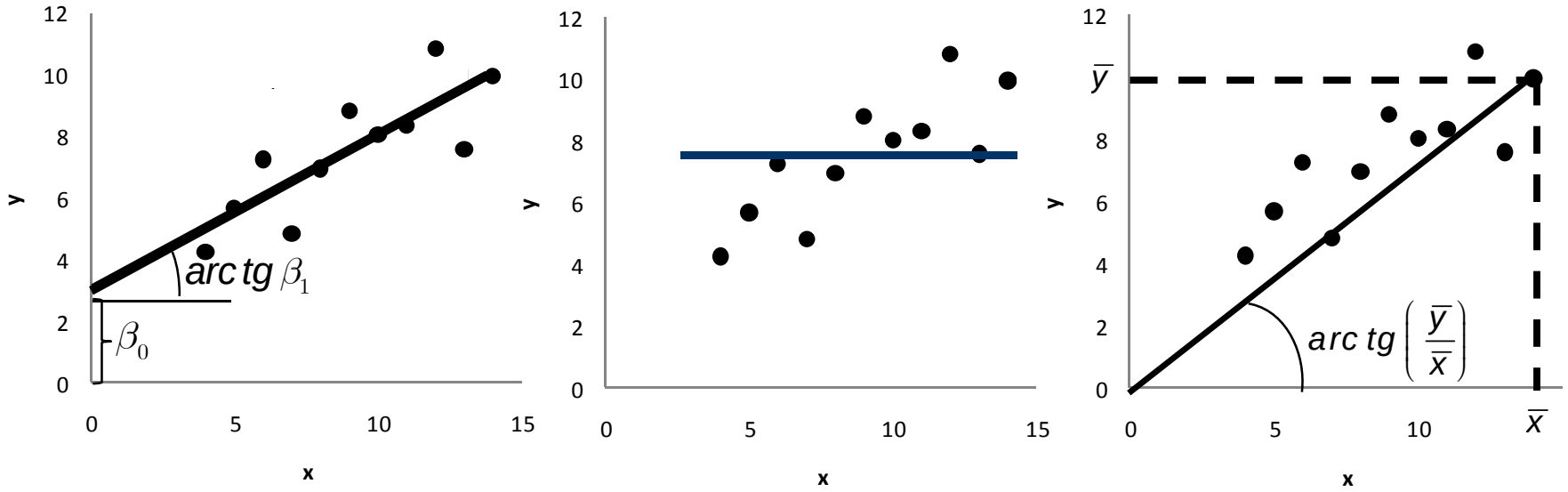
Los parámetros del modelo son desconocidos en general, por lo que necesitamos un criterio de estimación

A menudo usaremos el criterio de Mínimos Cuadrados Ordinarios (MCO) consistente en calcular el valor de los parámetros que minimiza la suma de residuos al cuadrado



- El símbolo "^" denota estimaciones
- Los valores ajustados por la regresión son expectativas de *y* condicionales al valor de *x*

El modelo de regresión (III)



Ajustar un modelo de regresión no es algo exótico: todo el mundo lo hace, consciente o inconscientemente

Calcular una media equivale a ajustar una regresión con término constante y sin variables explicativas

Calcular un ratio es lo mismo que ajustar una regresión sin término constante a un sólo dato

El modelo de regresión (IV.a): Interpretación de los coeficientes

Cuando las variables son continuas, los coeficientes de regresión son derivadas parciales de la variable endógena con respecto a las explicativas

Si las variables tienen algún tipo de transformación, esta interpretación general puede concretarse de varias formas. Por ejemplo:

Modelo	**Interpretación matemática y conceptual**	
$y_t = \beta x_t + \varepsilon_t$	$\beta = \frac{dy_t}{dx_t}$	Cambio esperado en y_t cuando x_t aumenta en una unidad
$\ln y_t = \beta \ln x_t + \varepsilon_t$	$\beta = \frac{d \ln y_t}{d \ln x_t} = \frac{x_t}{y_t}\frac{dy_t}{dx_t}$	Cambio porcentual esperado en y_t cuando x_t aumenta un uno por ciento
$\ln y_t \times 100 = \beta x_t + \varepsilon_t$	$\beta = \frac{1}{dx_t}\left(\frac{dy_t}{y_t} \times 100\right)$	Cambio porcentual esperado en y_t cuando x_t aumenta en una unidad
$y_t = \beta \ln x_t \times 100 + \varepsilon_t$	$\beta = dy_t \left(\frac{dx_t}{x_t} \times 100\right)^{-1}$	Cambio esperado en y_t cuando x_t aumenta un uno por ciento

El modelo de regresión (IV.b): Interpretación de los coeficientes

Cuando las variables explicativas son dicotómicas, los coeficientes de regresión son medias o cambios de media

En este caso podemos dar una interpretación válida para todos los casos: hay que tener en cuenta el diseño de las variables y la formulación del modelo.

Supongamos por ejemplo dos variables dicotómicas:

$$Hombre_t = \begin{cases} 1 \text{ si } t \text{ es hombre} \\ 0 \text{ si } t \text{ es mujer} \end{cases} ; \; Mujer_t = \begin{cases} 0 \text{ si } t \text{ es hombre} \\ 1 \text{ si } t \text{ es mujer} \end{cases}$$

Modelo	**Interpretación matemática y conceptual**	
$y_t = \beta + \varepsilon_t$	$\beta = E(y_t)$	Esperanza incondicional de la variable endógena
$y_t = \beta_H Hombre_t + \beta_M Mujer_t + \varepsilon_t$	$\beta_H = \mathrm{E}(y_t \mid Hombre_t = 1)$ $\beta_M = \mathrm{E}(y_t \mid Mujer_t = 1)$	Esperanza de la variable endógena condicionada al sexo del individuo
$y_t = \beta_H + \delta_M Mujer_t + \varepsilon_t$	$\beta_H = \mathrm{E}(y_t \mid Hombre_t = 1)$ $\beta_H + \delta_M = \mathrm{E}(y_t \mid Mujer_t = 1)$	Esperanza (o cambio en la esperanza) de la variable endógena condicionada al sexo del individuo

El modelo de regresión (V): Formulaciones matriciales

En notación vectorial, la expresión (1) puede escribirse como:

$$y_t = \boldsymbol{x}_t^T \beta + \varepsilon_t \quad (t = 1,2,\ldots,n)$$

en donde:

$\boldsymbol{x}_t^T$: vector (1x*k*) de observaciones de cada una de las *k* variables explicativas correspondientes al caso *t*-ésimo, y

β : vector (*k*x1) de parámetros

o, de forma más compacta, como:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \qquad (2)$$

en donde:

$\boldsymbol{y}$: vector (*n*x1) de observaciones de la variable endógena, y

$\boldsymbol{X}$: matriz (*n*xk) que recoge en cada fila los regresores correspondientes a cada valor de la variable endógena y en cada columna las observaciones de cada regresor

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} ; \ \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} ; \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

El modelo de regresión (VI): Cuestiones abiertas

El modelo de regresión plantea, inicialmente, tres problemas estadísticos:

- Estimación, que consiste en obtener una buena aproximación al valor de los parámetros a partir de una muestra de las variables
- Inferencia (o contraste de hipótesis), acerca del verdadero valor de los parámetros
- Previsión, *backcasting*, *nowcasting*, consiste en estimar valores no observados de la variable endógena a partir del modelo estimado y de los valores de los regresores

Para resolver estos problemas es necesario hacer una serie de hipótesis acerca del MLG. Esto da lugar un cuarto problema:

- Diagnosis que consiste en, detectar incumplimientos de las hipótesis, valorar su impacto sobre el análisis y, si es necesario, resolver el incumplimiento o mitigar sus efectos negativos

Índice

- Introducción
- Formulación del modelo
- Estimación
- Inferencia
- Estimación restringida e inferencia

Estimación (I): Hipótesis previas

[H.1] El modelo está correctamente especificado:

[H.1.1] La relación entre la variable endógena y las variables explicativas es lineal

[H.1.2] El modelo incluye todas las variables explicativas relevantes

[H.1.3] El modelo no incluye ninguna variable explicativa irrelevante

[H.2] Los parámetros del modelo son constantes

[H.3] Suficientes grados de libertad. El número de observaciones es al menos igual que el número de parámetros que se desea estimar

[H.4] Regresores deterministas (las variables explicativas no son aleatorias)

[H.5] Ausencia de colinealidad. No existen relaciones lineales exactas entre las variables explicativas o, equivalentemente, $\left|\boldsymbol{X}^T\boldsymbol{X}\right| \neq 0$

[H.6] Perturbaciones esféricas:

[H.6.1] Esperanza nula: $E(\varepsilon_t) = 0 \;\; (t = 1,2,\ldots,n)$

[H.6.2] Homoscedasticidad: $\text{var}(\varepsilon_t) = E(\varepsilon_t^2) = \sigma^2 \;\; (t = 1,2,\ldots,n)$

[H.6.3] Ausencia de autocorrelación: $\text{cov}(\varepsilon_t,\varepsilon_\tau) = E(\varepsilon_t\varepsilon_\tau) = 0 \;\; (t \neq \tau)$

[H.7] Normalidad. La distribución de probabilidad del término de error es normal

[H.6] y [H.7] a menudo se resumen como $\varepsilon \left| \boldsymbol{X} \sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}) \right.$ y, consecuentemente, $\boldsymbol{y} \left| \boldsymbol{X} \sim N(\boldsymbol{X}\beta,\sigma^2\boldsymbol{I}) \right.$

Estimación (II): El estimador MCO

A partir del MLG en forma vectorial [expresión (2)], y una estimación concreta de β, que denotaremos $\hat{\beta}$, definimos:

- el vector de valores ajustados o "previsiones intramuestrales", $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$
- el correspondiente vector de residuos como: $\hat{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$

Un posible criterio consiste en calcular el valor de $\hat{\beta}$ que minimiza la suma de los residuos al cuadrado. Este criterio se conoce como de mínimos cuadrados ordinarios (MCO).

Los resultados principales de estimación MCO son:

$$\hat{\beta}_{MCO} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (3)$$

$$cov(\hat{\beta}_{MCO}) = \sigma_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Para estimar la varianza del término de error puede usarse la expresión:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k}\hat{\varepsilon}^T\hat{\varepsilon} \qquad (4)$$

Que proporciona estimaciones insesgadas. Consecuentemente, la matriz de covarianzas del estimador MCO puede estimarse usando la expresión:

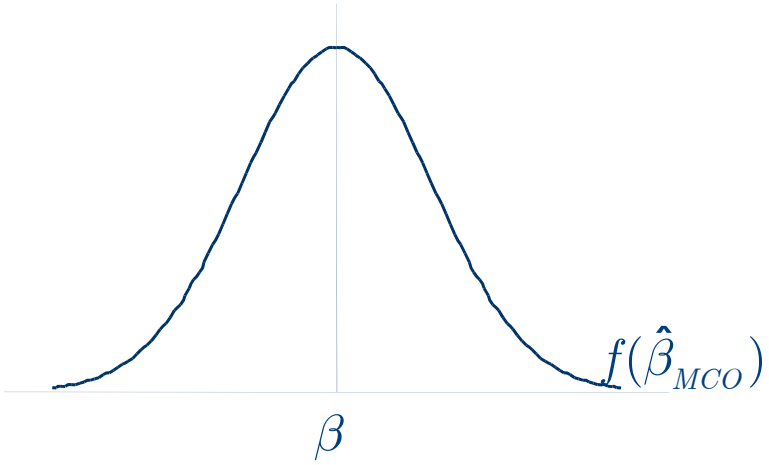
$$\hat{cov}(\hat{\beta}_{MCO}) = \hat{\sigma}_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \qquad (5)$$

Estimación (III.a): Propiedades de la estimación MCO

Bajo las hipótesis anteriores, el estimador MCO:

- Se distribuye como una normal
- Es insesgado, ya que esperanza coincide con el valor que se quiere estimar
- Si los errores son normales, es eficiente en el sentido máximo-verosímil, de forma que es el estimador insesgado más eficiente (no se exige linealidad).
- Incluso si los errores no son normales, es eficiente en el sentido de Gauss-Markov, ya que es más eficiente (tiene menos varianza) que cualquier estimador lineal e insesgado

Estas propiedades pueden presentarse visualmente en el siguiente gráfico:



Cada estimación de β es una muestra de la variable $\hat{\beta}_{MCO}$. La insesgadez significa que esta muestra probablemente saldrá del entorno del centro de la distribución, que coincide con el verdadero valor. La eficiencia implica que la distribución está más concentrada en torno a la media que la de cualquier otro estimador.

Además de todo lo anterior, el estimador MCO es consistente lo que quiere decir que, a medida que la muestra crece, las estimaciones tienden al verdadero valor

Estimación (III.a): Propiedades de la estimación MCO

Las propiedades del estimador MCO dependen de que las hipótesis se cumplan o no:

	Insesgadez	**Eficiencia**	**Consistencia**
Especificación correcta [H.1.1] y [H.1.2]	Requerida, con excepciones	Requerida	Requerida, con excepciones
Especificación correcta [H.1.3]	No requerida	Requerida	No Requerida
Parámetros constantes	Requerida		
Grados de libertad	Requerida para calcular las estimaciones MCO		
Regresores deterministas	La insesgadez y la eficiencia no se pueden probar		No requerida
Ausencia de colinealidad	No requerida		
Esperanza nula de los errores	Requerida	Requerida	No requerida
Homoscedasticidad	No requerida	Requerida	No requerida
Errores no autocorrelados	No requerida	Requerida	No requerida
Normalidad	No requerida	No requerida para Gauss-Markov	No requerida

Estimación (IV.a): Medidas de ajuste

Las medidas de ajuste sirven para:

- cuantificar la reducción de incertidumbre que proporciona el modelo y
- comparar modelos alternativos para la misma muestra.

La medida de ajuste más conocida es el coeficiente de determinación o R^2, que mide el porcentaje de la varianza de la variable dependiente que explica el modelo. Se define como:

$$R^2 = \frac{\left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_{\hat{y}}\right)^T \left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_{\hat{y}}\right)}{\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{y}\right)^T \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{y}\right)} = \frac{\hat{\text{var}}(\hat{\boldsymbol{y}})}{\hat{\text{var}}(\boldsymbol{y})} \tag{6}$$

o bien:

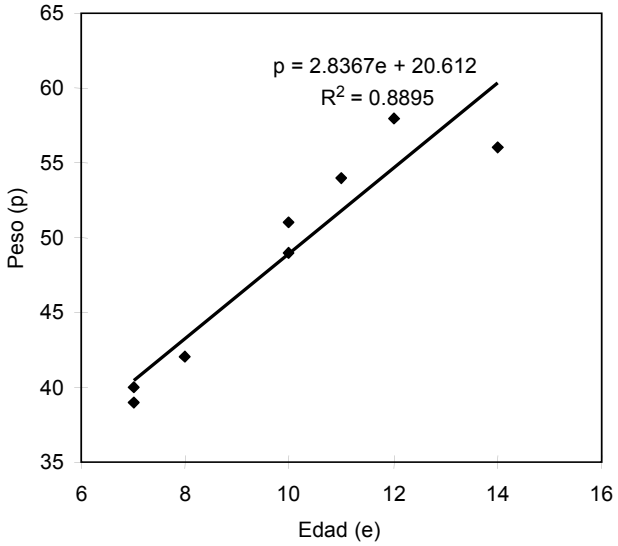
$$R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{y}\right)^T \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{y}\right)} = 1 - \frac{\hat{\text{var}}(\hat{\boldsymbol{\varepsilon}})}{\hat{\text{var}}(\boldsymbol{y})} \tag{7}$$

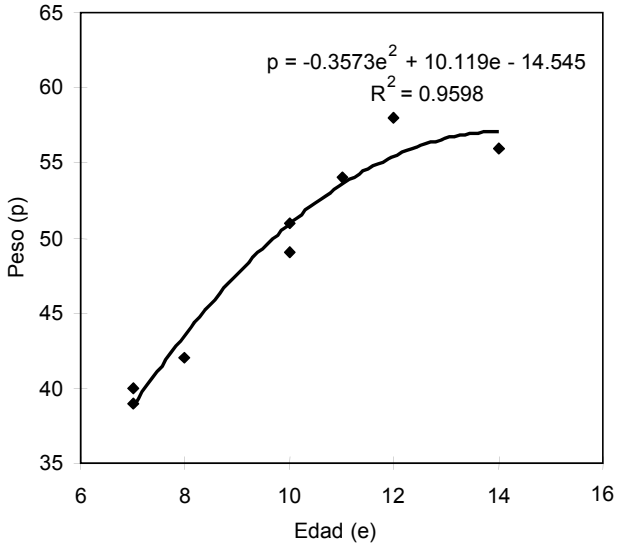
La expresión (7) sólo es válida si: (a) el modelo tiene un término constante, o (b) la variable endógena está expresada en desviaciones con respecto a su media. En estos casos el R^2 puede calcularse también como el cuadrado de la correlación muestral entre $\boldsymbol{y}$ e $\hat{\boldsymbol{y}}$.

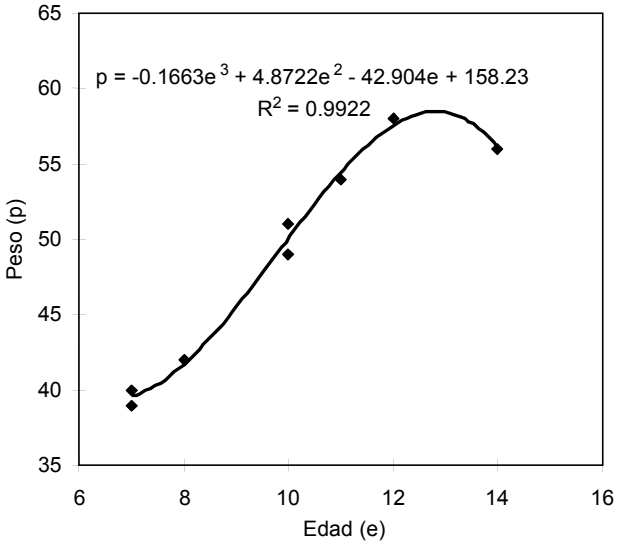
En principio, un modelo es tanto mejor cuanto mayor sea su correspondiente R^2 , ya que un valor alto supone que el modelo explica gran parte de la variabilidad de $\boldsymbol{y}$.

Estimación (IV.b): Ajuste y sobreajuste

Las siguientes figuras muestran el ajuste de tres modelos distintos a una muestra que recoge el peso y la edad de un grupo de ocho niños. Como puede verse, el mayor R^2 no siempre corresponde al mejor modelo.







Una regresión lineal explica el 89% de la varianza del peso

El modelo es imperfecto (¿qué peso predice para un niño de 0 años?)

Estas imperfecciones pueden deberse a: (a) limitaciones de la muestra y (b) no linealidad de la relación

Una regresión cuadrática mejora el R^2 (96%) y muestra un perfil intuitivamente razonable

A cambio, el nuevo modelo es más complejo, ya que requiere estimar tres parámetros en vez de dos

Una regresión cúbica proporciona un ajuste de más del 99%

Este es un buen resultado, siempre que estemos dispuestos a aceptar que los niños adelgazan a partir de los 13 años

Estimación (IV.c): Ajuste frente a complejidad

El uso mecánico del R^2 induce a sobreajustar la muestra. Para resolver este problema, a veces se usa un estadístico alternativo: el R^2 corregido con grados de libertad:

$$\bar{R}^2 = 1 - \frac{\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k}}{\frac{\left(\mathbf{y}-\hat{\mu}_y\right)^T\left(\mathbf{y}-\hat{\mu}_y\right)}{n-1}} = 1 - \frac{n-1}{n-k}(1-R^2) \qquad (8)$$

Este coeficiente: a) utiliza estimadores insesgados de la varianza residual y de la varianza de la variable dependiente y b) penaliza los modelos con muchos parámetros.

Actualmente disponemos de medidas más sofisticadas para comparar modelos, como por ejemplo los criterios de información de Akaike (AIC) y Schwartz (SBC)

$$AIC = n\ln(2\pi) + n\ln(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n}) + n + 2(k+1) \qquad (9)$$

$$SBC = n\ln(2\pi) + n\ln(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n}) + n + (k+1)\ln(n) \qquad (10)$$

Los criterios de información son útiles para comparar modelos alternativos para la misma variable endógena. Para usarlos, hay que tener en cuenta que:

- El criterio de decisión consiste en elegir el modelo que tenga un menor valor del criterio
- El AIC prima la capacidad predictiva del modelo y tiende a sobreparametrizarlo, el SBC prima la especificación correcta.

Índice

- Introducción
- Formulación del modelo
- Estimación
- Inferencia
- Estimación restringida e inferencia

Inferencia (I): Planteamiento general

- El contraste de hipótesis es una metodología de inferencia diseñada para valorar si una propiedad de la población es compatible con la información muestral
- Se denomina hipótesis nula, H_0, a la hipótesis que se desea contrastar frente a una hipótesis alternativa, H_1, que es la negación de la nula. El adjetivo «nula» indica que ésta es la hipótesis que se mantendrá como cierta a no ser que los datos indiquen su falsedad
- A partir de una muestra de la población estudiada, se calcula un estadístico (esto es, un valor que es función de la muestra) que debe:
  - ser una medida escalar de la distancia a la que la muestra se sitúa del cumplimiento exacto de la hipótesis nula,
  - … tener una distribución de probabilidad conocida si la hipótesis nula es cierta, y
  - … tener una distribución de probabilidad distinta (cuanto más distinta, mejor) si la nula es falsa
- Los resultados de inferencia que vamos a revisar son muy frágiles, ya que los únicos incumplimientos de hipótesis que admiten son presencia de variables irrelevantes y colinealidad (de grado)

Inferencia (II): Contrastes a partir del nivel de significación

Una estrategia de contraste consiste en fijar una región de rechazo, esto es, un conjunto de valores del estadístico que se consideran suficientemente raros como para rechazar H_0

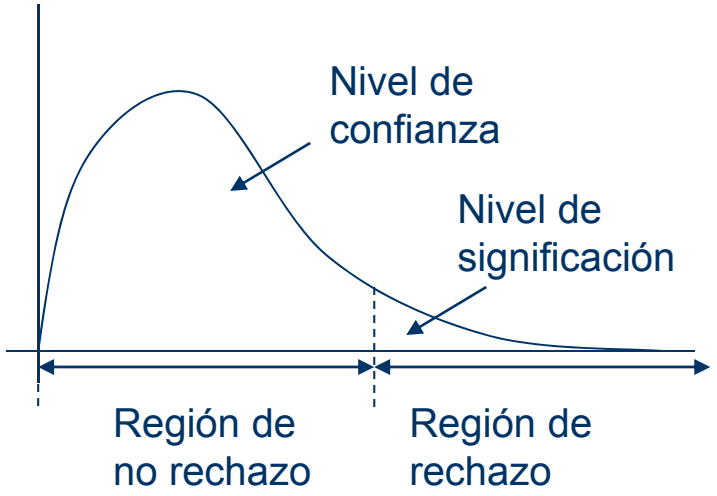
Si el estadístico cae ...

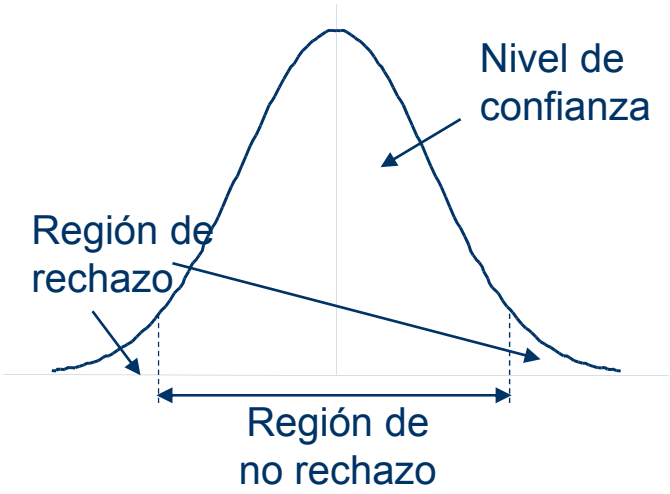
- en la región de rechazo, se rechaza la hipótesis nula en favor de la alternativa
- fuera de la región de rechazo, la hipótesis nula no se rechaza (no es lo mismo que aceptarla)

En Estadística se define que el "Error de tipo I" consiste en rechazar la hipótesis nula, siendo ésta cierta. El riesgo asociado a este error se conoce como "significación" y se mide mediante la probabilidad de que la distribución bajo la nula proporcione un valor del estadístico en la región de rechazo

Las regiones de rechazo pueden ser una o dos, dependiendo de si la distribución tiene una o dos colas

Los límites de rechazo deben determinarse teniendo en cuenta la forma de la distribución y la severidad del error de tipo I



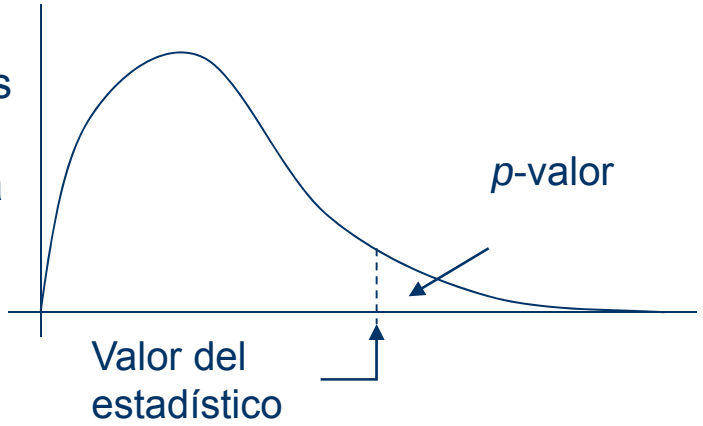


Inferencia (III): Contrastes a partir del *p*-valor

Otra estrategia consiste en medir el nivel de significación marginal (o *p*-valor) del estadístico

El *p*-valor es la probabilidad de que, siendo la hipótesis nula cierta, el estadístico de contraste tome un valor mayor o igual que el ha obtenido a partir de la muestra

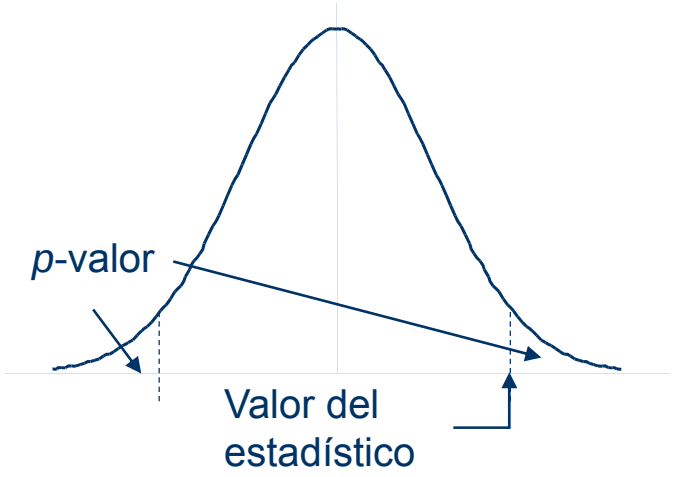
Puede interpretarse como: (a) una medida de la evidencia en favor de H_0, o (b) una estimación del riesgo de cometer el error de tipo I, dado el valor del estadístico



Si el *p*-valor es:

- Pequeño (p.ej. menor que el 5% o el 1%) entonces la muestra proporciona poca evidencia favorable a la nula y debe considerarse el rechazo
- Grande (p.ej. más del 5%) hay mucha evidencia a favor de la nula y debe considerarse el no-rechazo

Una vez más el límite de rechazo debe decidirlo el analista, teniendo en cuenta la severidad del error de tipo I



Inferencia (IV): Hipótesis simples en el modelo de regresión

Nos planteamos contrastar una sola hipótesis acerca del valor de los parámetros:

$$H_0 : \boldsymbol{a}^T \beta = c$$

en donde $\boldsymbol{a}$ es un vector k x1 de números reales y c es un escalar.

Puede demostrarse que, bajo las hipótesis realizadas:

$$t = \frac{\boldsymbol{a}^T \hat{\beta} - c}{\hat{\sigma}_\varepsilon \sqrt{\boldsymbol{a}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{a}}} \underset{H_0}{\sim} t_{n-k} \qquad (11)$$

Un contraste muy importante de este tipo es el de significación individual de un parámetro, en donde la hipótesis nula es $H_0 : \beta_i = 0$ y, por tanto, el estadístico de contraste se simplifica a:

$$t = \frac{\hat{\beta}_i}{s.\hat{d}.(\hat{\beta}_i)} \underset{H_0}{\sim} t_{n-k} \qquad (12)$$

Inferencia (V): Intervalos y regiones de confianza

Asimismo, se puede demostrar que:

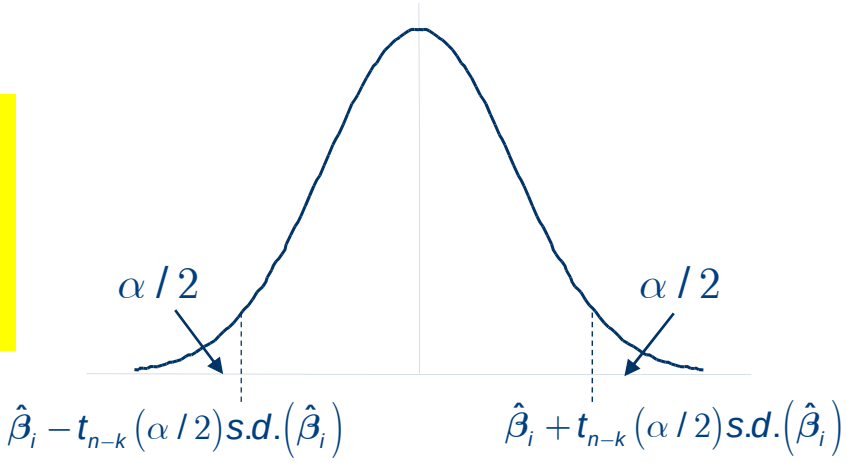
$$\frac{\hat{\beta}_i - \beta_i}{s.d.\left(\hat{\beta}_i\right)} \sim t_{n-k}$$

por tanto, fijando un nivel de significación arbitrario α, se obtiene un intervalo de confianza para el verdadero valor de un sólo parámetro:

$$P\left[\hat{\beta}_i - t_{n-k}(\alpha/2)s.d.\left(\hat{\beta}_i\right) \leq \beta_i \leq \hat{\beta}_i + t_{n-k}(\alpha/2)s.d.\left(\hat{\beta}_i\right)\right] = 1-\alpha$$

siendo $t_{n-k}(\alpha/2)$ el percentil 1- α/2 de una *t* de Student de *n-k* grados de libertad

La probabilidad de que el verdadero valor del parámetro se encuentre situado entre los límites superior e inferior del intervalo de confianza es 1- α



Índice

- Introducción
- Formulación del modelo
- Estimación
- Inferencia
- Estimación restringida e inferencia

Estimación restringida e inferencia (I): Planteamiento

A veces interesa imponer restricciones a los parámetros, porque:

- combinar información muestral y extramuestral mejora la precisión de las estimaciones, o
- porque se desea contrastar estas restricciones, comparando los resultados que se obtienen estimando el modelo libremente y con restricciones

Para estimar un modelo de regresión restringido, basta despejar algunos parámetros a partir de las restricciones, sustituirlos en el modelo y reordenar éste

Ejemplo. Sea el modelo:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t \quad (t = 1,2,\ldots,n)$$

Supongamos que se quiere imponer la restricción $\beta_2 = 1$. Esto puede hacerse sustituyendo la restricción en el modelo:

$$y_t = \beta_0 + \beta_1 x_t + z_t + \varepsilon_t$$

…y reordenando los términos para dejar en el lado derecho todos los sumandos que dependen de los parámetros desconocidos, de manera que el modelo resultante pueda estimarse por OLS:

$$y_t - z_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Estimación restringida e inferencia (II): Propiedades de MCRL

En comparación con MCO, las propiedades del estimador restringido (MCRL) son las siguientes:

1. Si las restricciones son ciertas y el estimador MCO es insesgado, el estimador MCRL también es insesgado
2. El estimador MCRL tiene menos incertidumbre que el MCO y, consecuentemente,
3. …si las restricciones son ciertas, el estimador MCRL es más eficiente que el MCO.

La propiedad previa no contradice el Teorema de Gauss-Markov, ya que MCRL utiliza más información que MCO

4. La suma de cuadrados de los residuos del estimador MCRL es mayor o igual que la del estimador MCO

A partir de esta última propiedad, la contrastación de hipótesis se reduce a estudiar si imponerlas como restricciones cambia significativamente el ajuste del modelo.

Estimación restringida e inferencia (III): Contrastes

Si la hipótesis nula es cierta, el contraste de hipótesis puede realizarse a partir del estadístico *F* habitual que, en este contexto de estimación restringida, puede calcularse de dos formas:

$$F = \frac{n-k}{m}\frac{SSR_{CLS} - SSR_{LS}}{SSR_{LS}} \underset{H_0}{\sim} F_{m,n-k} \qquad F = \frac{n-k}{m}\frac{R^2_{LS} - R^2_{CLS}}{1-R^2_{LS}} \underset{H_0}{\sim} F_{m,n-k}$$

…que coinciden con el valor de (2), quizá con pequeños errores de redondeo.

Ejemplo. En el modelo de regresión múltiple, la significación conjunta de todas las pendientes:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

…puede contrastarse particularizando el estadístico basado en el R^2 a m=k-1 y $R^2_{CLS} = 0$

$$F = \frac{n-k}{k-1}\frac{R^2_{LS}}{1-R^2_{LS}} \underset{H_0}{\sim} F_{k-1,n-k}$$

Miguel Jerez (mjerez@ccee.ucm.es)

Departamento de Fundamentos del Análisis Económico II
(Economía Cuantitativa)

Facultad de Ciencias Económicas, UCM