

Econometría

El modelo de regresión lineal simple

Alejandro Mosiño - *Universidad de Guanajuato*

v.2026.08.11

Construcción del modelo

La curiosidad

Consideremos la siguiente pregunta de investigación:

¿Cómo se relaciona el ingreso per cápita con la esperanza de vida? En particular, ¿un aumento en el ingreso hace que las personas vivan más?

Datos

Consideremos los siguientes datos sobre la esperanza de vida y el ingreso per cápita:

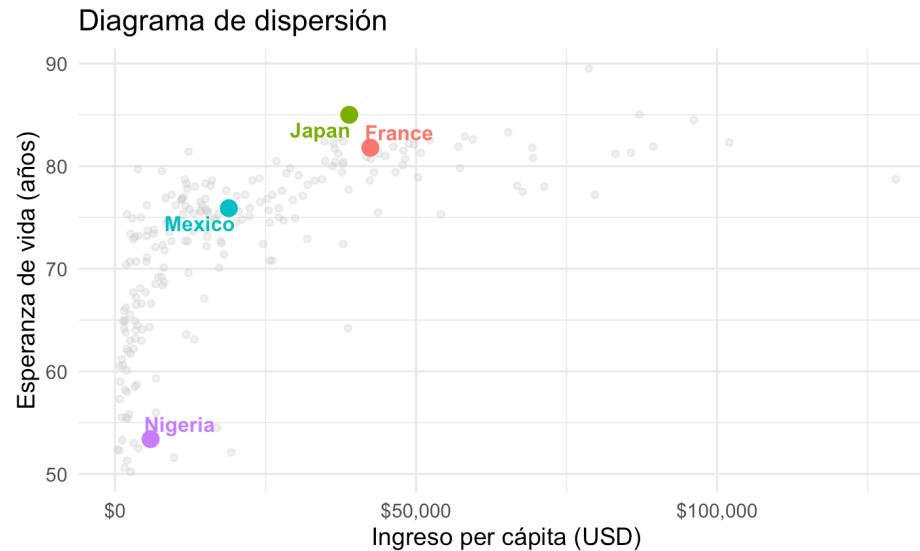
Relación entre ingreso per cápita y esperanza de vida

País	Ingreso per cápita	Esperanza de vida
France	\$42,400	81.8
Japan	\$38,900	85.0
Mexico	\$18,900	75.9
Nigeria	\$5,900	53.4

Fuente: CIA World Factbook.

Diagrama de dispersión (1/2)

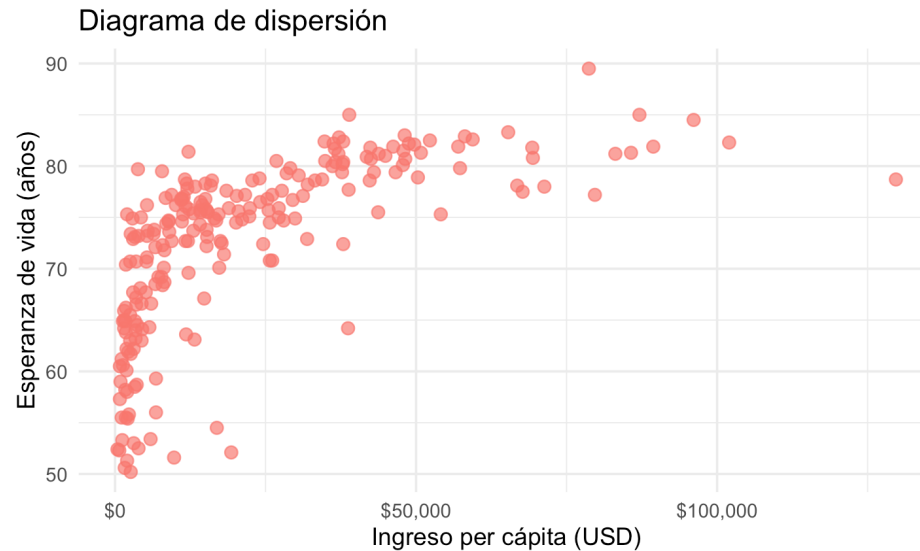
Al representar estos datos en un diagrama de dispersión comienza a aparecer un patrón:



- Los países con mayor ingreso per cápita tienden a presentar una mayor esperanza de vida.
- Los países con menor ingreso per cápita tienden a presentar una menor esperanza de vida.

Diagrama de dispersión (2/2)

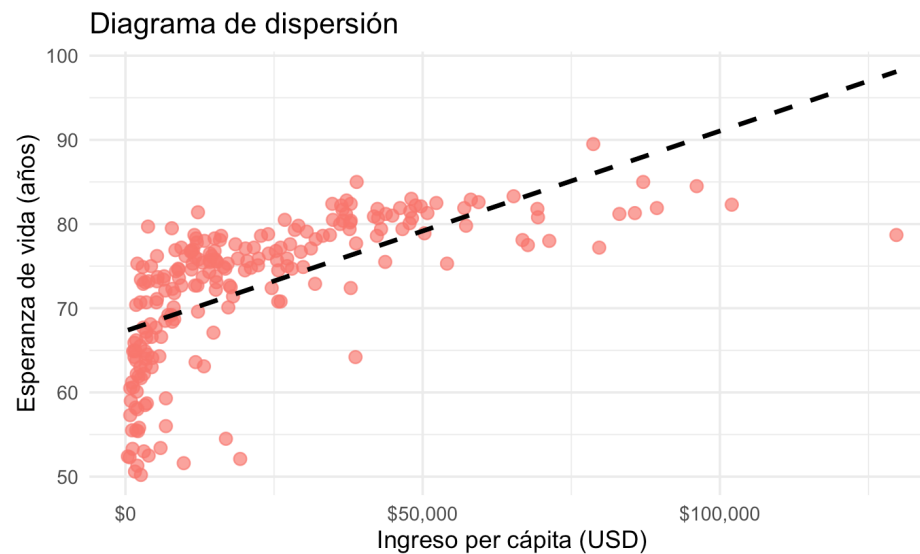
Con solo cuatro observaciones es difícil identificar un patrón general. Para comprender mejor la asociación entre ambas variables utilizaremos información de un conjunto mucho mayor de países.



Nota que, al incorporar más observaciones, la asociación *positiva* entre el ingreso per cápita y la esperanza de vida se hace mucho más evidente.

Línea de tendencia

Para resumir la asociación observada entre el ingreso per cápita y la esperanza de vida, dibujaremos una línea recta sobre el diagrama de dispersión. Esta es la **línea de tendencia**.



Todavía no sabemos cómo construir la línea de tendencia a partir de los datos. *¡Este será uno de los principales objetivos de la regresión lineal!*

Modelo

Si decidimos representar la asociación entre el ingreso per cápita y la esperanza de vida mediante una línea recta (como haremos durante todo este capítulo), entonces el modelo es:

$$y_i = \beta_1 + \beta_2 x_i$$

(¿Recuerdas la fórmula de una línea?)

donde:

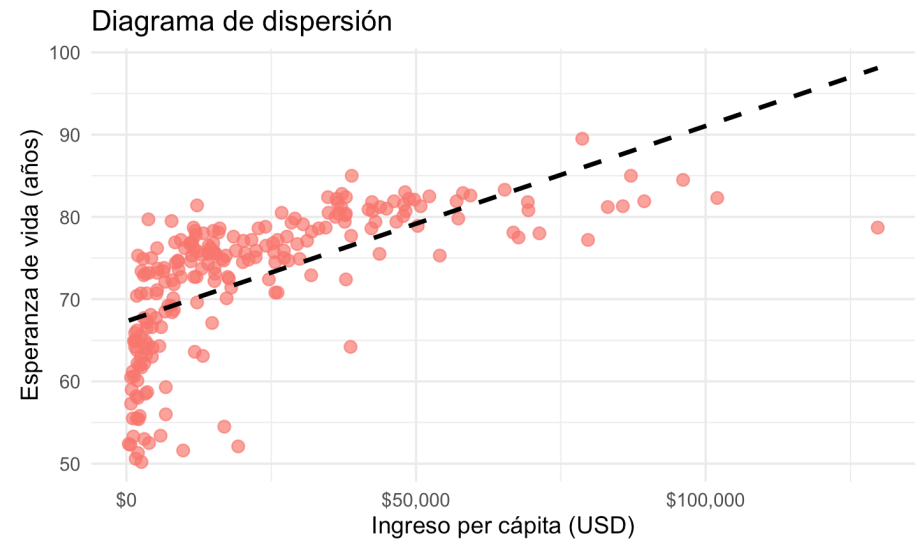
- y_i es la esperanza de vida (variable en el eje y)
- x_i es el ingreso per cápita (variable en el eje x)
- β_1, β_2 son parámetros.

¿Es suficiente una línea recta?

Nota.

La línea de tendencia recta es una buena aproximación, pero todavía presenta limitaciones:

- Esta supone que la asociación entre el ingreso per cápita y la esperanza de vida es *exactamente lineal*.
- Sin embargo, si esto fuera cierto, *todos los puntos en el diagrama de dispersión deberían estar sobre ella*.
- Como este no es el caso, ¿cómo incorporamos esas diferencias al modelo?



El modelo de regresión simple

El término de error

Para incorporar al modelo las diferencias entre la línea recta y los valores efectivamente observados, agregamos un **término de error**, también conocido como **perturbación** o **ruido**, u_i . El modelo tiene ahora dos componentes:

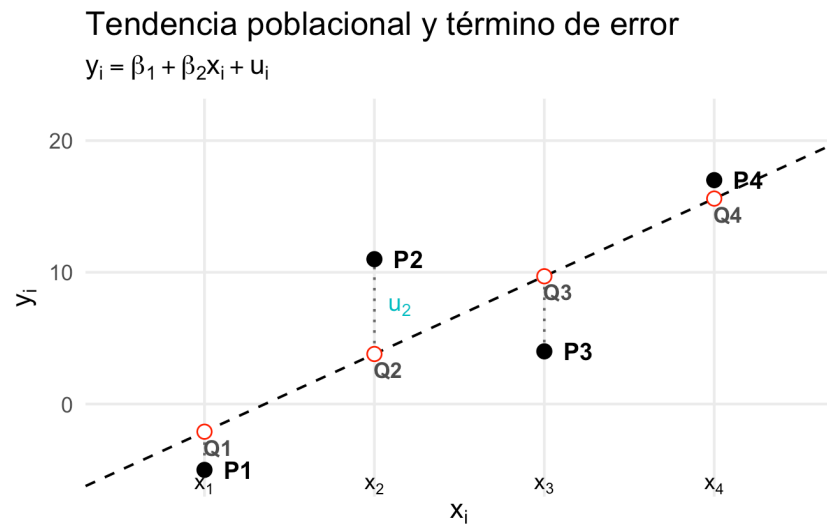
$$y_i = \beta_1 + \beta_2 x_i + u_i$$

1. El elemento $\beta_1 + \beta_2 x_i$ es el *componente sistemático* (parte no aleatoria o determinista).
2. El elemento u_i es el *componente aleatorio*.

Ahora el modelo reconoce explícitamente que x_i no explica por sí sola todas las diferencias observadas en y_i .

Componente sistemático y el término de error

Si conociéramos los *valores verdaderos* de β_1 y β_2 , podríamos localizar exactamente los puntos $Q_1 - Q_4$, que representan la parte sistemática del modelo. La línea $Q_1 - Q_4$ puede pensarse como la **tendencia poblacional**.



El problema es que en la práctica solo observamos los puntos $P_1 - P_4$. Por lo tanto, *dado que los u_i son desconocidos*, no podemos saber con certeza la localización de los puntos $Q_1 - Q_4$.

En realidad sólo Dios sabe dónde se encuentran los puntos $Q_1 - Q_4$.

¿Por qué existe u_i ?

El término u_i recoge todos los factores que afectan a y_i y que no están representados explícitamente por x_i . Por ejemplo:

- Variables explicativas no incluidas;
- Características no observables;
- Errores de medición;
- Choques aleatorios;
- Aproximaciones derivadas de la forma funcional utilizada.

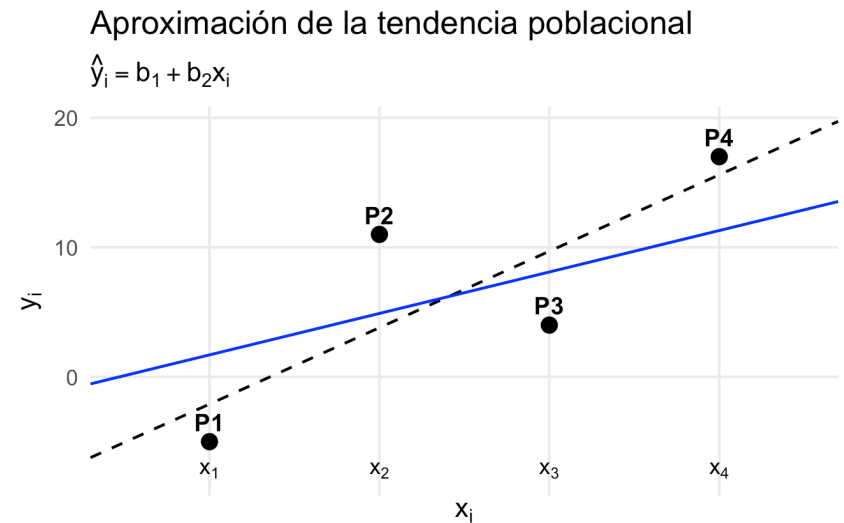
¿Cómo aproximamos la tendencia poblacional? (1/3)

Como no conocemos β_1 y β_2 , debemos estimarlos a partir de los puntos que sí observamos. ¿Cómo podemos hacerlo?

Una opción es identificar los puntos $P_1 - P_4$ en la gráfica y trazar la línea *a ojo de buen cubero*.

En este ejemplo, los valores estimados para β_1 y β_2 son b_1 y b_2 , respectivamente. El valor estimado de y_i se calcula entonces como:

$$\hat{y}_i = b_1 + b_2 x_i$$



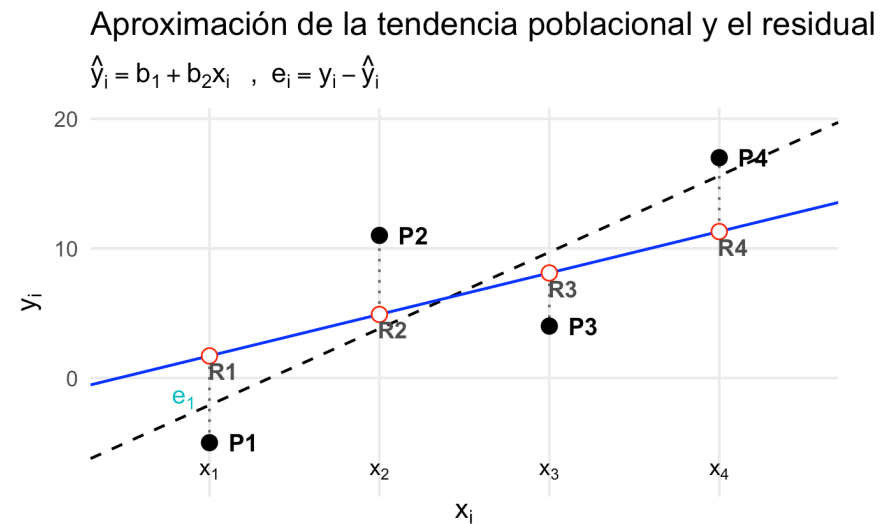
¿Cómo aproximamos la tendencia poblacional? (2/3)

Podemos usar técnicas más sofisticadas para encontrar b_1 y b_2 . Estas requieren el cálculo del (de los) **residual** (**residuales**) de estimación, e_i .

Tenemos que:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - b_1 - b_2 x_i \end{aligned}$$

Nota. El error u_i mide la distancia respecto de la tendencia poblacional; el residual e_i mide la distancia respecto de la **tendencia estimada**.



Nuestro objetivo será encontrar b_1 y b_2 de manera que, en conjunto, los residuales sean lo más pequeños posible.

¿Cómo aproximamos la tendencia poblacional? (3/3)

Ejercicio

Una primera posibilidad sería seleccionar los valores de b_1 y b_2 que hagan que la suma de los residuales se minimice:

$$\begin{aligned}(b_1, b_2) &= \arg \min \left\{ \sum_{i=1}^n e_i \right\} \\ &= \arg \min \left\{ \sum_{i=1}^n (y_i - b_1 - b_2 x_i) \right\}\end{aligned}$$

donde n es el número de observaciones. ¿Crees que este criterio funciona? Explica y comenta.

El método de los mínimos cuadrados ordinarios

El método de los mínimos cuadrados ordinarios

El criterio anterior presenta un problema: los residuales positivos y negativos pueden cancelarse. Entonces, necesitamos una forma de medir la magnitud de los residuales que evite esta cancelación.

El método de **mínimos cuadrados ordinarios** (MCO u OLS, por sus siglas en inglés) propone una solución sencilla: elevar los residuales al cuadrado y elegir la recta que minimiza su suma. Entonces:

$$(b_1, b_2) = \arg \min RSS$$

donde el **RSS** (Residual Sum of Squares) es la suma del cuadrado de los residuales:

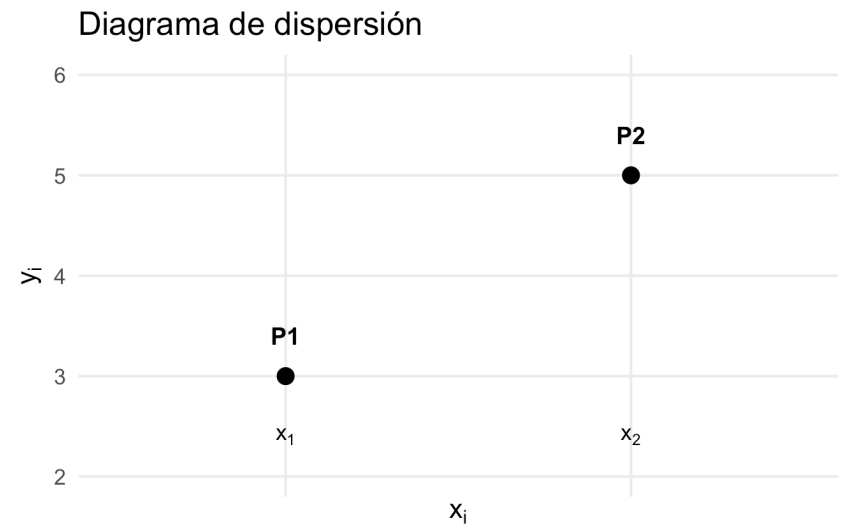
$$\begin{aligned} RSS &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2 \end{aligned}$$

MCO: dos observaciones

Ejercicio

Considera los siguientes datos:

x	y
1	3
2	5



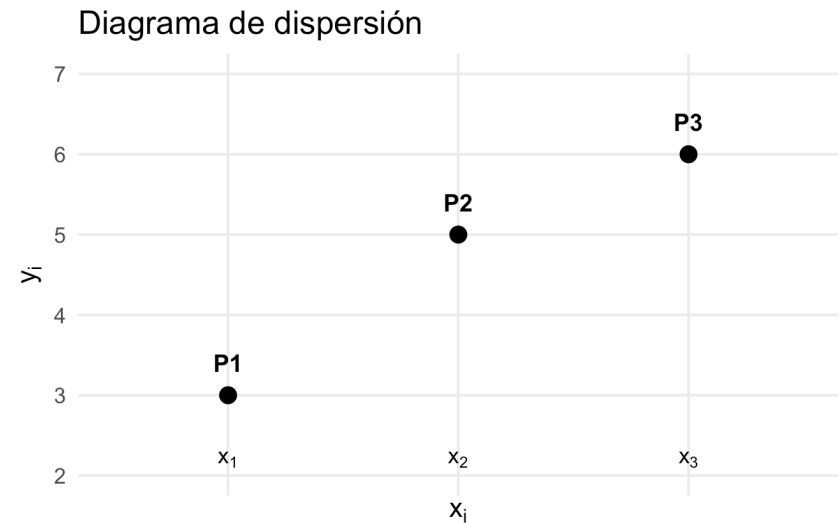
Encuentra los estimadores de mínimos cuadrados ordinarios b_1 y b_2 . Comenta tus resultados.

MCO: Cuando ningún ajuste es perfecto

Ejercicio

Considera los siguientes datos:

x	y
1	3
2	5
3	6



Encuentra los estimadores de mínimos cuadrados ordinarios b_1 y b_2 . Comenta tus resultados. En particular ¿qué cambió respecto al ejemplo anterior?

MCO: Caso general (1/3)

Para n observaciones, el problema:

$$\min_{b_1, b_2} \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$$

lleva a las siguientes **ecuaciones normales**:

$$nb_1 + b_2 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$$

$$b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0$$

MCO: Caso general (2/3)

Nota que las ecuaciones normales de la regresión no son otra cosa que las condiciones de primer orden del problema de minimización. Resolviendo estas, obtenemos los valores de b_1 y b_2 que minimizan el RSS:

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\text{Cov}(x, y)}{\text{Var}(x)} \end{aligned}$$

Y:

$$b_1 = \bar{y} - b_2 \bar{x}$$

MCO: Caso general (3/3)

Ejercicio

Considera los siguientes datos que miden la relación entre el promedio escolar y el ingreso familiar:

Relación entre promedio escolar e ingreso familiar (miles de pesos)

Promedio	Ingreso_familiar
8	21
6	15
7	15
4	9
6	12
7	18
5	6
5	12

Encuentra los estimadores de mínimos cuadrados ordinarios b_1 y b_2 . Comenta tus resultados. ¿Cuál sería la intuición económica de la recta estimada?

Fórmulas alternativas para b_2

Ejercicio*

La expresión anterior no es la única formas de escribir b_2 . Las siguientes representaciones serán útiles para estudiar posteriormente las propiedades del estimador:

$$b_2 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Y:

$$b_2 = \sum_{i=1}^n w_i y_i;$$
$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Propiedades algebraicas de MCO

Dos descomposiciones de la variable dependiente (1/3)

Podemos descomponer la variable dependiente, y_i , de dos maneras alternativas:

- Modelo poblacional: descomposición de la variable dependiente en un *componente sistemático y el término de error*.
- Modelo estimado: descomposición de la variable dependiente en su *valor estimado y el residual*.

Modelo poblacional

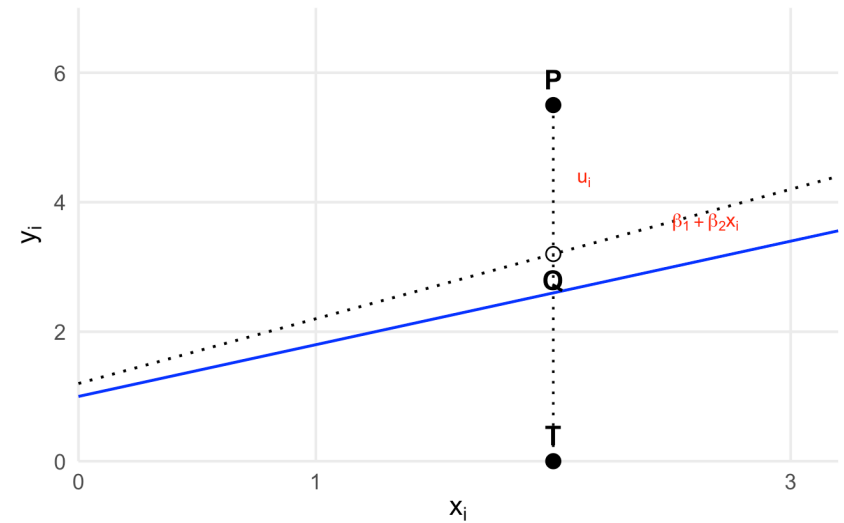
Descomposición de la variable dependiente en un *componente sistemático y el término de error*.

Tenemos:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Nota que, en esta descomposición, cada punto en la gráfica resulta de:

$$P_i = Q_i + u_i$$



Modelo estimado

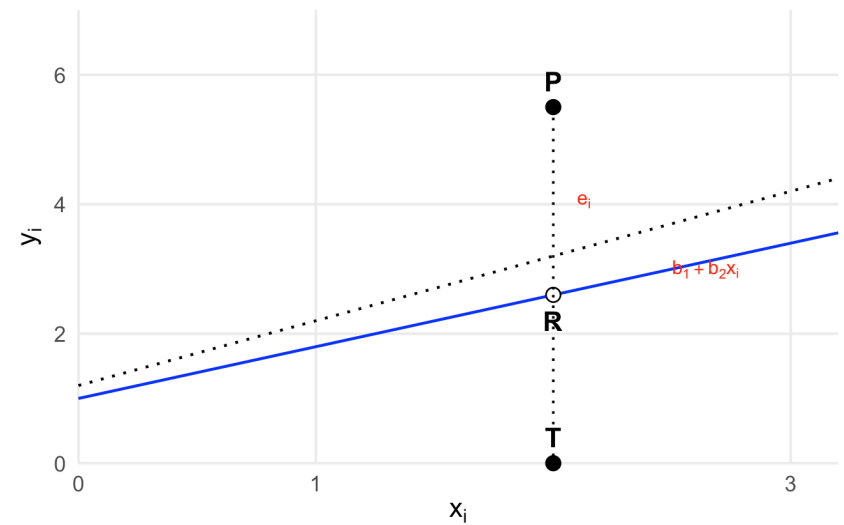
Descomposición de la variable dependiente en su *valor estimado y el residual*.

Tenemos:

$$\begin{aligned}y_i &= \hat{y}_i + e_i \\ &= b_1 + b_2 x_i + e_i\end{aligned}$$

Nota que, en esta descomposición, cada punto en la gráfica resulta de:

$$P_i = R_i + e_i$$



Propiedades algebraicas de MCO: residuales

Podemos demostrar que:

1. El valor promedio de los residuales es cero:

$$\frac{1}{n} \sum_{i=1}^n e_i = 0$$

2. La correlación muestral entre x_i y el residual es cero:

$$\text{Cov}(x, e) = 0$$

Propiedades algebraicas de MCO: valores estimados

Ejercicio

Demuestra que:

1. El valor estimado promedio de la variable dependiente es igual al valor promedio observado:

$$\overline{\hat{y}} = \bar{y}$$

2. La correlación muestral entre $\hat{y}_i$ y e_i es cero:

$$\text{Cov}(\hat{y}, e) = 0$$

Interpretación y transformaciones

Interpretación de los coeficientes estimados (1/3)

Hasta ahora hemos aprendido a construir la ecuación estimada. Pero el análisis no termina aquí. Aun debemos interpretar qué nos dice la estimación acerca del fenómeno que estamos estudiando.

En general:

$$\hat{y}_i = b_1 + b_2 x_i$$

- Si x_i aumenta en una unidad, el valor estimado de y_i cambia en b_2 unidades (signo positivo o negativo según el caso).
- El valor estimado de y_i es igual a b_1 cuando $x_i = 0$.

Interpretación de los coeficientes estimados (2/3)

Ejemplo

Relación entre ingreso per cápita y esperanza de vida

Término	Coeficiente	Desv. Estándar	Estadístico t	Valor p
Constante	67.299334	0.612710	109.839	<0.001
Ingreso per cápita	0.000238	0.000019	12.212	<0.001

$$\begin{aligned}\widehat{\text{Esperanza de vida}}_i &= b_1 + b_2 \text{Ingreso per cápita}_i \\ &= 67.35 + 0.00024 \text{Ingreso per cápita}_i\end{aligned}$$

¿Cómo lees este resultado?

Interpretación de los coeficientes estimados (3/3)

De igual forma, los coeficientes estimados pueden utilizarse para hacer predicciones:

$$\begin{aligned}\mathbb{E}(y_i \mid x_i = x_0) &= \hat{y}_0 \\ &= b_1 + b_2 x_0\end{aligned}$$

Donde x_0 es algún valor de x .

¿Cuál es la esperanza de vida para un país cuyo ingreso per capita es de \$18,000?

Transformaciones de escala

Ejercicio*

Supongamos que el modelo es:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

¿Qué pasa con los coeficientes estimados, b_1 y b_2 , si cambias las unidades de medición? Específicamente, explica qué pasa si regresamos el modelo:

$$\begin{aligned} y_i^* &= \beta_1^* + \beta_2^* x_i + u_i^* \\ y_i^* &= \lambda_1 + \lambda_2 y_i \end{aligned}$$

Transformaciones de centrado

Ejercicio*

Supongamos que el modelo es:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Explica qué pasa con los coeficientes estimados si regresamos el modelo:

$$y_i = \beta_1^* + \beta_2^* x_i^* + u_i^*$$
$$x_i^* = x_i - \bar{x}$$

Propiedades de los coeficientes estimados

¿Qué tan buenos son los estimadores de MCO?

Hasta ahora hemos aprendido a construir e interpretar los estimadores b_1 y b_2 . Pero... ¿Podemos confiar en ellos?

Para responder a esta pregunta, debemos estudiar sus *propiedades matemáticas*.

¿Qué necesitamos suponer? (1/4)

Las propiedades de los estimadores de MCO dependen de una serie de supuestos sobre el modelo. Estos supuestos son los siguientes:

1. El modelo es lineal en sus parámetros y está correctamente especificado:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

¿Qué necesitamos suponer? (2/4)

Los siguientes son los supuestos relacionados con el *término de error*, u_i . Estos se conocen como **supuestos de Gauss-Markov**:

2. El valor esperado del error es igual a cero para todas las observaciones:

$$\mathbb{E}(u_i) = 0; \forall i$$

3. La varianza del error es constante para todas las observaciones (este es el supuesto de **homoscedasticidad**):

$$\begin{aligned}\mathbb{V}ar(u_i) &= \mathbb{E}(u_i^2) \\ &= \sigma_i^2 \\ &= \sigma^2; \forall i\end{aligned}$$

¿Qué necesitamos suponer? (3/4)

4. La covarianza entre los errores de cualesquiera dos observaciones es cero. Este es el supuesto de la **no autocorrelación**:

$$\begin{aligned}\text{Cov}(u_i, u_j) &= \mathbb{E}(u_i u_j) \\ &= 0; \quad \forall i, j\end{aligned}$$

5. Las x_i y las u_i *ortogonales* (las x_i son **exógenas**). Este supuesto viene en dos versiones:

- Versión fuerte: Las x_i son *deterministas*.
- Versión débil: Las x_i son estocásticas, pero siguen una distribución independiente de la distribución de las u_i .

¿Qué necesitamos suponer? (4/4)

6. (*Conveniencia*). Finalmente, se supone que el error, u_i , sigue una distribución normal.

Entonces, si se cumplen todos los supuestos sobre el error, tenemos:

$$u_i \sim iid \square(0, \sigma^2)$$

Componentes de los coeficientes estimados (1/3)

Recordemos que la variable y_i puede descomponerse en: 1) Un componente aleatorio y 2) Un componente determinista. Es de esperar que los coeficientes estimados admitan también la misma descomposición. Por ejemplo, para β_2 podemos mostrar que:

$$b_2 = \beta_2 + \frac{\text{Cov}(x, u)}{\text{Var}(x)}$$

1. El elemento β_2 es la parte sistemática (determinista).

2. El elemento $\frac{\text{Cov}(x, u)}{\text{Var}(x)}$ es la parte aleatoria.

Componentes de los coeficientes estimados (2/3)

Ejercicio

Muestra que:

$$b_2 = \beta_2 + \sum_{i=1}^n w_i u_i$$

donde, igual que antes:

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Componentes de los coeficientes estimados (2/3)

Ejercicio

Muestra que:

$$\begin{aligned}\sum_{i=1}^n w_i &= 0 \\ \sum_{i=1}^n w_i^2 &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \sum_{i=1}^n w_i x_i &= 1\end{aligned}$$

¿Qué significa que un estimador sea bueno?

Un buen estimador debería cumplir tres propiedades fundamentales:

- **Insesgadez**. Esto es, *en promedio*, el estimador es igual a su valor verdadero. En el caso de los estimadores de MCO, esto significa que:

$$\mathbb{E}(b_1) = \beta_1; \quad \mathbb{E}(b_2) = \beta_2$$

- **Eficiencia**. Buscamos los estimadores con la mínima varianza.
- **Consistencia**. Los estimadores convergen en probabilidad a sus valores verdaderos cuando $n \rightarrow \infty$.

Los estimadores de MCO son insesgados

Ejercicio

Demuestra que:

$$\mathbb{E}(b_1) = \beta_1$$

$$\mathbb{E}(b_2) = \beta_2$$

Los estimadores de MCO son eficientes (1/5)

Para saber si los estimadores de MCO son eficientes, primero tenemos que calcular su varianza:

$$\mathbb{V}ar(b_1) = \mathbb{E}([b_1 - \mathbb{E}(b_1)]^2)$$

$$\mathbb{V}ar(b_2) = \mathbb{E}([b_2 - \mathbb{E}(b_2)]^2)$$

Los estimadores de MCO son eficientes (2/5)

Ejercicio

Muestra que:

$$\mathbb{V}ar(b_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Luego muestra que*:

$$\mathbb{V}ar(b_1) = \sigma^2 \left[\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right]; \quad \mathbb{C}ov(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

Los estimadores de MCO son eficientes (3/5)

Para demostrar que los estimadores de MCO tienen la varianza más pequeña, definimos toda la gama posible de estimadores lineales insesgados para β_2 :

$$\tilde{b}_2 = \sum_{i=1}^n c_i y_i$$

Nota que esta gama incluye a los estimadores de MCO en tanto que:

$$c_i = w_i$$

Los estimadores de MCO son eficientes (4/5)

Ejercicio

Muestra que la gama de estimadores que hemos definido es insesgada en tanto que:

$$\sum_{i=1}^n c_i = 0; \quad \sum_{i=1}^n c_i x_i = 1$$

Muestra que, si los estimadores son insesgados:

$$\mathbb{V}ar(\tilde{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2 \sum_{i=1}^n (c_i - w_i)^2$$

Los estimadores de MCO son eficientes (5/5)

Concluimos que los estimadores de MCO son eficientes dado que:

$$\begin{aligned}\mathbb{V}ar(\tilde{\beta}_2) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2 \sum_{i=1}^n (c_i - w_i)^2 \\ &= \mathbb{V}ar(b_2) + \sigma^2 \sum_{i=1}^n (c_i - w_i)^2 \\ &> \mathbb{V}ar(b_2)\end{aligned}$$

Teorema de Gauss–Markov

Las propiedades anteriores se resumen en uno de los resultados fundamentales de la econometría clásica:

Si los supuestos 1–5 del modelo de regresión simple se satisfacen, los estimadores de MCO, b_1 y b_2 , son MELI (Mejores Estimadores Lineales Insesgados).

Nota. Los estimadores son BLUE (*Best Linear Unbiased Estimators*). Este es el conocido Teorema de Gauss - Markov.

Los estimadores son consistentes (1/3)

A diferencia de la insesgadez y la eficiencia, la consistencia describe el comportamiento de los estimadores cuando el tamaño de muestra crece indefinidamente. Para saber si los estimadores son consistentes, debemos recordar primero algunas de las reglas de los límites en probabilidad, **plim**. Sean $\mathbf{x}$ y $\mathbf{y}$ dos series de valores:

1. Si $\mathbf{z} = \mathbf{x}/\mathbf{y}$:

$$\text{plim}(\mathbf{z}) = \frac{\text{plim}(\mathbf{x})}{\text{plim}(\mathbf{y})}$$

2. Si $\mathbf{z} = \mathbf{x}\mathbf{y}$:

$$\text{plim}(\mathbf{z}) = \text{plim}(\mathbf{x}) \cdot \text{plim}(\mathbf{y})$$

Los estimadores son consistentes (2/3)

3. Si $z = x + y$:

$$\text{plim}(z) = \text{plim}(x) + \text{plim}(y)$$

4. Si a es constante:

$$\text{plim}(a) = a$$

5. **Teorema de Slutsky**: si $g(\cdot)$ es continua:

$$\text{plim}(g(x)) = g(\text{plim}(x))$$

Los estimadores son consistentes (3/3)

Ejercicio

Utiliza las reglas anteriores y el hecho que los estimadores para la varianza y covarianza son ambos consistentes, para mostrar que:

$$\text{plim}(b_1) = \beta_1$$

$$\text{plim}(b_2) = \beta_2$$

Distribución de los estimadores de MCO

Ejercicio*

Muestra que si el supuesto 6 del modelo de regresión simple se satisface, entonces:

$$b_1 \sim \square(\beta_1, \text{var}(b_1))$$

$$b_2 \sim \square(\beta_2, \text{var}(b_2))$$

Estimadores para las varianzas de b_1 y b_2

Estimadores para las varianzas de b_1 y b_2 (1/4)

Nota.

Las expresiones obtenidas para $\mathbb{V}ar(b_1)$ y $\mathbb{V}ar(b_2)$ dependen de σ^2 , la varianza del término de error. Naturalmente, como u_i es desconocido, también lo será σ^2 .

Se puede demostrar que un *estimador insesgado de la varianza de los errores* es:

$$\begin{aligned}s^2 &= \frac{1}{n-2} \sum_{i=1}^n e_i^2 \\ &= \frac{RSS}{n-2}\end{aligned}$$

¿Porqué no utilizar simplemente la varianza de los residuales como estimador de σ^2 ?

Estimadores para las varianzas de b_1 y b_2 (2/4)

Ejercicio*

Encontrar un estimador insesgado para σ^2 .

Pista:

- Demuestra primero que los residuales pueden calcularse como:

$$e_i = (x_i - \bar{x})(\beta_2 - b_2) + (u_i - \bar{u})$$

- Calcula luego:

$$\mathbb{E} \left(\sum_{i=1}^n e_i^2 \right)$$

Estimadores para las varianzas de b_1 y b_2 (3/4)

Una vez estimada la varianza de los errores, podemos calcular:

$$\widehat{\text{Var}}(b_2) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\widehat{\text{Var}}(b_1) = s^2 \left[\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

Estimadores para las varianzas de b_1 y b_2 (4/4)

Ejemplo

Relación entre ingreso per cápita y esperanza de vida

Término	Coeficiente	Desv. Estándar	Estadístico t	Valor p
Constante	67.299334	0.612710	109.839	<0.001
Ingreso per cápita	0.000238	0.000019	12.212	<0.001

Nota: La mayoría de los software de econometría calculan la desviación estándar estimada, que no es más que la raíz cuadrada de las fórmulas presentadas anteriormente.

Bondad de ajuste: R^2

¿Qué tan bien explica el modelo los datos?

Hasta ahora hemos aprendido a construir una regresión e interpretar sus coeficientes, pero todavía queda una pregunta importante:

¿Qué tan bien se ajusta el modelo a los datos observados?

Suma total de los cuadrados

¿Qué tan bien se ajusta el modelo a los datos observados?

Para responder, calculamos primero la **Suma total de cuadrados (TSS)**:

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

La TSS mide la *variabilidad total* de la variable dependiente. En particular, suma los cuadrados de las desviaciones de cada observación respecto de su media.

El TSS Representa toda la *variabilidad que existe en los datos antes de estimar cualquier modelo*.

Suma de cuadrados explicados y RSS (1/2)

Recordemos que:

$$y_i = \hat{y}_i + e_i$$

A partir de esta, podemos mostrar que:

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

Esto es, la suma total de cuadrados (TSS) se descompone en la **suma de cuadrados explicados (ESS)** y la suma del cuadrado de los residuales (RSS):

$$TSS = ESS + RSS$$

Suma de cuadrados explicados y RSS (2/2)

$$TSS = ESS + RSS$$

- La ESS mide la *variabilidad explicada* por el modelo. Representa la *parte de la variabilidad total que la regresión logra explicar mediante las variables incluidas en el modelo*.
- La RSS mide la *variabilidad que el modelo no logra explicar*. Representa la *parte de la variabilidad total que permanece sin explicar después de estimar la regresión*.

Análisis de varianza, la R^2 (1/4)

La R^2 , o **coeficiente de determinación**, mide qué proporción de la variabilidad observada en y_i logra explicar el modelo estimado. Es decir:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

Nota que también es posible calcular la R^2 mediante la fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Análisis de varianza, la R^2 (2/4)

A partir de su definición, es claro que:

$$0 \leq R^2 \leq 1$$

Tenemos que:

- Cuando la R^2 se acerca a 1, más de la variabilidad de y_i se explica por el modelo. Entonces, *a mayor R^2 mejor ajuste*.
- Cuando la R^2 se acerca a 0, menos de la variabilidad de y_i se explica por el modelo (más de la variabilidad de y_i es inexplicable). Entonces, *a menor R^2 peor ajuste*.

Nota. Una R^2 no implica, por sí sola, que la relación sea causal ni que el modelo sea correcto. ¿O tú qué piensas?

Análisis de varianza, la R^2 (3/4)

Ejemplo

Para el modelo de la esperanza de vida y el ingreso per capita tenemos:

Media de la vble. dep.	72.515247	D.T. de la vble. dep.	8.471071
Suma de cuad. residuos	9,511.58	D.T. de la regresión	6.560400
R-cuadrado	0.402933	R-cuadrado corregido	0.400231
F(1, 221)	149.142563	Valor p (de F)	1.49e-26
Log-verosimilitud	-734.8933	Criterio de Akaike	1,475.787
Criterio de Schwarz	1,486.008	Crit. de Hannan-Quinn	1,476.537

¿Cómo interpretas este resultado? En particular, ¿qué tan bien responde este modelo a la pregunta inicial?

Análisis de varianza, la R^2 (4/4)

Ejercicio*

Calcular R^2 para el siguiente conjunto de datos:

x	y
1	3
2	5
3	6

Relación entre la R^2 y el coeficiente de correlación

Ejercicio*

Consideremos el coeficiente de correlación entre la variable dependiente y su valor estimado:

$$r_{y,\hat{y}} = \frac{\text{Cov}(y, \hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}}$$

Muestra que:

$$r_{y,\hat{y}} = \sqrt{R^2}$$

Resumen

¿Qué hemos aprendido?

¿Cómo se relaciona el ingreso per cápita con la esperanza de vida?

Para responderla aprendimos a:

- Representar la relación mediante un modelo lineal;
- Incorporar un término de error;
- Estimar el componente sistemático mediante mínimos cuadrados ordinarios;
- Demostrar las propiedades de los estimadores;
- Interpretar los coeficientes estimados;
- Evaluar qué tan bien explica el modelo los datos.

¿Qué sigue?

Hasta ahora, hemos aprendido a construir evidencia cuando una variable explicativa es suficiente para representar el fenómeno de interés. En el siguiente capítulo veremos cómo extender ese razonamiento a problemas donde intervienen múltiples variables de manera simultánea.