

ECONOMETRIA

Tema 2: El Modelo de Regresión Lineal Simple

César Alonso

UC3M

Curso 2009/2010

- Como economistas, nos interesa la relación entre dos o más variables económicas. Por ello, nos concentramos en poblaciones, al menos, bivariantes.
- La teoría económica postula, en general, relaciones del tipo

$$Y = f(X)$$

donde $f(\cdot)$ es una función.

- Dichas relaciones son **exactas o determinísticas**, de manera que a cada valor de X le corresponde un único valor de Y .
- Si tuviéramos más variables exógenas, el razonamiento sería idéntico

$$Y = f(X_1, \dots, X_K)$$

a cada combinación de valores de X_1, \dots, X_K le corresponde un único valor de Y .

- ¿Qué sucede en general con los datos reales de variables económicas?
- **Ejemplo:** Relación entre tasa de ahorro (Y) y renta (X) (Goldberger, Capítulo 1 de “A Course in Econometrics”, 1991. Harvard U. Press.)
 - La teoría económica predice una relación creciente entre tasa de ahorro y renta
 - Datos de 1027 familias de EE.UU. en los años 1960 a 1962.
 - Para simplificar, hemos agrupado los datos en intervalos para ambas variables, poniendo el punto medio del intervalo.
Para cada combinación de X e Y presentamos la frecuencia relativa (en tanto por uno).

Relaciones empíricas y teóricas

Distribución conjunta de frecuencias de X e Y

$P(X, Y)$	X (renta en miles de dólares)					
Y (tasa de ahorro)	1.4	3.0	4.9	7.8	14.2	$P(Y)$ (suma de filas)
0.45	0.015	0.026	0.027	0.034	0.033	0.135
0.18	0.019	0.032	0.057	0.135	0.063	0.306
0.05	0.059	0.066	0.071	0.086	0.049	0.331
-0.11	0.023	0.035	0.045	0.047	0.015	0.165
-0.25	0.018	0.016	0.016	0.008	0.005	0.063
$P(X)$ (suma de columnas)	0.134	0.175	0.216	0.310	0.165	1.000

- Dada la evidencia empírica, ¿podemos afirmar que existe una relación determinística entre tasa de ahorro y renta?
 - Para que ello fuera cierto, deberíamos encontrar en cada columna (para cada nivel de renta X) una única frecuencia distinta de 0.
 - Claramente, esto NO es cierto: para cada nivel de renta, existen familias que ahorran mucho y familias que desahorran mucho.
- NO hay una función que relacione ahorro y renta: tenemos una distribución, con valores más y menos probables:
 - Observamos una proporción mayor de familias con tasas de ahorro más altas cuanto mayor es su renta.

- Para verlo mejor, podemos concentrarnos en las distribuciones condicionales de la tasa de ahorro para cada nivel de renta.
 - Para ello, tenemos que dividir las frecuencias relativas de cada columna por la suma de éstas

Distribuciones condicionales de frecuencias de Y
para cada valor de X

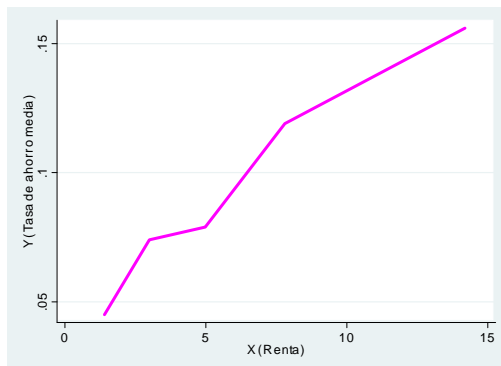
$P(Y X)$	X (renta en miles de dólares)				
Y (tasa de ahorro)	1.4	3.0	4.9	7.8	14.2
0.45	0.112	0.149	0.125	0.110	0.200
0.18	0.142	0.183	0.264	0.435	0.382
0.05	0.440	0.377	0.329	0.277	0.297
-0.11	0.172	0.200	0.208	0.152	0.091
-0.25	0.134	0.091	0.074	0.026	0.030
Suma de columnas	1	1	1	1	1
Media cond. $\hat{\mu}_{Y X}$	0.045	0.074	0.079	0.119	0.156

Relaciones empíricas y teóricas

- Vemos que, en términos relativos, las tasas de ahorro negativas son más frecuentes para rentas bajas.
- Parece existir una contradicción entre la relación funcional exacta predicha por la teoría económica y la evidencia empírica:
 - La teoría afirma que las familias de igual renta deberían presentar la misma tasa de ahorro
 - PERO vemos que no es cierto en realidad.
 - Y no podemos argumentar que lo que observamos es una mera desviación del comportamiento óptimo.
(implicaría que la mayoría de las familias “se equivocan” sistemáticamente).
- Por supuesto, cabe argumentar que hay otras características en las que difieren familias de igual renta.
 - Ello requeriría condicionar en otras características.
 - Ello reduciría la dispersión (tendríamos celdas con valores cercanos a 0).
 - PERO seguiríamos teniendo tasas de ahorro distintas para familias parecidas.

- **CONCLUSIÓN:** las relaciones empíricas entre variables económicas NO son determinísticas, sino **estocásticas**
- Para reconciliar teoría y datos, debemos reinterpretar la teoría económica:
 - Cuando la teoría postula que Y es función de X , entenderemos que el valor medio de Y es una función de X .
- En el ejemplo, vemos que las distribuciones condicionales del ahorro para cada nivel de renta varían con la renta:
 - Cuanto mayor es la renta, las tasas de ahorro tienden a ser mayores.
 - Ello implica que la tasa de ahorro media, condicional a la renta, aumenta con la renta.

- **Interpretación:** la media de la tasa de ahorro Y es una función creciente de la renta X . Gráficamente:



- Dada la distribución de probabilidad conjunta de (X, Y) (por ejemplo, tasa de ahorro y renta familiar), supongamos que nos preguntan la tasa de ahorro de una familia tomada aleatoriamente de la población de interés.
- Supongamos que nuestro criterio para medir el error en la predicción $c(X)$ es la minimización de $E(U^2)$, siendo:

$$U = Y - c(X)$$

el error de predicción, pudiendo emplear en la predicción de Y el valor de X correspondiente.

Conceptos Previos

Mejor Predicción Constante

- Supongamos que no conocemos la renta de la familia considerada (X).
- Entonces, nuestra elección de predictores queda restringida a la información sobre la distribución marginal de la tasa de ahorro Y .
- En el ejemplo anterior, para calcular la distribución marginal de Y debemos sumar las frecuencias observadas para cada fila.

Y (tasa de ahorro)	$P(Y)$
0.45	0.135
0.18	0.306
0.05	0.331
-0.11	0.165
-0.25	0.063

Conceptos Previos

Mejor Predicción Constante

- En este caso, ignoramos cómo se comporta Y de acuerdo con X .
- La predicción que podemos hacer sobre Y se limita a las constantes.
- El error de predicción será $U = Y - c$. Se elegirá c tal que minimice $E(U^2) = \sum_k (Y_k - c)^2 p_k$. Dicho valor no es otro que:

$$c = E(Y) = \mu_Y$$

- La **media poblacional** μ_Y es el **mejor predictor constante** de Y en una distribución de probabilidad bivalente (véase Capítulo 3 de Goldberger).
- En el ejemplo, suponiendo que la distribución presentada se refiere a una población,

$$\begin{aligned} E(Y) &= 0.45 \times 0.135 + 0.18 \times 0.306 + 0.05 \times 0.331 \\ &\quad - 0.11 \times 0.165 - 0.25 \times 0.063 \\ &= 0.09848 = 9.85\% \end{aligned}$$

Conceptos Previos

Mejor Predicción Lineal

- Supongamos que conocemos la renta (X) de la familia para la que queremos predecir su tasa de ahorro (Y).
- Además, sólo podemos elegir predictores que sean funciones lineales de X , es decir,

$$c(X) = c_0 + c_1 X,$$

siendo c_0 y c_1 constantes.

- El error de predicción será $U = Y - c_0 - c_1 X$. Se elegirán aquellas constantes c_0 y c_1 que minimicen $E(U^2) = \sum_k (Y_k - c_0 - c_1 X_k)^2 p_k$.
- Sean α_0 , α_1 , dichas constantes, de manera que $c(X) = \alpha_0 + \alpha_1 X$, verificando que

$$c_0 = \alpha_0 = E(Y) - \alpha_1 E(X) = \mu_Y - \alpha_1 \mu_X,$$

$$c_1 = \alpha_1 = \frac{C(X, Y)}{V(X)} = \frac{\sigma_{XY}}{\sigma_X^2}.$$

- La recta $\alpha_0 + \alpha_1 X$ es la **proyección lineal (o mejor predicción lineal)** de Y dado X

$$L(Y | X) = \alpha_0 + \alpha_1 X$$

- En nuestro ejemplo

$$C(X, Y) = E(XY) - E(X)E(Y)$$

tenemos que calcular los $5 \times 5 = 25$ valores resultantes de multiplicar cada uno de los valores de X e Y , respectivamente, y presentar la celda correspondiente a la probabilidad de ocurrencia de cada valor:

Conceptos Previos

Mejor Predicción Lineal

Distribución marginal de XY

XY	XY	XY	XY	XY	XY	XY	XY	XY	XY
-3.55	.005	-0.75	.016	0.07	.059	0.54	.032	1.40	.135
-1.95	.008	-0.54	.045	0.15	.066	0.63	.015	2.21	.027
-1.56	.015	-0.35	.018	0.25	.071	0.71	.049	2.56	.063
-1.23	.016	-0.33	.035	0.25	.019	0.88	.057	3.51	.034
-0.86	.047	-0.15	.023	0.39	.086	1.35	.026	6.39	.033

donde

$$E(XY) = \sum_{i=1}^5 \sum_{j=1}^5 X_i Y_j \Pr(XY = X_i Y_j) = 0.782607$$

y

$$\begin{aligned} E(X) &= 1.4 \times 0.134 + 3.0 \times 0.175 + 4.9 \times 0.216 \\ &\quad + 7.8 \times 0.310 + 14.2 \times 0.165 = 6.532 \end{aligned}$$

y por tanto,

$$C(X, Y) = 0.782607 - 6.532 \times 0.09848 = 0.13934.$$

- En consecuencia, teniendo en cuenta que

$$E(X^2) = 1.4^2 \times 0.134 + 3.0^2 \times 0.175 + 4.9^2 \times 0.216 \\ + 7.8^2 \times 0.310 + 14.2^2 \times 0.165 = 59.155$$

entonces

$$V(X) = E(X^2) - [E(X)]^2 = 59.155 - 6.532^2 = 16.488$$

con lo cual

$$c_1 = \alpha_1 = \frac{C(X, Y)}{V(X)} = \frac{0.13934}{16.488} = 0.008451$$

$$c_0 = \alpha_0 = E(Y) - \alpha_1 E(X) = 0.09848 - 0.008451 \times 6.532 = 0.0432$$

y por tanto la función de proyección lineal es

$$L(Y | X) = 0.043278 + 0.008451X$$

- Aplicada únicamente a los valores de renta X , podemos escribir la proyección lineal como

$$L(Y|X) = \begin{cases} 0.043278 + 0.008451 \times 1.4 = 0.055 & \text{si } X = 1.4 \\ 0.043278 + 0.008451 \times 3.0 = 0.069 & \text{si } X = 3.0 \\ 0.043278 + 0.008451 \times 4.9 = 0.085 & \text{si } X = 4.9 \\ 0.043278 + 0.008451 \times 7.8 = 0.1092 & \text{si } X = 7.8 \\ 0.043278 + 0.008451 \times 14.2 = 0.1633 & \text{si } X = 14.2 \end{cases}$$

- Supongamos que conocemos la renta (X) de la familia antes de hacer la predicción de su tasa de ahorro (Y).
- Además, podemos elegir como función de predicción cualquier función de X , $c(X)$.
- El error de predicción será $U = Y - c(X)$. Se elegirá $c(X)$ de forma que minimice $E(U^2)$, resultando que $c(X) = E(Y | X)$.
- El **mejor predictor** de Y dado X es su **esperanza condicional**, $E(Y | X)$.
 - **Solamente** cuando la función de esperanza condicional es lineal, la función de proyección lineal $L(Y | X)$ y la función de esperanza condicional $E(Y | X)$ coinciden.
 - De lo contrario, cuando la función de esperanza condicional no es lineal, entonces la proyección lineal no es el mejor predictor, pero es la **mejor aproximación lineal** a la función de esperanza condicional.

- La función de esperanza condicional viene dada por las medias de cada una de las distribuciones condicionales de Y para cada uno de los valores de X .
- En el ejemplo,

Distribuciones condicionales de frecuencias de Y
para cada valor de X

$P(Y X)$	X (renta en miles de dólares)				
Y (tasa de ahorro)	1.4	3.0	4.9	7.8	14.2
0.45	0.112	0.149	0.125	0.110	0.200
0.18	0.142	0.183	0.264	0.435	0.382
0.05	0.440	0.377	0.329	0.277	0.297
-0.11	0.172	0.200	0.208	0.152	0.091
-0.25	0.134	0.091	0.074	0.026	0.030
$\hat{\mu}_{Y X}$	0.045	0.074	0.079	0.119	0.156

- La función de media condicional se obtiene calculando $E(Y|X)$ para cada uno de los valores de X :

$$E(Y|X = 1.4) = 0.45 \times 0.112 + 0.18 \times 0.142 + 0.05 \times 0.440 \\ - 0.11 \times 0.172 - 0.25 \times 0.134 = 0.045$$

$$E(Y|X = 3.0) = 0.45 \times 0.149 + 0.18 \times 0.183 + 0.05 \times 0.377 \\ - 0.11 \times 0.200 - 0.25 \times 0.091 = 0.074$$

$$E(Y|X = 4.9) = 0.45 \times 0.125 + 0.18 \times 0.264 + 0.05 \times 0.329 \\ - 0.11 \times 0.208 - 0.25 \times 0.074 = 0.079$$

$$E(Y|X = 7.8) = 0.45 \times 0.110 + 0.18 \times 0.435 + 0.05 \times 0.277 \\ - 0.11 \times 0.152 - 0.25 \times 0.026 = 0.119$$

$$E(Y|X = 14.2) = 0.45 \times 0.200 + 0.18 \times 0.382 + 0.05 \times 0.297 \\ - 0.11 \times 0.091 - 0.25 \times 0.030 = 0.156$$

- De manera que la función de esperanza condicional se puede escribir como

$$E(Y|X) = \begin{cases} 0.045 & \text{si } X = 1.4 \\ 0.074 & \text{si } X = 3.0 \\ 0.079 & \text{si } X = 4.9 \\ 0.119 & \text{si } X = 7.8 \\ 0.156 & \text{si } X = 14.2 \end{cases}$$

- En resumen

X (renta en miles de dólares)	Predictores de tasa de ahorro		
	C	$L(Y X)$	$E(Y X)$
1.4	0.0985	0.055	0.045
3.0	0.0985	0.069	0.074
4.9	0.0985	0.085	0.079
7.8	0.0985	0.1092	0.119
14.2	0.0985	0.1633	0.156

- Las predicciones asociadas a la proyección lineal son distintas de las basadas en la función de esperanza condicional, porque ésta no es lineal.
 - En el gráfico presentado anteriormente, puede verse que la función de esperanza condicional no es lineal.
 - $L(Y|X)$ proporciona una aproximación bastante buena a $E(Y|X)$. Ello implica que $L(Y|X)$ puede ser, en casos como éste, un buen predictor, aunque no coincida con $E(Y|X)$.
 - Pero mientras que $E(Y|X)$ caracteriza momentos (medias condicionales) de las correspondientes distribuciones condicionales de Y dado X , $L(Y|X)$ NO.
 - Ello implica que $E(Y|X)$ puede tener una interpretación causal, pero $L(Y|X)$ NO.

Introducción al modelo de regresión lineal simple

- El Modelo de Regresión Lineal Simple se puede emplear para estudiar la relación entre dos variables, aunque tiene limitaciones como herramienta para el análisis empírico.
- Objeto de estudio: Y y X son dos variables que representan alguna población y estamos interesados en “explicar Y en términos de X ” o en “estudiar cómo varía Y ante variaciones en X ”.
Por ejemplo, $Y =$ ventas, $X =$ gastos en publicidad; $Y =$ tasa ahorro, $X =$ renta.
- Al tratar de formular un modelo que “explique Y en términos de X ” debemos afrontar varias cuestiones:
 - ¿Cómo tenemos en cuenta otros factores que afecten a Y además de X ?
 - ¿Cuál es la forma funcional de la relación entre Y y X ?
 - ¿Estamos captando con nuestro modelo una relación ceteris-paribus entre Y y X ?

- El Modelo de Regresión Lineal Simple nos permite “explicar Y en términos de X ” resolviendo las cuestiones anteriores.
- Sea

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

donde:

- Y : Variable dependiente, endógena, explicada, de respuesta...
- X : Variable independiente, exógena, explicativa, de control, regresor..
- β_0 y β_1 : Parámetros poblacionales
- ε : Término de error o perturbación inobservable. Representa los factores que influyen en Y además de X , el componente aleatorio de Y que no viene explicado por $\beta_0 + \beta_1 X$.

- *Ejemplo 1:*

Si $Y =$ salario y $X =$ años de estudio, entonces el término de error puede recoger factores inobservables como:

- experiencia laboral
- capacidad o habilidad
- antigüedad en la empresa

- *Ejemplo 2:*

Si $Y =$ cosecha y $X =$ cantidad de abono, entonces el término de error puede recoger factores como:

- calidad de la tierra
- lluvia.

1 **Linealidad en los parámetros** ($Y = \beta_0 + \beta_1 X + \varepsilon$).

- Este supuesto implica que un cambio unitario en X tiene el mismo efecto sobre Y con independencia del valor inicial de X .
- Puede no ser realista para algunas aplicaciones económicas.
(por ejemplo, en el caso de salario y educación podemos pensar en la existencia de rendimientos crecientes)
- Esta limitación puede superarse formulando modelos lineales en parámetros que recogen relaciones no lineales entre variables.

2 $E(\varepsilon|X) = 0$, es decir:

Para cualquier valor de X , la media de los inobservables es siempre la misma e igual a cero

(que es la media de los inobservables para el total de la población)

• **Implicaciones:**

- $E(\varepsilon) = 0$

Por la ley de esperanzas iteradas,

$$E(\varepsilon) = E[E(\varepsilon|X)] = 0$$

Supuestos del modelo de regresión simple

- Que $E(\varepsilon|X) = 0$ implica que $C(h(X), \varepsilon) = 0$, donde $h(\cdot)$ es cualquier función de X .

Por tanto, ε no está correlacionado con ninguna función de X .

- En particular, $C(X, \varepsilon) = 0$

$$C(X, \varepsilon) = E(X\varepsilon) - E(X)E(\varepsilon) \text{ donde}$$

$$E(X\varepsilon) = E[E(X\varepsilon|X)] = E[X E(\varepsilon|X)] = 0$$

$$E(X)E(\varepsilon) = 0 \text{ dado que } E(\varepsilon) = 0$$

- Nótese que $E(\varepsilon|X) = 0 \Rightarrow C(X, \varepsilon) = 0$, pero $C(X, \varepsilon) = 0 \not\Rightarrow E(\varepsilon|X) = 0$ (que $C(X, \varepsilon) = 0$ es cond. necesaria, pero no suficiente, para $E(\varepsilon|X) = 0$).

Supuestos del modelo de regresión simple

- $E(Y|X) = \beta_0 + \beta_1 X$
 - La **función de esperanza condicional** o **función de regresión poblacional** es **lineal**.

Entonces:

$$C(Y, X) = C[E(Y|X), X] = \beta_1 V(X) \Rightarrow \beta_1 = \frac{C(Y, X)}{V(X)}$$

$$E(Y) = E[E(Y|X)] = \beta_0 + \beta_1 E(X) \Rightarrow \beta_0 = E(Y) - \beta_1 E(X)$$

- Nótese que:
- Hemos de utilizar esperanzas condicionales en X dado el carácter estocástico de dicha variable.
 - Si X fuera determinística (como ocurriría en el caso de datos experimentales), bastaría con aplicar esperanzas marginales.

Supuestos del modelo de regresión simple

- Al ser la función de esperanza condicional lineal en X ,

$$E(Y|X) = L(Y|X) = \beta_0 + \beta_1 X$$

donde $L(\cdot)$ denota la proyección lineal de Y dado X .

- β_0 y β_1 son los parámetros que minimizan la varianza del error $\varepsilon = Y - \beta_0 - \beta_1 X$, es decir, resuelven el problema

$$\min [E(Y - \beta_0 - \beta_1 X)^2]$$

cuyas condiciones de primer orden son

$$E(Y - \beta_0 - \beta_1 X) \equiv E(\varepsilon) = 0 \Rightarrow \beta_0 = E(Y) - \beta_1 E(X)$$

$$E[(Y - \beta_0 - \beta_1 X)X] = E(X\varepsilon) = 0 \Rightarrow \beta_1 = \frac{C(Y, X)}{V(X)}$$

3. $V(\varepsilon|X) = \sigma^2$ para todo X
(Homocedasticidad condicional)

Implicaciones:

- $V(\varepsilon) = \sigma^2$.

Para verlo,

$$V(\varepsilon|X) = E(\varepsilon^2|X) - E[E(\varepsilon|X)^2] = E(\varepsilon^2|X) = \sigma^2$$

$$E(\varepsilon^2) = E[E(\varepsilon^2|X)] = \sigma^2$$

$$V(\varepsilon) = E(\varepsilon^2) - [E(\varepsilon)]^2 = \sigma^2$$

- $V(Y|X) = \sigma^2$

La varianza de Y dado X es constante.

- Supongamos, que el supuesto **1.** de **linealidad en parámetros** se cumple, de manera que nuestra especificación es

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

- Queremos ver cómo podemos interpretar los parámetros de este modelo.
- La interpretación depende de que se cumpla o no el supuesto **2.**
 $E(\varepsilon|X) = 0$
- Si $E(\varepsilon|X) = 0$, entonces tenemos que

$$\begin{aligned} E(Y|X) &= \beta_0 + \beta_1 E(X|X) + E(\varepsilon|X) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

- En este caso, $E(Y|X) = L(Y|X)$: la función de esperanza condicional es lineal, y por tanto coincide con la proyección lineal de Y dado X .
- Por tanto, la **pendiente** β_1 tiene una interpretación causal:

$$\beta_1 = \frac{\Delta E(Y|X)}{\Delta X}$$

Cuando X aumenta en una unidad, Y varía, en media, β_1 unidades de Y .

- La pendiente β_1 mide el cambio promedio en Y ante un cambio unitario en X .
- En otras palabras, β_1 mide la diferencia de medias entre la distribución condicional $f(Y|X = x)$ y la distribución condicional $f(Y|X = x + \Delta x)$.

- En cuanto a la **constante** (también llamada **término constante**) β_0 , puede verse que

$$E(Y|X = 0) = \beta_0,$$

es decir: β_0 es el valor medio de Y cuando $X = 0$.

- Geométricamente, es el valor de la recta de regresión en el eje de ordenadas.
- En la práctica, β_0 no tiene a menudo interpretación, en aquellos casos en que no tiene sentido que $X = 0$.
- Sin embargo, el término constante β_0 debe incluirse siempre en el modelo, para controlar por el hecho de que X e Y no tienen porqué tener media 0.

- Si $E(\varepsilon|X) \neq 0$, entonces tenemos que

$$\begin{aligned} E(Y|X) &= \beta_0 + \beta_1 E(X|X) + E(\varepsilon|X) \\ &= \beta_0 + \beta_1 X + E(\varepsilon|X) \\ &\neq \beta_0 + \beta_1 X \end{aligned}$$

- En este caso, si $E(\varepsilon|X) \neq 0$,

$$E(Y|X) \neq L(Y|X).$$

- Los parámetros β_0 y β_1 son en este caso solamente los parámetros de la proyección lineal, $L(Y|X)$.
- Pero β_0 y β_1 no tienen una interpretación causal.
- En resumen:
Si $E(\varepsilon|X) \neq 0$, $Y = \beta_0 + \beta_1 X + \varepsilon$ caracteriza una proyección lineal, pero no una esperanza condicional, y NO tiene interpretación causal.

- Nuestro objetivo consiste en estimar los parámetros poblacionales, los “betas”, a partir de un conjunto de datos.
- Supondremos que nuestros datos (y_i^*, x_i^*) , con $i = 1, \dots, n$, son una realización de una **muestra aleatoria** de tamaño n de una población, (Y_i, X_i) .
- Sea el modelo:

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

donde:

- $E(\varepsilon|X) = 0$
- $V(\varepsilon|X) = \sigma^2$

- ¿Cómo podemos estimar los parámetros β_0 , β_1 y σ^2 ?
 - Necesitaremos una muestra de la población
- Dada una muestra aleatoria de tamaño n de la población, podemos escribir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

donde para todo $i = 1, \dots, n$:

- $E(\varepsilon_i | X_i) = 0$
 - $V(\varepsilon_i | X_i) = \sigma^2$
- Vamos a ver cómo podemos obtener estimadores de los parámetros β_0 , β_1 y σ^2 , denotados como $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma}^2$

- Los parámetros de interés son características de la población, que son funciones de momentos poblacionales. El principio de analogía consiste en utilizar como estimador la característica análoga en la muestra.
- *Ejemplo:* **media marginal**

Sea una muestra aleatoria de observaciones de Y , $\{y_i\}_{i=1}^n$. Para estimar la media marginal de Y , $E(Y)$, utilizamos la media muestral $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$

- Recuérdese que, bajo los supuestos que hacíamos en el modelo simple en la población:

$$E(Y|X) = L(Y|X) = \beta_0 + \beta_1 X$$

y que por tanto β_0 y β_1 se obtienen de minimizar:

$$E(\varepsilon^2) = E(Y - \beta_0 - \beta_1 X)^2$$

siendo:

$$\beta_0 = E(Y) - \beta_1 E(X)$$
$$\beta_1 = \frac{C(Y, X)}{V(X)}$$

- Aplicando el **principio de analogía**, sustituyendo momentos poblacionales por muestrales, obtenemos estimadores de β_0 , β_1 :

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\frac{1}{n} \sum_i (X_i - \bar{X}) Y_i}{\frac{1}{n} \sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}\end{aligned}$$

- Podemos ver también este mismo estimador de la siguiente forma: bajo los supuestos que hacíamos en el modelo simple en la población, β_0 y β_1 son los parámetros que minimizan la varianza del error $\varepsilon = Y - \beta_0 - \beta_1 X$, es decir, resuelven el problema

$$\min E(\varepsilon^2),$$

o de forma equivalente,

$$\min [E(Y - \beta_0 - \beta_1 X)^2].$$

- Para una observación de la muestra, el análogo muestral del **término de error o perturbación** (desviación entre el valor observado y valor esperado) $\varepsilon_i = Y_i - E(Y_i|X_i)$, donde la función de esperanza condicional es lineal, se conoce como **residuo** (desviación entre el valor observado Y_i y el **valor ajustado o valor predicho**

$$\hat{Y}_i = (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{L}(Y_i|X_i),$$

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- Por tanto, el análogo muestral del problema de minimizar $E(\varepsilon^2)$ es

$$\min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2,$$

- Por tanto, el estimador que hemos obtenido antes a partir del principio de analogía puede interpretarse también como un estimador que minimiza la suma de los cuadrados de los residuos, lo que se conoce como estimador de **mínimos cuadrados ordinarios (MCO)**
- Las condiciones de primer orden son:

$$\sum_i \hat{\varepsilon}_i = 0,$$
$$\sum_i X_i \hat{\varepsilon}_i = 0.$$

- 0, de forma equivalente,

$$\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i = 0 \text{ (media muestral de los residuos 0)}$$

$$\frac{1}{n} \sum_i x_i \hat{\varepsilon}_i = 0 \text{ (covarianza muestral entre residuos y regresores 0)}$$

donde $x_i = X_i - \bar{X}$ (desviación respecto a la media muestral).

- Nótese que estas condiciones de primer orden son el **análogo muestral** de las condiciones de primer orden para el modelo de regresión clásico referido a los β 's en la población:

$$\begin{aligned} E(\varepsilon) &= 0, \\ C(X, \varepsilon) &= 0. \end{aligned}$$

- El sistema de ecuaciones normales nos queda como:

$$\begin{aligned}n\hat{\beta}_0 + \hat{\beta}_1 \sum_i X_i &= \sum_i Y_i \\ \hat{\beta}_1 \sum_i x_i^2 &= \sum_i y_i x_i\end{aligned}$$

- En el modelo de regresión simple

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n,$$

los estimadores MCO de β_0 y β_1 , es decir, $\hat{\beta}_0$ y $\hat{\beta}_1$, serían los argumentos que minimizan:

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

- Las condiciones de primer orden serían:

$$\sum_i \hat{\varepsilon}_i = 0 \quad \Rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\sum_i X_i \hat{\varepsilon}_i = 0 \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

- Los **valores predichos** o **valores ajustados** en base a los estimadores MCO resultantes, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, verifican que

$$\sum_i \hat{Y}_i \hat{\varepsilon}_i = 0 \quad (\text{covarianza } 0 \text{ entre valores ajustados MCO y residuos MCO}).$$

(es inmediato de comprobar explotando las condiciones de primer orden $\sum_i \hat{\varepsilon}_i = 0$ y $\sum_i X_i \hat{\varepsilon}_i = 0$, puesto que \hat{Y}_i es una función lineal de X_i).

Propiedades de los estimadores MCO

Linealidad (en las observaciones de Y)

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n} \sum_i Y_i - \hat{\beta}_1 \bar{X}$
- $\hat{\beta}_1 = \frac{\sum_i x_i Y_i}{\sum_i x_i^2} = \sum_i c_i Y_i$, donde $x_i = X_i - \bar{X}$, $c_i = \frac{x_i}{\sum_i x_i^2}$

- Esta propiedad se verifica si se cumplen los supuestos **1.** (linealidad) y **2.** ($E(\varepsilon|X) = 0$).
- $E(\hat{\beta}_0) = \beta_0$ (Véase ejercicio)
- $E(\hat{\beta}_1) = \beta_1$
 - Demostración: debemos probar que $E(\hat{\beta}_1|X) = \beta_1$; después, es inmediato que $E(\hat{\beta}_1) = E_X[E(\hat{\beta}_1|X)] = \beta_1$.
(el carácter estocástico de X nos obliga a utilizar esperanzas condicionales).

En primer lugar, podemos escribir $\hat{\beta}_1$ como

$$\begin{aligned}\hat{\beta}_1 &= \sum_i c_i Y_i = \sum_i c_i (\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 \sum_i c_i + \beta_1 \sum_i c_i X_i + \sum_i c_i \varepsilon_i\end{aligned}$$

- Pero

$$\sum_i c_i = \frac{1}{\sum_i x_i^2} \left(\sum_i x_i \right) = 0 \text{ porque } \sum_i x_i = \sum_i X_i - n\bar{X} = 0$$

$$\sum_i c_i X_i = \sum_i \frac{x_i X_i}{\sum_i x_i^2} = \frac{1}{\sum_i x_i^2} \sum_i x_i^2 = 1$$

- Por tanto,

$$\hat{\beta}_1 = \beta_1 + \sum_i c_i \varepsilon_i$$

y

$$\begin{aligned} E(\hat{\beta}_1 | X) &= \beta_1 + E(\sum_i c_i \varepsilon_i | X) = \beta_1 + \sum_i c_i \underbrace{E(\varepsilon_i | X)}_{= 0} \\ &= \beta_1 \end{aligned}$$

- Recordemos que la propiedad de insesgadez indica que si disponemos de un número infinito de muestras de tamaño n de la misma población y estimamos el mismo modelo con cada una de las muestras:
 - tendremos una distribución de valores estimados de β_j , con una realización numérica distinta para cada muestra,
 - la media de la distribución de dichos valores estimados de β_j , coincidirá con el parámetro poblacional β_j .

- Además de los supuestos **1.** y **2.**, utilizaremos el supuesto **3.**
($V(\varepsilon|X) = \sigma^2$ para todo X).
- $V(\hat{\beta}_0) = (\sum_i X_i^2 / n) V(\hat{\beta}_1)$ (Véase ejercicio)
- $V(\hat{\beta}_1) = \frac{\sigma^2}{n} E\left(\frac{1}{S_x^2}\right)$.

- Demostración:

$$\begin{aligned}V(\widehat{\beta}_1) &= E \left[\left(\widehat{\beta}_1 - \beta_1 \right) \right]^2 = E \left(\sum_i c_i \varepsilon_i \right)^2 = \sum_i E (c_i^2 \varepsilon_i^2) \\&= E \left[\sum_i E (c_i^2 \varepsilon_i^2 | X) \right] = E \left[\sum_i c_i^2 E (\varepsilon_i^2 | X) \right] \\&= \sigma^2 E \left[\sum_i c_i^2 \right] \text{ (por el supuesto 3.)} \\&= \sigma^2 E \left[\sum_i \left(\frac{x_i}{\sum_i x_i^2} \right)^2 \right] = \sigma^2 E \left[\frac{1}{(\sum_i x_i^2)^2} \sum_i x_i^2 \right] \\&= \sigma^2 E \left[\frac{\frac{1}{n}}{\frac{1}{n} (\sum_i x_i^2)} \right] = \frac{\sigma^2}{n} E \left(\frac{1}{S_X^2} \right)\end{aligned}$$

Propiedades de los estimadores MCO

El Teorema de Gauss-Markov

- En el contexto del modelo de regresión lineal, bajo los supuestos **1.** a **3.**, $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores de menor varianza entre los estimadores lineales e insesgados.
(Demostración: Goldberger p. 65-68, para el modelo simple)
- Por tanto, cuando se cumplen los supuestos del modelo clásico, el estimador de MCO es el más **eficiente** dentro de la familia de estimadores lineales e insesgados.

Propiedades de los estimadores MCO

Consistencia de los estimadores MCO

- Los estimadores MCO $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores consistentes de β_0 y β_1 :

$$p \lim_{n \rightarrow \infty} \hat{\beta}_j = \beta_j, \quad j = 0, 1.$$

es decir:

$$\lim_{n \rightarrow \infty} \Pr \left(\left| \hat{\beta}_j - \beta_j \right| < \delta \right) = 1, \quad \forall \delta > 0$$

- Intuición:**

- Los estimadores MCO se obtienen a partir de los análogos muestrales de momentos poblacionales. En concreto, explotan los análogos muestrales de las condiciones de momentos:

$$E(\varepsilon) = 0, \quad C(X, \varepsilon) = 0, \quad (*)$$

es decir:

$$\frac{1}{n} \sum_i \hat{\varepsilon}_i = 0, \quad \frac{1}{n} \sum_i \hat{\varepsilon}_i x_i = 0,$$

Propiedades de los estimadores MCO

Consistencia de los estimadores MCO

- Pero dichos análogos muestrales son funciones de medias muestrales de variables aleatorias, que bajo condiciones bastante generales son estimadores consistentes de sus análogos poblacionales.
- La condición esencial para consistencia es que las condiciones sobre los momentos poblacionales (*) se cumplan.
(Lo que ocurre si se cumplen los supuestos **1.** y **2.** del modelo de regresión)

Estimación de las varianzas

Estimación de varianza

- Las varianzas de los estimadores MCO, $\hat{\beta}_0$ y $\hat{\beta}_1$, dependen de $\sigma^2 = V(\varepsilon) = E(\varepsilon^2)$
- El problema es que los errores ε_i ($i = 1, \dots, n$) son inobservables.
- Una vez estimado el modelo por MCO, observamos los residuos $\hat{\varepsilon}_i$:

$$\begin{aligned}\hat{\varepsilon}_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (\beta_0 + \beta_1 X_i + \varepsilon_i) - \hat{\beta}_0 - \hat{\beta}_1 X_i \\ &= \varepsilon_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) X_i \quad (i = 1, \dots, n)\end{aligned}$$

- Si bien $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $\hat{\varepsilon}_i \neq \varepsilon_i$. Además, $E(\hat{\varepsilon}_i - \varepsilon_i) \neq 0$.
- Si observáramos, para nuestra muestra de tamaño n , los errores ε_i ($i = 1, \dots, n$), entonces el estimador natural de σ^2 sería el análogo muestral de $E(\varepsilon^2)$, es decir, $\frac{1}{n} \sum_i \varepsilon_i^2$. PERO este estimador no es factible.

Estimación de las varianzas

Estimación de varianza

- Si reemplazamos los errores por sus análogos muestrales, los residuos, podemos calcular como estimador de σ^2 :

$$\tilde{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n}.$$

- Este estimador sí es factible, pero es sesgado. La razón es que, los residuos verifican 2 restricciones lineales, $\frac{1}{n} \sum_i \hat{\varepsilon}_i = 0$ y $\frac{1}{n} \sum_i \hat{\varepsilon}_i x_i = 0$, de manera que sólo hay $(n - 2)$ residuos independientes (lo que se conoce como **grados de libertad**).

Estimación de las varianzas

Estimación de varianza

- Alternativamente, podemos obtener un estimador insesgado (que para n grande es muy similar):

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2}.$$

(Demostración: véase Wooldridge, Teorema 2.3)

- Tanto $\tilde{\sigma}^2$ como $\hat{\sigma}^2$ son estimadores consistentes de σ^2
- En general, para tamaños muestrales moderados, es irrelevante cuál de los dos estimadores utilizar, porque siempre que n no sea muy pequeño, proporcionan estimaciones numéricas muy parecidas.

Estimación de las varianzas

Estimación de las varianzas de los estimadores MCO

- Hemos visto que

- $V(\hat{\beta}_0) = (\sum_i X_i^2 / n) V(\hat{\beta}_1)$

- $V(\hat{\beta}_1) = \frac{\sigma^2}{n} E\left(\frac{1}{S_x^2}\right)$.

- Como estimador de $V(\hat{\beta}_1)$, podemos aproximar $E\left(\frac{1}{S_x^2}\right)$ mediante $\frac{1}{S_x^2}$ así como un estimador consistente de σ^2 :

$$\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{nS_x^2}$$

- Y por tanto, para estimar $V(\hat{\beta}_0)$,

$$\hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum_i X_i^2}{n^2 S_x^2}$$

Medidas de bondad del ajuste

Error estándar de la regresión

- En el caso poblacional, vemos que la función de esperanza condicional $E(Y|X)$ es la proyección lineal de Y dado X . en el sentido de que minimiza $E(\varepsilon^2)$
- Por analogía, en el caso de la estimación MCO a partir de una muestra de los coeficientes del modelo de regresión clásico

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n,$$

donde:

$$E(\varepsilon|X) = 0$$

$$V(\varepsilon|X) = \sigma^2$$

- De forma más conveniente, suele considerarse la raíz cuadrada de $\hat{\sigma}^2$, $\hat{\sigma}$, que se denomina **error estándar de la regresión**, como medida de la bondad del ajuste

Medidas de bondad del ajuste

El coeficiente de determinación

- Una medida más popular de capacidad predictiva del modelo es el R^2 o **coeficiente de determinación**, que se define como

$$R^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i y_i^2},$$

donde $y_i = Y_i - \bar{Y}_i$, $\hat{y}_i = \hat{Y}_i - \bar{Y}_i$, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

(la segunda igualdad es cierta siempre que el modelo tenga término constante)

- El R^2 se interpreta como la proporción de la variación muestral de Y explicada por el modelo. (Véase Goldberger, pp. 82 y 83).
- El R^2 verifica que $0 \leq R^2 \leq 1$.
 - Cuando $R^2 = 0$, el modelo explica el 0% de la variación de Y .
 - Cuando $R^2 = 1$, el modelo explica el 100% de la variación de Y .

Medidas de bondad del ajuste

El coeficiente de determinación

- Puede verse que

$$R^2 = (\hat{\rho}_{Y\hat{Y}})^2 = \left(\frac{S_{Y\hat{Y}}}{S_Y S_{\hat{Y}}} \right)^2$$

es decir: el R^2 es el cuadrado del coeficiente de correlación muestral entre Y_i e \hat{Y}_i .

- En el caso de datos de sección cruzada, no es inusual que el R^2 de una regresión presente valores bajos.
 - Ello implica que el modelo deja sin explicar una proporción importante de la variación de Y .
 - PERO un R^2 bajo no implica necesariamente que las estimaciones son poco útiles o inadecuadas.
- El R^2 puede ser útil para comparar distintos modelos para la misma variable dependiente Y .