ECONOMETRIA I

M. Angeles Carnero

Departamento de Fundamentos del Análisis Económico
Despacho 36
acarnero@merlin.fae.ua.es
TUTORIAS: Martes: De 10 a 12:30 y de 15 a 17
Jueves: De 10 a 12

Curso 2011-12

¿Qué es la Econometría?

- Es la ciencia que tiene como objetivo analizar datos económicos mediante la utilización de herramientas estadísticas.
- El análisis de regresión nos permite describir e interpretar algunos aspectos de la realidad que nos rodea. Tratamos de analizar de qué forma se relacionan las variables económicas, determinar cuál es el efecto que una variación en una o varias variables puede provocar sobre otra variable.
- Resolver estas cuestiones ayuda a comprender mejor la realidad en que vivimos y puede ser de utilidad en:
  - la formulación de políticas públicas
  - la planificación estratégica de las empresas
  - la toma de decisiones individuales

¿Ejemplos?

1. ...
2. ...
3. ...

Ejemplos

1. ¿Es el salario de los trabajadores licenciados mayor que el de trabajadores con estudios primarios?
2. ¿Sufren las mujeres discriminación salarial?
3. ¿Es relevante la región de origen a la hora de explicar la edad del matrimonio?

Más ejemplos

- El Gobierno de un país se plantea el objetivo social de reducir la desigualdad entre sexos; para ello, pretende articular medidas que lleven a un aumento de la participación laboral femenina. Una posible medida es destinar recursos públicos a subvencionar guarderías infantiles. ¿Cuál es el efecto esperado de esta medida sobre la participación laboral de las mujeres?
- El Gobierno de un país desea reducir el número de contagios de SIDA entre sus ciudadanos y para ello decide incrementar en un 5 % los recursos públicos destinados a concienciación social. ¿Qué efecto tendrá esta medida sobre la tasa de contagio?
- El Gobierno de un país decide mejorar los resultados escolares de los niños y para ello destina más recursos públicos a la contratación de personal docente. ¿Realmente esta medida ayuda a mejorar los resultados escolares?

Más ejemplos

- Una empresa quiere saber qué efecto producirá sobre sus ventas un incremento de sus gastos en publicidad
- Una empresa desea conocer cuáles son los factores que determinan la satisfacción de sus trabajadores y cuál es la importancia de cada uno de ellos
- ¿El hecho de vivir en zona rural o en zona urbana es un factor relevante para explicar los salarios de los trabajadores?
- ¿Cómo afecta un impuesto sobre el tabaco a su consumo? ¿Es el tabaco un producto de demanda elástica o inelástica?

Los modelos de regresión nos permiten responder este tipo de cuestiones.

Especificación del modelo:

Objetivo: Cuantificar la relación que existe entre las siguientes variables

$$Y - \text{ variable endógena, dependiente}$$

$$X = (X_1, ..., X_k) - \text{ variable(s) exógena(s), explicativa(s), regresor(es)}$$

$$Y = f(X_1, ..., X_k)$$

a través de una **relación lineal**, usando una muestra de tamaño T y suponiendo una **relación aleatoria** (no exacta):

$$Y = f(X_1, ..., X_k) \overset{+\text{ error}}{\Rightarrow} Y_t = \underbrace{\beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt}}_{\text{parte determinista}} + \underbrace{u_t}_{\text{parte aleatoria}}$$

$$\text{para } t = 1, 2, ..., T$$

u_t : término aleatorio, término de error, parte inobservable

El modelo que se plantea es solamente una aproximación al verdadero modelo.

Razones para no tener una relación exacta:

- Aleatoriedad en el comportamiento humano
- Omisión de variables relevantes
- Errores de medida en las variables

Mismas razones por las que se incluye el término u. (Perturbación o innovación aleatoria)

Notación:

(a) $Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t, \quad t = 1, 2, ..., T$

(b) Notación vectorial: $Y_t = X_t'\beta + u_t$ donde $X_t' = \begin{bmatrix} 1 & X_{2t} & ... & X_{kt} \end{bmatrix}$

(c) Notación matricial: $Y = X\beta + u$, donde

$$\underset{(T\times 1)}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_T \end{bmatrix}, \underset{(T\times k)}{X} = \begin{bmatrix} 1 & X_{21} & ... & X_{k1} \\ 1 & X_{22} & ... & X_{k2} \\ ... & ... & ... & ... \\ 1 & X_{2T} & ... & X_{kT} \end{bmatrix},$$

$$\underset{(k\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}, \underset{(T\times 1)}{u} = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_T \end{bmatrix}$$

Es necesario que $T > k$ para que haya un número finito de soluciones.

(a) $$Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t, \quad t = 1, 2, ..., T$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_T \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 X_{21} + ... + \beta_k X_{k1} \\ \beta_1 + \beta_2 X_{22} + ... + \beta_k X_{k2} \\ ... \\ \beta_1 + \beta_2 X_{2T} + ... + \beta_k X_{kT} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_T \end{bmatrix} \rightarrow \text{(b)}$$

$$\underset{T\times 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_T \end{bmatrix}} = \underset{T\times k}{\begin{bmatrix} 1 & X_{21} & ... & X_{k1} \\ 1 & X_{22} & ... & X_{k2} \\ ... & ... & ... & ... \\ 1 & X_{2T} & ... & X_{kT} \end{bmatrix}} \underset{k\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}} + \underset{T\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_T \end{bmatrix}}$$

(c) $$Y = X \cdot \beta + u$$

Hipótesis básicas:

1. X es una matriz de constantes conocida. Esto es, la matriz de regresores no es aleatoria.
2. rango(X)=k, es decir, X tiene rango completo de columnas por lo que existe $(X'X)^{-1}$. Esto significa que no hay relaciones lineales exactas entre las variables (no hay multicolinealidad exacta).
3. Hipótesis de linealidad: Los coeficientes son constantes a lo largo de la muestra y aparecen en el modelo de forma lineal.
4. $E(u) = 0$ y por tanto $E(u_t) = 0 \; \forall t, t = 1, 2, ...T$. Equivalentemente: $E(Y_t) = X_t'\beta, t = 1, 2, ..T \Leftrightarrow E(Y) = X\beta$
5. Var(u)=E(uu') $= \sigma^2 I_T$; Var(u_t)=E(u_t^2) $= \sigma^2$ (Supuesto de homocedasticidad) y Covar(u_t, u_s)=E($u_t u_s$) $= 0, \forall t \neq s$ (ausencia de correlación serial). Equivalentemente:
$$\left\{ \begin{array}{l} var(Y_t) = \sigma^2, \; t = 1, 2, ..T \text{ (homocedasticidad)} \\ cov(Y_t, Y_s) = 0, \forall t \neq s \text{ (ausencia de correlación serial)} \end{array} \right\} \Leftrightarrow$$
$$Var(Y) = \sigma^2 I_T$$
6. Hipótesis adicional de normalidad: $u \sim N(0, \sigma^2 I); u_t \sim N(0, \sigma^2)$ $\Leftrightarrow Y \sim N(X\beta, \sigma^2 I)$

Notas:

1. Cuando hablamos de las hipótesis básicas del MRL nos estamos refiriendo a los supuestos 1-5 y cuando hablemos del MRL con errores normales nos referiremos a los supuestos 1-6.
2. Interpretación de los coeficientes: $\beta_j = \frac{\partial E(Y_t)}{\partial X_{jt}}$ puede interpretarse como el cambio en el valor esperado de Y ante un aumento de X_j en una unidad, manteniendo constante el resto de las variables explicativas. El término constante puede interpretarse como la media de la variable dependiente Y cuando todas las variables explicativas tomen el valor cero. Así, si $X_{2t} = ... = X_{kt} = 0$, $E(Y_t) = \beta_1$.En muchos modelos, puede no tener sentido que todas las variables explicativas sean cero, en cuyo caso, la constante carecerá de interpretación.
3. Un modelo de regresión sin término constante

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t, t = 1, 2, ..., T$$

se denomina "modelo de regresión por el origen".

Estimación MCO en el modelo de regresión simple

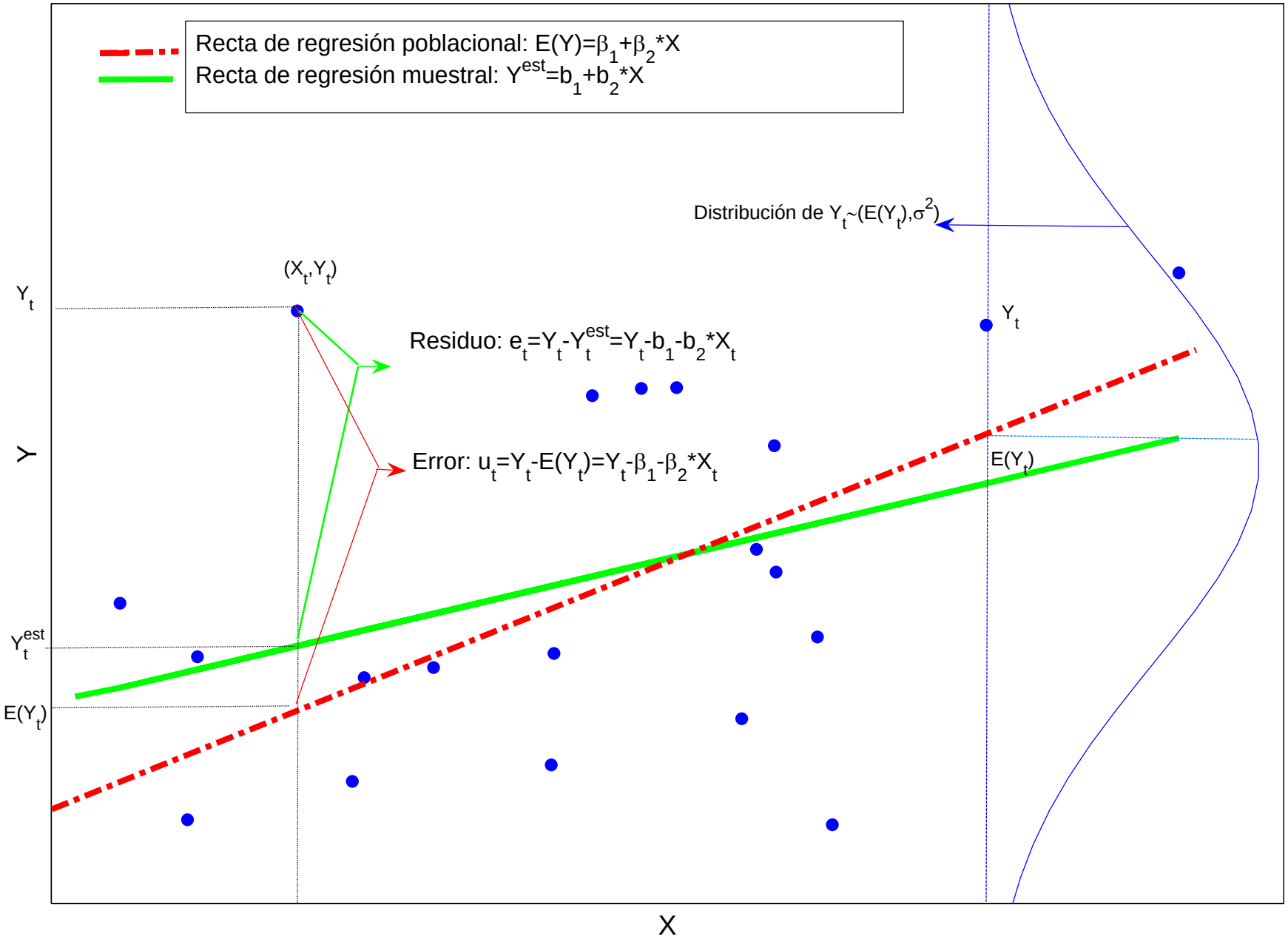
- Consideremos el modelo de regresión simple

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

y dibujemos la nube de puntos asociada a una determinada muestra de tamaño T y una recta cualquiera

$$y = b_1 + b_2 x$$

El método de estimación por Mínimos Cuadrados Ordinarios (MCO) consiste en elegir los valores de b_1 y b_2 de forma que la recta esté lo más "próxima"posible a los puntos de la nube según un determinado criterio de proximidad.


Recta de regresión poblacional: E(Y)=β1+β2*X
Recta de regresión muestral: Y^est=b1+b2*X
Distribución de Yt~(E(Yt),σ2)
(Xt,Yt)
Yt
Residuo: et=Yt-Yt^est=Yt-b1-b2*Xt
Error: ut=Yt-E(Yt)=Yt-β1-β2*Xt
Yt
E(Yt)
Yt^est
E(Yt)
Y
X

Gráficamente podemos ver que la distancia vertical del punto (X_t, Y_t) a la recta $y = b_1 + b_2 x$ viene dada por

$$Y_t - b_1 - b_2 X_t$$

y por tanto la función objetivo que tenemos que minimizar es

$$s(b_1, b_2) = \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_t)^2$$

Los coeficientes estimados se obtienen igualando a cero las derivadas parciales de la función objetivo. Las derivadas parciales vienen dadas por:

$$\begin{aligned} \frac{\partial s(b_1, b_2)}{\partial b_1} &= -2 \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_t) \\ \frac{\partial s(b_1, b_2)}{\partial b_2} &= -2 \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_t) X_t \end{aligned}$$

Igualando a cero y simplificando obtenemos las condiciones de primer orden que se denominan ecuaciones normales

$$\begin{aligned} T\widehat{\beta}_1 + \widehat{\beta}_2 \sum_{t=1}^{T} X_t &= \sum_{t=1}^{T} Y_t \\ \widehat{\beta}_1 \sum_{t=1}^{T} X_t + \widehat{\beta}_2 \sum_{t=1}^{T} X_t^2 &= \sum_{t=1}^{T} Y_t X_t \end{aligned}$$

y despejando obtenemos las expresiones de los estimadores MCO de β_1 y β_2

$$\begin{aligned} \widehat{\beta}_2 &= \frac{S_{XY}}{S_X^2} \\ \widehat{\beta}_1 &= \overline{Y} - \widehat{\beta}_2 \overline{X} \end{aligned}$$

donde $\overline{X}$ es la media muestral de las observaciones para X, $\overline{Y}$ es la media muestral de las observaciones para Y, S_{XY} es la covarianza muestral entre X e Y, y S_X es la varianza muestral de X.

Esto es

$$\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t \qquad \overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$$

$$S_{XY} = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})(Y_t - \overline{Y}) \quad S_X^2 = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})^2$$

- Las distancias verticales de los puntos a la recta de regresión

$$e_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t, \quad t = 1, 2, .., T$$

se denominan residuos MCO.

- Los valores estimados, valores ajustados o predicciones para la variable dependiente en función del modelo de regresión los denotaremos por:

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t, \quad t = 1, 2, .., T$$

Notas:

- Podemos calcular la predicción para la variable dependiente para cualquier valor de X aunque dicho valor no corresponda con ninguno de los valores observados en la muestra.
- Las distancias verticales de los puntos de la nube a la recta definidas por

$$Y_t - b_1 - b_2 X_t$$

pueden ser positivas o negativas, y por tanto un criterio que consistiera en minimizar la suma de las distancias no sería apropiado.

- Podrían considerarse otros criterios alternativos como por ejemplo minimizar la suma de los valores absolutos de las distancias verticales

$$\min_{b_1, b_2} \sum_{t=1}^{T} |Y_t - b_1 - b_2 X_t|$$

El problema de utilizar este criterio es que no es diferenciable y por tanto es más complicado calcular el mínimo.

Estimación MCO en el modelo de regresión múltiple

- Análoga al caso de regresión simple. La función objetivo que tenemos que minimizar es

$$s(b_1, b_2, ..b_k) = \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_{2t} - .. - b_K X_{kt})^2$$

y los coeficientes estimados se obtienen igualando a cero las derivadas parciales de la función objetivo dadas por:

$$\begin{aligned}
\frac{\partial s(b_1, b_2, ..b_k)}{\partial b_1} &= -2 \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_{2t} - .. - b_K X_{kt}) \\
\frac{\partial s(b_1, b_2, ..b_k)}{\partial b_2} &= -2 \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_{2t} - .. - b_K X_{kt}) X_{2t} \\
&\vdots \\
\frac{\partial s(b_1, b_2, ..b_k)}{\partial b_k} &= -2 \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_{2t} - .. - b_K X_{kt}) X_{Kt}
\end{aligned}$$

Igualando a cero y simplificando obtenemos las condiciones de primer orden (ecuaciones normales)

$$\begin{aligned}
\widehat{\beta}_1 T + \widehat{\beta}_2 \sum_{t=1}^{T} X_{2t} + \widehat{\beta}_3 \sum_{t=1}^{T} X_{3t} + ... + \widehat{\beta}_k \sum_{t=1}^{T} X_{kt} &= \sum_{t=1}^{T} Y_t \\
\widehat{\beta}_1 \sum_{t=1}^{T} X_{2t} + \widehat{\beta}_2 \sum_{t=1}^{T} X_{2t}^2 + \widehat{\beta}_3 \sum_{t=1}^{T} X_{2t}X_{3t} + ... + \widehat{\beta}_k \sum_{t=1}^{T} X_{2t}X_{kt} &= \sum_{t=1}^{T} X_{2t}Y_t \\
&... \qquad\qquad ... \\
\widehat{\beta}_1 \sum_{t=1}^{T} X_{kt} + \widehat{\beta}_2 \sum_{t=1}^{T} X_{2t}X_{kt} + \widehat{\beta}_3 \sum_{t=1}^{T} X_{3t}X_{kt} + ... + \widehat{\beta}_k \sum_{t=1}^{T} X_{kt}^2 &= \sum_{t=1}^{T} X_{kt}Y_t
\end{aligned}$$

A la hora de obtener la expresión del estimador MCO de los parámetros $\beta_1, \beta_2, .., \beta_k$ es más sencillo reescribir el sistema en notación matricial.

Se puede demostrar que

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

donde

$$\underset{(k\times 1)}{\widehat{\beta}} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \cdots \\ \widehat{\beta}_k \end{bmatrix}$$

X es la matriz de observaciones de las variables explicativas

$$\underset{(T\times k)}{X} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & & X_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2T} & \cdots & X_{kT} \end{bmatrix}$$

e Y es el vector de observaciones para la variable dependiente

$$\underset{(T\times 1)}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_ \end{bmatrix}$$

- Análogamente al caso del modelo de regresión simple los residuos MCO se definen como

$$e_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2t} - .. - \widehat{\beta}_K X_{kt}, \quad t = 1, 2, .., T$$

y el vector de residuos MCO es

$$\underset{(T\times 1)}{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_T \end{bmatrix}$$

- Análogamente al caso del modelo de regresión simple, los valores estimados, valores ajustados o predicciones para la variable dependiente en función del modelo de regresión los denotaremos por:

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_{2t} + .. + \widehat{\beta}_K X_{kt}, \quad t = 1, 2, .., T$$

- También, igual que en el modelo de regresión simple, podemos calcular la predicción para la variable dependiente para cualquier valor de $X_2, .., X_k$ aunque dichos valores no correspondan con ninguno de los valores observados en la muestra.

Interpretación de los parámetros estimados

Los valores estimados para las pendientes miden los efectos marginales estimados de cada variable sobre la variable dependiente ya que

$$\frac{\partial \widehat{Y}}{\partial X_j} = \widehat{\beta}_j, \quad j = 2, ..k$$

Por tanto $\widehat{\beta}_j$ mide el efecto que tendría sobre la variable dependiente un aumento en una unidad en X_j, manteniendo constante las restantes variables explicativas del modelo. La característica fundamental del modelo de regresión lineal es que los efectos marginales son constantes.

- Ejemplo:

  Consideremos el siguiente modelo para el gasto en vestido y calzado estimado en base a una muestra de 7427 hogares españoles

$$\widehat{gvest}_t = 1{,}2 + 0{,}064 renta_t + 0{,}132 nad_t + 0{,}159 nhijos_t$$

  donde *gvest* es el gasto anual del hogar en vestido y calzado (en miles Euros), *renta* es la renta anual del hogar (en miles de Euros), *nad* es el número de adultos en el hogar y *nhijos* es el número de hijos menores de 18 años. Según este modelo, un aumento en 1000 Euros en la renta anual del hogar produciria un aumento estimado de 64 Euros (0,064 miles de Euros) al año en vestido y calzado, manteniendo constante el número de adultos y el número de hijos en el hogar. Un adulto adicional en el hogar supone un aumento estimado de 132 Euros (0,132 miles de Euros) en el gasto en vestido y calzado, si la renta anual no ha variado y no ha cambiado el número de hijos en el hogar. ¿Qué supondría un hijo más?

- Utilizando el modelo estimado también podemos calcular las diferencias estimadas en la variable dependiente entre "individuos" con distintos valores para las variables explicativas. En el ejemplo anterior, según el modelo estimado, la diferencia en el gasto en vestido y calzado entre dos familias con la misma renta, la familia A formada por una pareja con un hijo menor de 18 años y la familia B formada por una pareja con un hijo adulto es:
  Predicción para la familia A

$$\widehat{gvest}_A = 1{,}2 + 0{,}064 renta_A + 0{,}132 * 2 + 0{,}159$$

  Predicción para la familia B

$$\widehat{gvest}_B = 1{,}2 + 0{,}064 renta_B + 0{,}132 * 3$$

  puesto que la renta de las dos familias coincide, la diferencia estimada en el gasto en vestido y calzado es

$$0{,}132 * 2 + 0{,}159 - 0{,}132 * 3 = 0{,}027$$

  Es decir la familia A gastaría 27 Euros más que la B

- La interpretación de los coeficientes estimados depende de las variables explicativas incluidas en la regresión, porque los efectos se miden manteniendo constantes el resto de las variables incluidas en la regresión. Así pues, si en el ejemplo anterior no incluimos el número de adultos en la regresión, la estimación cambia de la siguiente forma

$$\widehat{gvest}_t = 1{,}56 + 0{,}067 renta_t + 0{,}121 nhijos_t$$

puesto que los coeficientes estimados están midiendo cosas distintas. Así pues, en el modelo anterior, el coeficiente estimado para *nhijos* mide la disminución en el gasto en vestido si el número de hijos disminuye en una unidad (por ejemplo, si los hijos se mudan a vivir fuera del hogar), permaneciendo constante la renta anual y el número de adultos. Esta disminución en el modelo anterior se estima que es de 159 euros. Sin embargo, en esta ultima estimación, el efecto marginal por un hijo menos es una disminución en el gasto en vestido y calzado de 121 euros.

- En este modelo, la estimación de la constante carece de una interpretación útil, pues indica que el gasto anual en vestido y calzado predicho para una familia sin ningún adulto, ningún hijo y con renta anual igual a cero es de 1200 euros. No es posible que exista un hogar con tales valores de las variables explicativas. Un ejemplo en el cuál la constante tiene interpretación es un modelo de regresión del salario sobre la experiencia laboral del individuo. Sea $salario_t$ el salario por hora del individuo t y exp_t los años de experiencia laboral del individuo t. Consideremos el siguiente modelo estimado

$$\widehat{salario}_t = 6{,}2 + 1{,}1 exp_t$$

La interpretación del término constante nos dice que los trabajadores sin ningún año de experiencia laboral, ganan en promedio un salario por hora de 6.2 euros. Por cada año adicional de experiencia en el mercado laboral, su salario por hora aumenta en 1.1 euros.

- Si cambiamos las unidades de medida de alguna o algunas de las variables explicativas y/o de la variable dependiente, en general, variarán los valores estimados de los parámetros. Sin embargo, podemos calcular los nuevos valores de los parámetros estimados sin tener que volver a estimar el modelo. Consideremos el modelo estimado

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_{2t} + .. + \widehat{\beta}_K X_{kt}$$

Si ahora medimos la variable X_2 en otras unidades distintas $X_2^* = dX_2$, y substituimos en el modelo estimado $X_2 = \frac{X_2^*}{d}$ tenemos

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 \frac{X_{2t}^*}{d} + .. + \widehat{\beta}_K X_{kt} = \widehat{\beta}_1 + \widehat{\beta}_2^* X_{2t}^* + .. + \widehat{\beta}_K X_{kt}$$

donde $\widehat{\beta}_2^* = \frac{\widehat{\beta}_2}{d}$. Por tanto el coeficiente estimado de X_2^* será igual al coeficiente estimado de X_2 dividido por d, mientras que la constante del modelo estimado y los coeficientes estimados de las restantes variables no cambian cuando cambiamos las unidades de medida de X_2.

- Si ahora medimos la variable dependiente en otras unidades distintas $Y^* = cY$, si substituimos en el modelo estimado $Y = \frac{Y^*}{c}$ tendremos

$$\widehat{Y}_t^* = c\widehat{\beta}_1 + c\widehat{\beta}_2 X_{2t} + .. + c\widehat{\beta}_K X_{kt} = \widehat{\beta}_1^* + \widehat{\beta}_2^* X_{2t}^* + .. + \widehat{\beta}_K^* X_{kt}$$

donde $\widehat{\beta}_1^* = c\widehat{\beta}_1, \widehat{\beta}_2^* = c\widehat{\beta}_2, .., \widehat{\beta}_k^* = c\widehat{\beta}_k$, y por tanto todos los nuevos coeficientes estimados serán iguales a los coeficientes estimados que teniamos anteriormente multiplicados por c.

Siguiendo con el ejemplo anterior, si ahora medimos la renta y el gasto en Euros (en lugar de en miles de Euros como antes) el modelo estimado será

$$\widehat{gvest}_t = 1200 + 0{,}064 renta_t + 132 nad_t + 159 nhijos_t$$

- Nótese que la interpretación de los coeficientes no cambia cuando hacemos un cambio de unidades.

- Dentro del contexto del modelo de regresión lineal (modelo lineal en parámetros) podemos considerar relaciones no lineales entre las variables de interés. Los ejemplos de relaciones no lineales que aparecen con más frecuencia en Economía son:
  - Modelo lineal en logarítmos:

$$\widehat{\log(Y_t)} = \widehat{\beta}_1 + \widehat{\beta}_2 \log(X_{2t}) + .. + \widehat{\beta}_K \log(X_{kt})$$

    ahora los $\widehat{\beta}_j$, $j = 2, .., k$ son las elasticidades estimadas, es decir $\widehat{\beta}_j$ mide la variación en tanto por ciento estimada para la variable dependiente ante un aumento de un 1 % en la variable explicativa X_j. La característica de este modelo es que las elasticidades son constantes.
  - Modelo semilogarítmico log-nivel:

$$\widehat{\log(Y_t)} = \widehat{\beta}_1 + \widehat{\beta}_2 X_{2t} + .. + \widehat{\beta}_K X_{kt}$$

    ahora, para $j = 2, ..k$, $100 * \widehat{\beta}_j$ mide la variación en tanto por ciento estimada para la variable dependiente ante un aumento en una unidad en la variable explicativa X_j.

Ejemplo

Consideremos ahora el siguiente modelo para el gasto en vestido y calzado estimado en base a la misma muestra del ejemplo anterior

$$\widehat{\log(gvest_t)} = -1{,}06 + 0{,}49\log(renta_t) + 0{,}042nad_t + 0{,}088nhijos_t$$

donde *gvest* es el gasto anual del hogar en vestido y calzado (en miles Euros), *renta* es la renta anual del hogar (en miles de Euros), *nad* es el número de adultos en el hogar y *nhijos* es el número de hijos menores de 18 años. Según este modelo, un aumento de un 1 % en la renta anual del hogar produciría un aumento estimado de un 0,49 % en el gasto en vestido y calzado. Un adulto adicional en el hogar supone un aumento estimado de 4,2 % en el gasto en vestido y calzado, mientras que un hijo más supone un aumento del 8,8 % en el gasto en vestido y calzado.

- Modelo semilogarítmico nivel-log:

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 \log(X_{2t}) + .. + \widehat{\beta}_K \log(X_{kt})$$

ahora, para $j = 2, ..k$, $\frac{\widehat{\beta}_j}{100}$ mide la variación en unidades estimada para la variable dependiente ante un aumento en un 1 % en la variable explicativa X_j.

- Modelo polinomial.

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t + \widehat{\beta}_3 X_t^2 + .. + \widehat{\beta}_K X_t^{k-1}$$

En este modelo los efectos marginales vendrían dados por

$$\frac{\partial \widehat{Y}}{\partial X} = \widehat{\beta}_2 + 2\widehat{\beta}_3 X + .. + (k-1)\widehat{\beta}_K X^{k-2}$$

Ejemplo

La relación entre el salario de los trabajadores y la edad no es generalmente lineal, ya que aunque el salario aumenta con la edad (al menos hasta una cierta edad) ese aumento no es constante. Muchos estudios consideran una relación cuadrática entre el salario y la edad. Consideremos el siguiente modelo para los salarios estimado en base a una muestra de 935 individuos

$$\widehat{salario}_t = -7{,}92 + 0{,}605 educ_t + 0{,}357 edad_t - 0{,}0022 edad2_t$$

donde *salario* es el salario mensual en cientos de dolares, *educ* es el nivel de educación en años, *edad* es la edad en años y *edad*2 es la edad al cuadrado. Según el modelo estimado, el efecto marginal es

$$\frac{\partial \widehat{salario}}{\partial edad} = 0{,}357 - 2 * 0{,}0022 * edad$$

Asi, un año más de edad supone para un trabajador de 30 años un aumento en el salario mensual de $0{,}357 - 2 * 0{,}0022 * 30 = 0{,}225$ cientos de dólares, es decir de 22,5 dólares.

- La forma de estimar el modelo no depende de cómo esten definidas las variables (en niveles, en logaritmos, etc.), pero si es muy importante tener en cuenta como están definidas para poder interpretar correctamente los resultados de la estimación.
- No todos los modelos se pueden tratar como modelos de regresión lineal. Por ejemplo el modelo $y = \frac{1}{\beta_1+\beta_2 x} + u$ es intrínsicamente no lineal y para estimar los parámetros hay que utilizar técnicas econométricas más complejas que no se van a estudiar en este curso.

Propiedades del ajuste MCO

Son propiedades algebraicas que se cumplen siempre que calculamos $\hat{\beta}, \hat{\sigma}, e, \hat{Y}$ a partir de los datos, con independencia de que se verifiquen o no las hipótesis básicas.

(a.1) $X'e = 0 \rightarrow$ Los residuos son ortogonales a las variables explicativas puesto que $X'e = X'\left(Y - \hat{Y}\right) = X'Y - X'X\hat{\beta} = 0$

(a.2) $\hat{Y}'e = 0; \quad \hat{Y} = X\hat{\beta}; \quad Y = \hat{Y} + e$

(a.3) $Y'Y = \hat{Y}'\hat{Y} + e'e \rightarrow \sum_{t=1}^{T} Y_t^2 = \sum_{t=1}^{T} \hat{Y}_t^2 + \sum_{t=1}^{T} e_t^2$

Propiedades del ajuste MCO

(a.4) $\sum_{t=1}^{T} e_t = 0 \rightarrow$ Es un caso particular del anterior punto (a.1), considerando que la primera columna de la matriz X es una columna de unos. Puesto que $X'e = 0$ para todas las filas, en concreto la primera: $e_1 + e_2 + ... + e_T = \sum_{t=1}^{T} e_t = 0$

(a.5) El hiperplano estimado pasa por $(\bar{Y}, \bar{X}_2, ..., \bar{X}_k)$; es decir, $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + ... + \hat{\beta}_k \bar{X}_k$.

(a.6) $\bar{Y} = \widehat{\bar{Y}}$, la media de la variable dependiente coincide con la media de los valores ajustados, puesto que $e = Y - \hat{Y} \rightarrow e_t = Y_t - \hat{Y}_t \rightarrow \sum_{t=1}^{T} e_t = \sum_{t=1}^{T} Y_t - \sum_{t=1}^{T} \hat{Y}_t = 0 \rightarrow \bar{Y} - \widehat{\bar{Y}} = 0$

- Medidas de bondad de ajuste:

Tras estimar el modelo, interesa saber cómo la función estimada se ajusta a los datos.
Definiciones:

1. Suma Cuadrática Total:
$$SCT = \sum_{t=1}^{T} (Y_t - \overline{Y})^2 = \sum Y_t^2 - T\overline{Y}^2 = Y'Y - T\overline{Y}^2;$$
2. Suma Cuadrática Explicada:
$$SCE = \sum_{t=1}^{T} (\widehat{Y}_t - \overline{\widehat{Y}})^2 \underset{\text{como } \overline{Y}=\overline{\widehat{Y}}}{=} \sum_{t=1}^{T} (\widehat{Y}_t - \overline{Y})^2 = \sum \widehat{Y}_t^2 - T\overline{Y}^2 =$$
$$\widehat{Y}'\widehat{Y} - T\overline{Y}^2;$$
3. Suma Cuadrática Residual:
$$SCR = \sum_{t=1}^{T} e_t^2 = e'e.$$

SCT, SCE y SCR son no negativos (pues son sumas de cuadrados) y además son medidas del grado de variabilidad de la variable dependiente, de los valores ajustados y de los residuos, respectivamente, pues son el numerador de la varianza muestral de cada una de estas variables (recuérdese que los valores ajustados tienen media $\overline{Y}$ y los residuos tienen media 0). Estas tres medidas están relacionadas entre sí de la forma

$$SCT = SCE + SCR$$

como puede comprobarse fácilmente:

$$SCT = Y'Y - T\overline{Y}^2 \underset{\text{Utilizando (a.3)}}{=} \widehat{Y}'\widehat{Y} - T\overline{Y}^2 + e'e = SCE + SCR$$

Supondremos ahora que SCT no es nula, lo que equivale a decir que las observaciones de la variable dependiente no son todas iguales, dividiendo los tres sumandos de la igualdad anterior por SCT nos queda que:

$$1 = \frac{SCE}{SCT} + \frac{SCR}{SCT}$$

Se define el coeficiente de determinación:

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

Mide la proporción de variabilidad de la variable dependiente Y, que viene explicada por el modelo

Dado que $SCT = SCE + SCR$, $R^2 \in [0,1]$:

- $SCR = 0 \rightarrow R^2 = 1 \rightarrow$ El modelo explica toda la variabilidad de Y.
- $SCR = SCT \rightarrow R^2 = 0 \rightarrow$ El modelo no explica nada de la variabilidad de Y.

Problema: No tiene en cuenta el número de regresores que entran en el modelo. Se mantiene constante o aumenta con la incorporación de variables explicativas.
Solución: El coeficiente de determinación corregido $\left(\bar{R}^2\right)$

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}\frac{SCR}{SCT} = 1 - \frac{T-1}{T-k}\left(1 - R^2\right)$$

Este coeficiente corrige SCT y SCR por sus grados de libertad, penalizando la incorporación de nuevas variables explicativas, haciendo disminuir $\bar{R}^2$ si está poco relacionada con la variable dependiente.
Si $R^2 = 1 \rightarrow \bar{R}^2 = 1$, si $R^2 = 0 \rightarrow \bar{R}^2 = \frac{1-k}{T-k} \leq 0$.

Se puede demostrar que R^2 y $\bar{R}^2$ no dependen de las unidades de medida.

Propiedades del ajuste que no se verifican en el modelo de regresión por el origen:

- La suma de los residuos no es necesariamente cero. El motivo es que cuando el modelo no tiene constante ninguna de las columnas de X es una columna de unos y por tanto no se verifica la propiedad (a.4).
- La media de la variable dependiente puede no coincidir con la media de los valores ajustados.

$$SCT \neq SCE + SCR$$

ya que como $\overline{\widehat{Y}} \neq \overline{Y}$, $SCE \neq \widehat{Y}'\widehat{Y} - T\overline{Y}^2$

- La interpretación del R^2 no está clara. De hecho, si definimos

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

R^2 puede tomar valores negativos.