## 14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

**BREIT** (Brescia Institute of Technology) · Aporta — Grupo Breca

# Clase 9

## Modelo Lineal Simple y Multivariado

*Teoría*

**Módulos oficiales del MicroMaster:**
M8 – Single and Multivariate Linear Models

B.Sc. Cristian Lazo Quispe
clazoq@uni.pe
linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

Material del *Advanced Program in Data Science & Global Skills*, diseñado y desarrollado por **BREIT (Brescia Institute of Technology)**, iniciativa de **Aporta** —plataforma de innovación e impacto social del **Grupo Breca**— en alianza con el **MIT IDSS** (Institute for Data, Systems, and Society).



11 de agosto de 2026

# Índice

# 1. De qué trata esta clase

En la Clase 4 vimos cómo medir si dos variables "se mueven juntas" (covarianza, correlación) y dejamos planteada la *mejor predicción lineal* de $Y$ a partir de $X$. En la Clase 7 estudiamos causalidad con un tratamiento que era, en el fondo, **una moneda** ($T = 1$ tratado, $T = 0$ control). Esta clase da el salto natural: ¿y si la variable explicativa no es una moneda, sino algo **continuo** —años de educación, ingreso del hogar, horas de estudio— que puede tomar todo un rango de valores?

Casi todo lo interesante en ciencias sociales vive en una **distribución conjunta (joint distribution)** de dos o más variables. La pregunta de fondo casi siempre es: *"dado un valor de $X$, ¿cuál es el valor promedio de $Y$?"*. Eso es la **esperanza condicional** $\mathbb{E}[Y \mid X]$. En la Clase 7, $X$ era una moneda y $\mathbb{E}[Y \mid X]$ tenía solo dos valores (media de tratados, media de controles). Aquí $X$ es continua, así que $\mathbb{E}[Y \mid X]$ es *una función* de $X$ —y el modelo lineal la aproxima con una recta.

¿Por qué nos importa esta función? Por tres razones clásicas (Duflo & Ellison):

- **Predicción (prediction):** con $X$ puedo predecir $Y$ (¿cuánto ganará alguien con 16 años de educación?).
- **Causalidad (determining causality):** bajo los supuestos correctos, el coeficiente mide el *efecto* de mover $X$.
- **Entender el mundo (understanding the world better):** cuantificar cómo se asocian las variables que nos interesan.

### El "caballo de batalla": el modelo lineal

El **modelo lineal (linear model)**, estimado por **regresión lineal (linear regression)**, es la herramienta más usada en estadística aplicada y econometría. Es el *workhorse* (caballo de batalla): simple, interpretable y sorprendentemente flexible. Esta clase responde dos preguntas:

1. ¿Cómo escribimos $\mathbb{E}[Y \mid X]$ como una recta y cómo *estimamos* sus parámetros? (**modelo simple**, mínimos cuadrados / OLS).
2. ¿Cómo generalizamos a *varias* variables explicativas a la vez? (**modelo multivariado**, notación matricial).

# 2. El modelo lineal simple

## 2.1. La ecuación del modelo

**Definición 2.1** (Modelo lineal simple (simple linear model))**.** Para $n$ observaciones $i = 1, \ldots, n$, el modelo lineal simple postula

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} \qquad i = 1, \ldots, n.$$

Sus piezas, con su nombre técnico en inglés:

- $Y_i$: la **variable dependiente** (dependent / explained variable, *regressand*). Lo que queremos explicar o predecir.
- $X_i$: la **variable explicativa** (explanatory / independent variable, *regressor*). Lo que usamos para explicar.
- $\beta_0, \beta_1$: los **coeficientes de regresión** (regression coefficients), los *parámetros a estimar*.
- $\varepsilon_i$: el **error** (error term), una variable aleatoria *no observada* que recoge todo lo demás que afecta a $Y_i$.

Son los mismos $\alpha, \beta$ de la Clase 4 (ahora con la notación estándar $\beta_0, \beta_1$). La gran idea: el modelo escribe la **media de $Y$ como función de $X$**. Tomando esperanza y usando $\mathbb{E}[\varepsilon_i] = 0$:

$$\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i.$$

Es decir, **la recta $\beta_0 + \beta_1 X$ *es* la esperanza condicional** $\mathbb{E}[Y \mid X]$. El error $\varepsilon_i$ es la desviación de cada punto respecto a esa recta.

**Interpretación de los coeficientes**

- $\beta_0$ (**intercepto**, intercept): el valor esperado de $Y$ cuando $X = 0$. Es "dónde arranca" la recta. A veces tiene sentido literal, a veces es solo el ancla matemática de la recta.
- $\beta_1$ (**pendiente**, slope): el cambio esperado en $Y$ por *cada unidad adicional* de $X$:

$$\beta_1 = \frac{\Delta\, \mathbb{E}[Y \mid X]}{\Delta X}.$$

  Es "cuánto sube (o baja) $Y$ si $X$ aumenta en 1".

La pendiente $\beta_1$ *lleva unidades*: si $Y$ es salario en soles y $X$ son años de educación, entonces $\beta_1$ está en "soles por año de educación". Traducirla a una frase concreta — "cada año extra de estudio se asocia con $\beta_1$ soles más de salario"— es la mitad del trabajo del analista. Y ojo: "se asocia" no es "causa" (ver el aviso de la Sección 4).

**Ejemplo 2.1** (Lectura de una recta: educación y salario)**.** Supongamos que estimamos, para trabajadores urbanos del Perú,

$$\widehat{\text{salario}} = 600 + 180 \cdot (\text{años de educación}).$$

Entonces $\beta_0 = 600$ soles es el salario predicho con 0 años de educación (extrapolación, tómese con pinzas), y $\beta_1 = 180$ soles/año dice que *cada año adicional de educación se asocia, en promedio, con* 180 *soles más de salario mensual*. Para alguien con 11 años de escolaridad la predicción es $600 + 180 \cdot 11 = 2580$ soles.

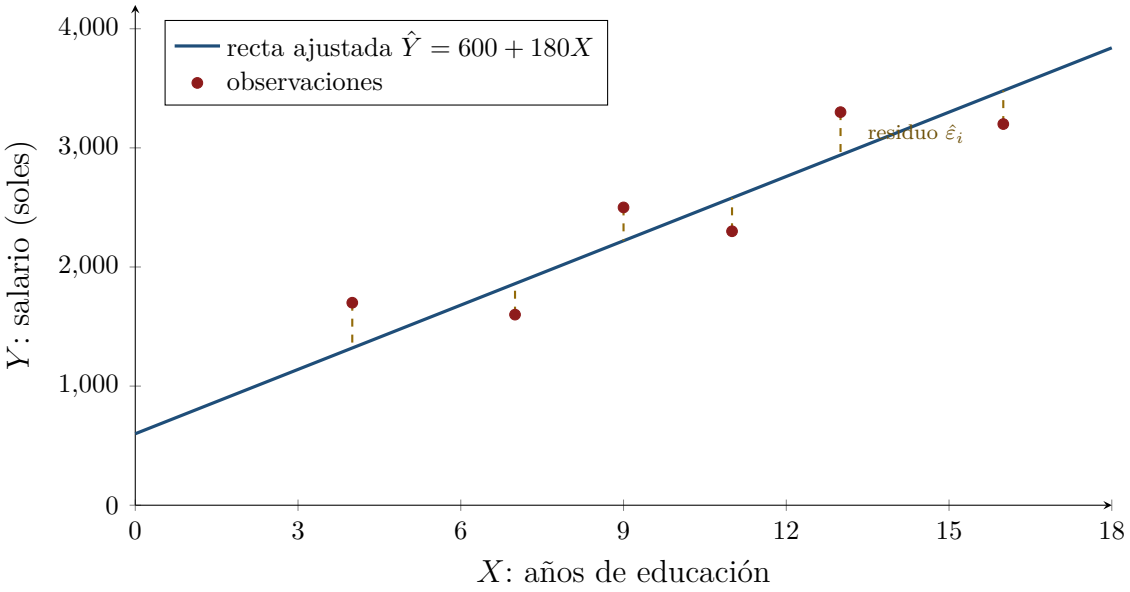


Figura 1: Nube de puntos (scatterplot) con la **recta ajustada** (azul). Cada segmento ámbar punteado es un **residuo** $\hat{\varepsilon}_i$: la distancia *vertical* entre el punto observado $Y_i$ y la recta $\hat{Y}_i$. Mínimos cuadrados (OLS) elige la recta que hace más pequeña la *suma de los cuadrados* de esos segmentos verticales.

# 3. Mínimos cuadrados ordinarios (OLS)

## 3.1. El criterio: minimizar la suma de cuadrados

Tenemos los datos $(X_i, Y_i)$; queremos *elegir* $\beta_0, \beta_1$ que mejor ajusten la recta. "Mejor" necesita una definición. La más usada es **mínimos cuadrados ordinarios (ordinary least squares, OLS)**: elegir la recta que minimiza la suma de los errores *al cuadrado*.

**Definición 3.1** (Estimador de mínimos cuadrados (OLS))**.** Los estimadores $\hat{\beta}_0, \hat{\beta}_1$ son los valores que minimizan la **suma de cuadrados de los residuos**:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2.$$

Podríamos minimizar la suma de *valores absolutos* $\sum |Y_i - \beta_0 - \beta_1 X_i|$ (eso se llama *least absolute deviations*). ¿Por qué se prefieren los cuadrados?

- **Penalizan más los errores grandes:** un residuo de 4 pesa 16, uno de 2 pesa 4. La recta "se esfuerza" por no dejar puntos muy lejos.
- **Son suaves (diferenciables):** se pueden derivar e igualar a cero, lo que da una **fórmula cerrada (closed form)** —no hace falta buscar numéricamente. El valor absoluto tiene un pico en 0 y complica esto.
- **Conectan con la varianza:** minimizar cuadrados es el lenguaje natural de medias y varianzas (y, bajo errores normales, OLS coincide con máxima verosimilitud).

## 3.2. Derivación de las fórmulas

Llamemos $S(\beta_0, \beta_1) = \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$. Para minimizar, igualamos a cero las derivadas parciales (las *ecuaciones normales*):

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_i \Big(Y_i - \beta_0 - \beta_1 X_i\Big) = 0,$$
$$\frac{\partial S}{\partial \beta_1} = -2 \sum_i X_i \Big(Y_i - \beta_0 - \beta_1 X_i\Big) = 0.$$

De la primera ecuación, dividiendo entre $n$: $\bar{Y} - \beta_0 - \beta_1 \bar{X} = 0$, o sea

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

(la recta de OLS **siempre pasa por el punto medio** $(\bar{X}, \bar{Y})$). Sustituyendo $\beta_0$ en la segunda ecuación y despejando $\beta_1$ se obtiene, tras un poco de álgebra,

**Fórmulas cerradas de OLS (simple)**

$$\boxed{\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\mathrm{Var}}(X)}} \qquad \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

La pendiente es la **covarianza muestral de $X$ e $Y$ dividida por la varianza muestral de $X$** —exactamente la "mejor predicción lineal" de la Clase 4, ahora con datos.

La covarianza $\widehat{\mathrm{Cov}}(X, Y)$ mide si $X$ e $Y$ suben juntas; su *signo* es el signo de la pendiente. Pero la covarianza depende de las unidades de $X$. Al dividir por $\widehat{\mathrm{Var}}(X)$ "normalizamos" por cuánto varía $X$ y obtenemos el cambio en $Y$ *por unidad de* $X$. Si $X$ casi no varía, el denominador es chico y la pendiente queda mal determinada (¡es el supuesto de *identificación* de la Sección 5!).

**Ejemplo 3.1** (OLS a mano: horas de estudio y nota)**.** Cinco estudiantes; $X$ = horas de estudio por semana, $Y$ = nota (sobre 10):

| $X_i$ | $Y_i$ | $X_i - \bar{X}$ | $Y_i - \bar{Y}$ | $(X_i - \bar{X})(Y_i - \bar{Y})$ | $(X_i - \bar{X})^2$ |
|---|---|---|---|---|---|
| 1 | 3 | $-2$ | $-2$ | 4 | 4 |
| 2 | 4 | $-1$ | $-1$ | 1 | 1 |
| 3 | 4 | 0 | $-1$ | 0 | 0 |
| 4 | 6 | 1 | 1 | 1 | 1 |
| 5 | 8 | 2 | 3 | 6 | 4 |
| | | | **sumas** | 12 | 10 |

Las medias son $\bar{X} = \frac{1+2+3+4+5}{5} = 3$ y $\bar{Y} = \frac{3+4+4+6+8}{5} = 5$. Entonces

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{12}{10} = 1.2, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 5 - 1.2 \cdot 3 = 1.4.$$

**Recta ajustada:** $\hat{Y} = 1.4 + 1.2\,X$. **Interpretación:** cada hora extra de estudio se asocia, en promedio, con 1.2 puntos más en la nota; con 0 horas el modelo predice 1.4 puntos. (Seguiremos con este mismo ejemplo abajo para residuos y $R^2$.)

# 4. Valores ajustados, residuos y bondad de ajuste ($R^2$)

## 4.1. Valores ajustados y residuos

**Definición 4.1** (Valor ajustado y residuo)**.** Con los estimadores $\hat{\beta}_0, \hat{\beta}_1$:

- el **valor ajustado (fitted value)** es la predicción sobre la recta: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$;
- el **residuo (residual)** es lo que el modelo no logra explicar: $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

El **error** $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ usa los parámetros *verdaderos* y es *inobservable*. El **residuo** $\hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$ usa los *estimados* y sí lo calculamos. El residuo es nuestra "estimación" del error. Por construcción de OLS, los residuos *siempre* suman cero: $\sum_i \hat{\varepsilon}_i = 0$.

## 4.2. Descomposición de la varianza y $R^2$

¿Qué tan bien ajusta la recta? La idea (*analysis of variance*) es repartir la variación total de $Y$ en una parte "explicada" por $X$ y una parte "residual".

## Descomposición de la suma de cuadrados

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{\text{SST: total}} = \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{\text{SSE: explicada}} + \underbrace{\sum_i (Y_i - \hat{Y}_i)^2}_{\text{SSR: residual}} \quad \Longleftrightarrow \quad \boxed{\text{SST} = \text{SSE} + \text{SSR}}$$

- **SST** (total sum of squares): cuánto varía $Y$ en total.
- **SSE** (explained sum of squares): variación captada por la recta.
- **SSR** (residual sum of squares): variación que sobra, $\sum \hat{\varepsilon}_i^2$.

Cuidado al leer libros y software: algunos textos (¡incluido el material original del curso!) usan **SSR** para *Sum of Squared Residuals* (la parte residual, como aquí) y **SSM/SSE** para la parte "del modelo". Otros usan **SSR** para *Sum of Squares due to Regression* (¡la parte explicada!). En estas notas fijamos: **SSR = residual**, **SSE = explicada**, **SST = total**. Lo importante es la idea: total = explicada + residual.

## Coeficiente de determinación $R^2$

$$\boxed{R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}, \qquad 0 \le R^2 \le 1.}$$

$R^2$ es la **proporción de la varianza de $Y$ explicada por el modelo**. $R^2 = 1$ significa ajuste perfecto (todos los puntos sobre la recta, SSR$= 0$); $R^2 = 0$ significa que $X$ no explica nada (la recta es plana en $\bar{Y}$). En regresión *simple*, $R^2 = r_{XY}^2$, el cuadrado de la correlación muestral.

**Ejemplo 4.1** ($R^2$ del ejemplo de horas de estudio)**.** Con $\hat{Y} = 1.4 + 1.2X$, calculamos ajustados y residuos:

| $X_i$ | $Y_i$ | $\hat{Y}_i = 1.4 + 1.2X_i$ | $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ | $\hat{\varepsilon}_i^2$ | $(Y_i - \bar{Y})^2$ |
|---|---|---|---|---|---|
| 1 | 3 | 2.6 | 0.4 | 0.16 | 4 |
| 2 | 4 | 3.8 | 0.2 | 0.04 | 1 |
| 3 | 4 | 5.0 | −1.0 | 1.00 | 1 |
| 4 | 6 | 6.2 | −0.2 | 0.04 | 1 |
| 5 | 8 | 7.4 | 0.6 | 0.36 | 9 |
| | | **sumas** | 0 | 1.6 | 16 |

Entonces SSR $= 1.6$ y SST $= 16$, de modo que

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{1.6}{16} = 1 - 0.1 = 0.9.$$

**Interpretación:** las horas de estudio explican el 90 por ciento de la variación en las notas de esta muestra. (Y, en efecto, la correlación es $r = 0.949$, con $r^2 = 0.9$: coincide con $R^2$.) Nota también que los residuos suman exactamente 0, como debe ser.

Un $R^2$ grande dice que la recta *predice* bien dentro de la muestra, **no** que $X$ *cause* $Y$. La venta de helados predice muy bien los ahogamientos (Clase 7) y eso no hace que el helado ahogue a nadie. Además, $R^2$ *siempre* sube al agregar variables (aunque sean ruido), así que no sirve para comparar modelos con distinto número de regresores —para eso existe el $R^2$ ajustado.

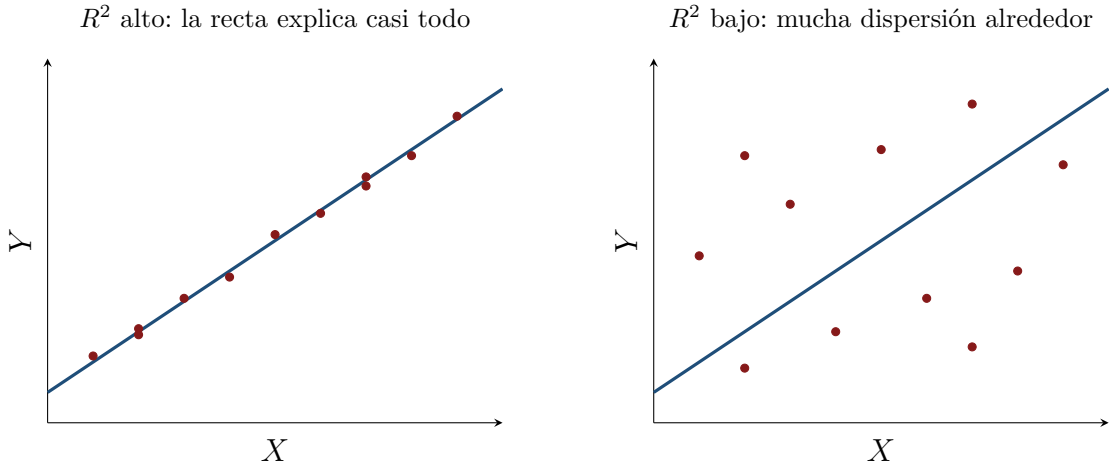


Figura 2: Dos ajustes con la *misma* recta. A la izquierda los puntos caen muy cerca de la recta: residuos pequeños, SSR baja, $R^2$ **alto** (la recta explica casi toda la variación). A la derecha hay mucha dispersión alrededor de la misma recta: residuos grandes, SSR alta, $R^2$ **bajo**. $R^2$ mide qué tan "pegados" están los datos a la recta.

# 5. Varianza del error, supuestos del modelo clásico y Gauss–Markov

## 5.1. Estimar la varianza del error $\sigma^2$

El modelo supone que cada error tiene la misma varianza $\mathrm{Var}(\varepsilon_i) = \sigma^2$. No la conocemos, así que la estimamos con los residuos:

**Estimador insesgado de $\sigma^2$**

$$s^2 = \frac{\mathrm{SSR}}{n-2} = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2}$$

Al estimar la media muestral dividíamos entre $n-1$ porque "gastamos" un grado de libertad estimando $\bar{X}$. Aquí estimamos *dos* parámetros ($\hat{\beta}_0$ y $\hat{\beta}_1$) antes de calcular los residuos, así que "gastamos" dos grados de libertad: quedan $n-2$. Dividir entre $n-2$ corrige el sesgo y hace que $\mathbb{E}[s^2] = \sigma^2$. En el ejemplo de horas de estudio: $s^2 = 1.6/(5-2) = 0.533$, de modo que $s \approx 0.73$ puntos es la dispersión típica de las notas alrededor de la recta.

## 5.2. Los supuestos del modelo clásico (CLRM)

**Supuestos del modelo lineal clásico (Classical Linear Regression Model)**

(I) $X$ **y** $\varepsilon$ **no correlacionados** (exogeneidad): el regresor no "contiene información" sobre el error. Si falla, $\hat{\beta}_1$ queda sesgado (es el problema de los confusores de la Clase 7).

(II) **Identificación (variación en $X$):** $\frac{1}{n}\sum_i (X_i - \bar{X})^2 > 0$. Necesitamos que $X$ *varíe*; si todos los $X_i$ fueran iguales no podríamos estimar una pendiente (el denominador de $\hat{\beta}_1$ sería 0).

(III) **Media cero del error:** $\mathbb{E}[\varepsilon_i] = 0$. El error no tiene un sesgo sistemático; lo que sobra de la recta se cancela en promedio.

(IV) **Homocedasticidad (homoskedasticity):** $\mathrm{Var}(\varepsilon_i) = \sigma^2$ *igual para todos*. La dispersión alrededor de la recta no cambia con $X$. Lo contrario (*heterocedasticidad*: más ruido para $X$ grandes, forma de embudo) afecta los errores estándar.

(V) **Sin autocorrelación (no serial correlation):** $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ para $i \neq j$. Los errores de distintas observaciones no están relacionados entre sí.

Una versión fuerte que resume (iii)–(v) es: $\varepsilon_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$.

(i) "el regresor juega limpio" (no esconde causas omitidas); (ii) "hay con qué medir la pendiente" ($X$ se mueve); (iii) "el error no empuja a $Y$ en una dirección"; (iv) "el ruido es parejo a lo largo de $X$"; (v) "cada observación aporta información nueva". Cuando se cumplen, OLS no solo es razonable: es *óptimo* en un sentido preciso (Gauss–Markov).

**Teorema 5.1** (Gauss–Markov: OLS es BLUE)**.** *Bajo los supuestos del CLRM (i)–(v), el estimador de OLS es* ***BLUE****: el* Best Linear Unbiased Estimator *—el estimador de* ***mínima varianza*** *dentro de la clase de los estimadores que son* lineales *en $Y$ e* insesgados*. Es decir: entre todas las reglas razonables (lineales e insesgadas), OLS es la más precisa.*

"BLUE" significa: si tu modelo cumple los supuestos, ningún otro estimador lineal e insesgado le gana a OLS en varianza —es el más eficiente de su clase. **No** promete que OLS sea bueno si los supuestos fallan (p. ej. si hay un confusor, OLS será insesgado para una *asociación*, no para un *efecto causal*). Bajo errores normales, además, $\hat{\beta}_0$ y $\hat{\beta}_1$ son normales, lo que habilita intervalos de confianza y pruebas $t$ sobre los coeficientes (tema de la Clase 9).

# 6. El modelo multivariado (multiple regression)

En la vida real $Y$ depende de *muchas* cosas a la vez. El salario depende de la educación $y$ de la experiencia $y$ de la región... La **regresión múltiple** incorpora varios regresores:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Con tantos subíndices, la notación por sumatorias se vuelve incómoda. La solución elegante es la **notación matricial**.

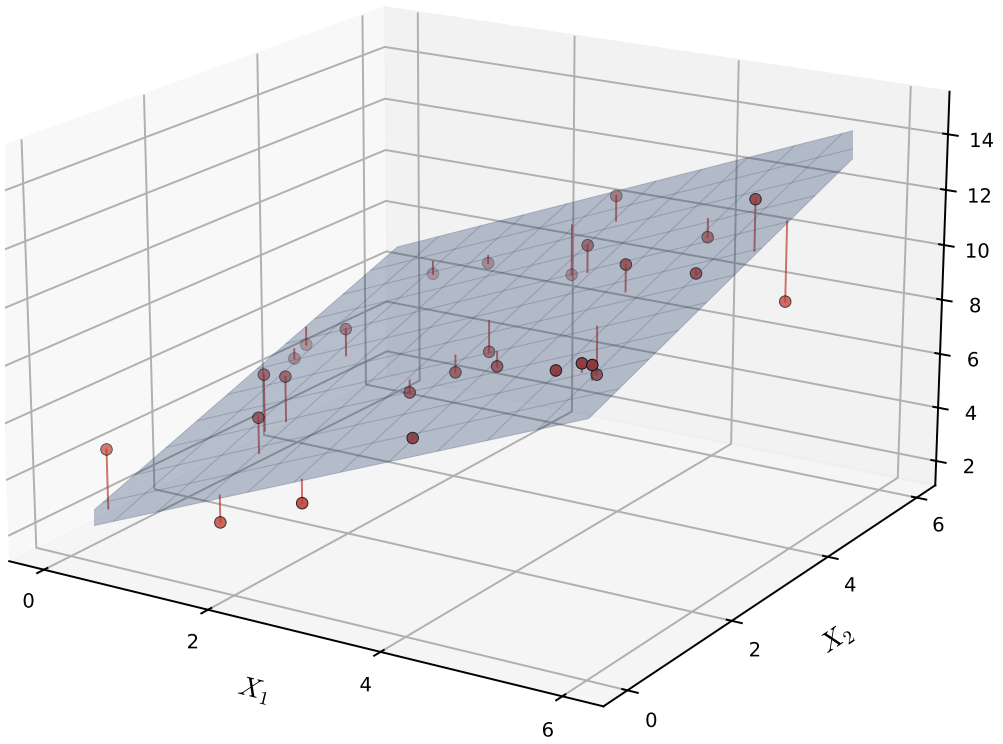


Figura 3: Con *un* predictor el modelo es una **recta**; con *dos* predictores $(X_1, X_2)$ es un **plano** en 3D; con $k$ predictores es un **hiperplano** en $k{+}1$ dimensiones. OLS elige el plano que minimiza la suma de los residuos al cuadrado (segmentos rojos verticales).

## 6.1. Notación matricial

### El modelo lineal en forma matricial

Apilando las $n$ observaciones,

$$\underset{n\times 1}{Y} = \underset{n\times(k+1)}{X}\ \underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\ X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

La primera columna de *unos* en $X$ (la "design matrix") corresponde al intercepto $\beta_0$. Los supuestos del error se escriben compactos: $\mathbb{E}[\varepsilon] = \mathbf{0}$ y $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ (varianza $\sigma^2$ en la diagonal, ceros fuera: homocedasticidad + sin autocorrelación).

### Estimador de OLS multivariado

Minimizar $\sum_i \varepsilon_i^2 = \varepsilon^\top \varepsilon = (Y - X\beta)^\top (Y - X\beta)$ respecto a $\beta$, derivar e igualar a cero da la *ecuación normal* $X^\top X\,\hat\beta = X^\top Y$, cuya solución es

$$\hat\beta = (X^\top X)^{-1} X^\top Y$$

siempre que $X^\top X$ sea **invertible**. (En el caso simple, esta fórmula se reduce exactamente a $\hat\beta_1 = \widehat{\mathrm{Cov}}/\widehat{\mathrm{Var}}$ y $\hat\beta_0 = \bar Y - \hat\beta_1 \bar X$.) Además $\mathbb{E}[\hat\beta] = \beta$ (insesgado) y $\mathrm{Var}(\hat\beta) = \sigma^2 (X^\top X)^{-1}$.

La belleza de $\hat\beta = (X^\top X)^{-1} X^\top Y$ es que *no cambia* con el número de regresores: un regresor o cincuenta, la fórmula es la misma. Es lo que R calcula por dentro cuando escribes `lm(y ~ x1 + x2 + ...)`. El álgebra matricial empaqueta "páginas de cálculos tediosos" en una línea.

Hay una segunda forma de "ver" OLS, y es muy iluminadora. Pensemos en cada *variable* como un vector en $\mathbb{R}^n$ (una coordenada por observación). Las columnas de $X$ generan un **subespacio** (todo lo que el modelo *puede* producir), llamado *espacio columna* $\mathrm{col}(X)$. El vector de datos $Y$ casi nunca cae dentro de ese subespacio. **OLS encuentra el punto del subespacio más cercano a** $Y$: la *proyección ortogonal* $\hat Y = X\hat\beta$. El residuo $\hat\varepsilon = Y - \hat Y$ sale **perpendicular** ($\perp$) al subespacio —por eso $X^\top \hat\varepsilon = \mathbf{0}$ (la ecuación normal). Minimizar $\|Y - X\beta\|^2$ *es* buscar esa proyección.

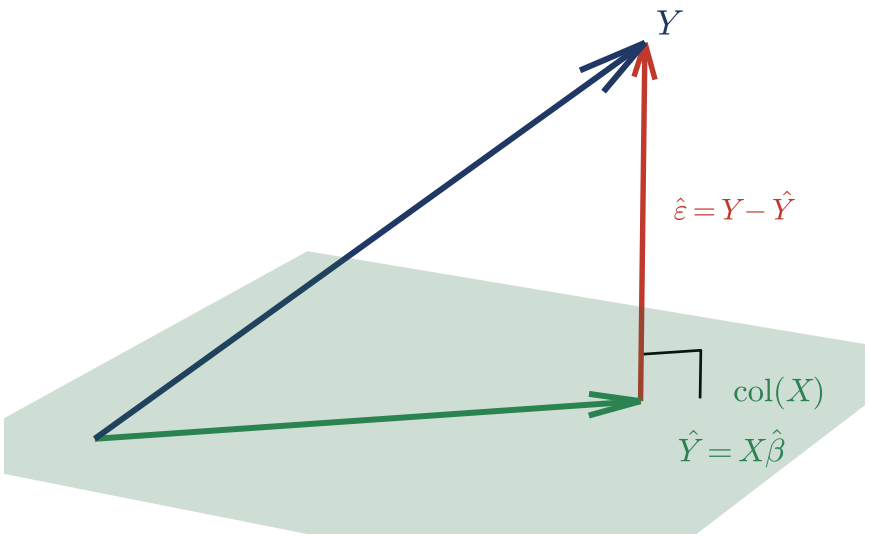


Figura 4: OLS como **proyección ortogonal** de $Y$ sobre el espacio columna $\mathrm{col}(X)$. El valor ajustado $\hat{Y} = X\hat{\beta}$ es la "sombra" de $Y$ sobre el subespacio; el residuo $\hat{\varepsilon} = Y - \hat{Y}$ es perpendicular a él (ángulo recto). Esta es la imagen del "hiperespacio" detrás de la fórmula $\hat{\beta} = (X^\top X)^{-1} X^\top Y$.

## 6.2. Interpretación: *ceteris paribus*

**Cada coeficiente, "manteniendo lo demás constante"**

En la regresión múltiple, $\beta_j$ es el cambio esperado en $Y$ por una unidad más de $X_j$, **manteniendo constantes todos los demás regresores** (*ceteris paribus*, *holding the others fixed*). Es el efecto "parcial" de $X_j$, ya *limpio* de las otras variables incluidas en el modelo.

**Ejemplo 6.1** (Salario, educación y experiencia: el valor de controlar)**.** Estimemos para trabajadores peruanos

$$\widehat{\text{salario}} = 500 + 180 \cdot \text{educación} + 60 \cdot \text{experiencia (soles)}.$$

- $\hat{\beta}_1 = 180$: cada año extra de educación se asocia con 180 soles más *entre personas con la misma experiencia*.
- $\hat{\beta}_2 = 60$: cada año extra de experiencia se asocia con 60 soles más *entre personas con la misma educación*.

La palabra clave es "con la misma". Si solo regresamos salario contra educación, el coeficiente mezclaría el efecto de la educación con el de la experiencia (las personas más educadas suelen, además, diferir en experiencia). Incluir ambas variables "separa" las contribuciones: *controlar* por experiencia es comparar peras con peras.

## 6.3. Multicolinealidad y la "trampa de las dummies"

**Multicolinealidad: cuándo $X^\top X$ no se puede invertir**

La fórmula de OLS exige que $X^\top X$ sea invertible, lo que equivale a que *no haya colinealidad perfecta*: ningún regresor puede ser una combinación lineal exacta de los otros (y necesitamos $n > k+1$ observaciones). Si dos columnas son "la misma información", el sistema no puede separar sus efectos y la inversa no existe.

Si la columna "edad" fuera exactamente "educación + experiencia +6" (porque casi todos entran al colegio a los 6), no hay forma de saber si un cambio en el salario se debe a la edad o a la suma de las otras dos: *matemáticamente* son indistinguibles. El modelo no puede repartir el crédito y se "cae". La colinealidad *alta* (no perfecta) no rompe el modelo, pero infla los errores estándar: los coeficientes quedan imprecisos.

Una **variable dummy (dummy / indicator variable)** vale 1 si la observación pertenece a una categoría y 0 si no. Si tus hogares son de Costa, Sierra o Selva y creas una dummy *para cada* región, las tres suman siempre 1 —justo la columna de unos del intercepto— y hay colinealidad perfecta. **Solución:** omitir una categoría (la "categoría base" o *reference*). Con dos dummies (Sierra, Selva) y Costa como base, cada coeficiente se lee como "cuánto difiere esa región respecto a la Costa". R hace esto automáticamente con `factor()`.

**Ejemplo 6.2** (Dummies regionales: leer respecto a la base)**.** Supongamos $\widehat{\text{gasto}} = 800 + 0.30 \cdot \text{ingreso} - 120 \cdot \text{Sierra} - 90 \cdot \text{Selva}$, con **Costa** como categoría base. Entonces, a igual ingreso, un hogar de la Sierra gasta en promedio 120 soles *menos* que uno de la Costa, y uno de la Selva 90 soles menos. El gasto predicho de un hogar costero con ingreso 2000 es $800 + 0.30 \cdot 2000 = 1400$ soles; el de un hogar serrano con el mismo ingreso, $1400 - 120 = 1280$ soles.

**Ejemplo 6.3** (De medias a coeficientes: por qué el dummy es una *resta*, no una media)**.** Un error común es pensar que "usar dummies es sacar la media de cada grupo". Casi: el coeficiente es la **diferencia** de medias respecto a la base, y es el *intercepto* el que guarda la media del grupo base. Con datos reales de salarios (base `Costa`), el modelo **solo con la región** $\widehat{\text{salario}} = \hat\beta_0 + \hat\beta_1 D_{\text{Selva}} + \hat\beta_2 D_{\text{Sierra}}$ devuelve $\hat\beta_0 = 3734$, $\hat\beta_1 = -116$, $\hat\beta_2 = -83$. Esos números no son inventados: coinciden *exactamente* con las medias por grupo,

$$\bar{Y}_{\text{Costa}} = 3734, \qquad \bar{Y}_{\text{Selva}} = 3618, \qquad \bar{Y}_{\text{Sierra}} = 3651,$$

porque el intercepto es $\bar{Y}_{\text{Costa}}$ y cada coeficiente es la resta $-116 = 3618 - 3734$ y $-83 = 3651 - 3734$. Para *recuperar* la media de un grupo se suma: $\bar{Y}_{\text{Selva}} = \hat\beta_0 + \hat\beta_1 = 3734 - 116 = 3618$. (Ojo: el modelo compara Selva *contra Costa*, no "Selva contra todo lo demás".)

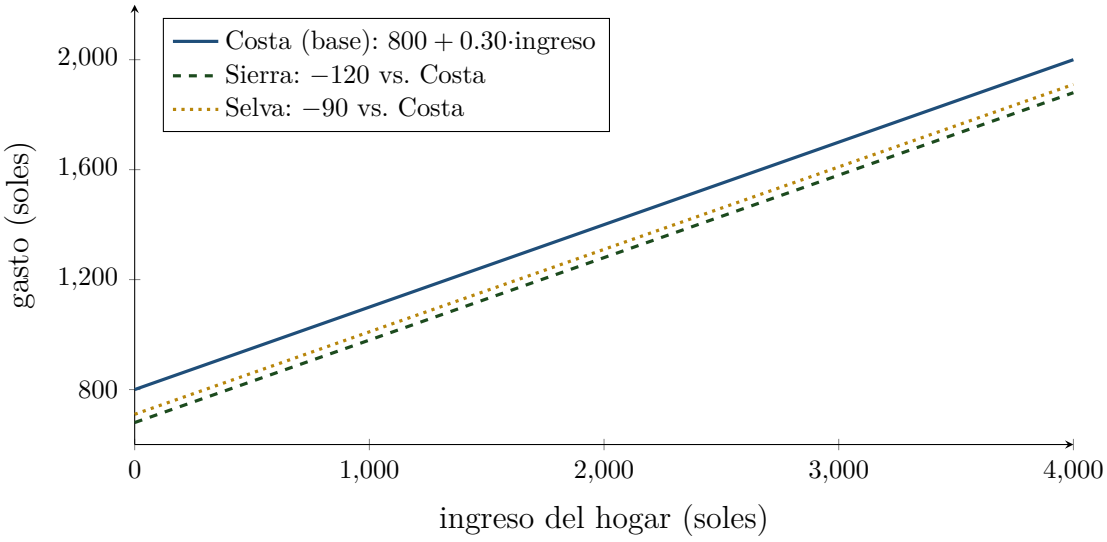


Figura 5: Las dummies regionales producen **rectas paralelas**: la misma pendiente (el efecto del ingreso, 0.30) y *distinto intercepto* según la región. La distancia vertical entre rectas es el coeficiente de cada dummy (−120 para Sierra, −90 para Selva), siempre leído *respecto a la categoría base* (Costa).

Si ahora sumamos `educacion`, el coeficiente de Selva pasa de −116 a −89: ya no es $\bar{Y}_{\text{Selva}} - \bar{Y}_{\text{Costa}}$, sino la brecha *a igual educación* (*ceteris paribus*). Cambia porque la educación promedio difiere por región (Selva 11.9, Costa 12.3, Sierra 12.6 años): parte de la brecha cruda se debía a que en la Selva se estudia menos, y al igualar educación esa penalización regional se achica. **Regla:** dummies *solas* ⇒ el coeficiente es la diferencia de medias; dummies *con controles* ⇒ es esa diferencia **ajustada**.

**Ejemplo 6.4** (La interacción lleva la idea más lejos: el DiD como "resta de restas")**.** El mismo principio —todo coeficiente es una combinación de medias— explica el estimador de **diferencias en diferencias (difference-in-differences, DiD)**. Con un grupo *tratado/control* y un periodo *antes/después*, el modelo con *interacción*

$$Y = \beta_0 + \beta_1\,\text{Tratado} + \beta_2\,\text{Post} + \beta_3\,(\text{Tratado}\cdot\text{Post}) + \varepsilon$$

tiene cuatro parámetros para las cuatro medias de celda. Con datos reales de un programa social:

| | antes | después |
|---|---|---|
| control | 9.73 | 12.11 |
| tratado | 11.05 | 18.32 |

se obtiene $\hat{\beta}_0 = 9.73$ (celda control-antes, la base), $\hat{\beta}_1 = 1.32$ (brecha inicial tratado vs. control), $\hat{\beta}_2 = 2.37$ (la tendencia del tiempo, medida en el control) y la **interacción**

$$\hat{\beta}_3 = (18.32 - 11.05) - (12.11 - 9.73) = 4.89,$$

que es el **DiD**: el cambio del grupo tratado *por encima* de la tendencia común. Si un dummy es "una resta de medias", el DiD es "una resta de restas". (Requiere el supuesto de *tendencias paralelas*: sin el programa, el tratado habría seguido la tendencia del control.)

# 7. Regresión lineal en R

En R, todo lo anterior cabe en la función `lm()` (*linear model*). La sintaxis `lm(y ~ x)` se lee "$y$ explicada por $x$".

**Regresión simple: horas de estudio y nota**

```
# datos del ejemplo a mano
horas <- c(1, 2, 3, 4, 5)
nota  <- c(3, 4, 4, 6, 8)

m1 <- lm(nota ~ horas)        # ajusta nota = b0 + b1*horas
summary(m1)                   # coeficientes, errores estandar, R^2

# Salida (resumida):
#               Estimate Std.Error t value Pr(>|t|)
# (Intercept)     1.400     ...       ...       ...      <- b0 = 1.4
# horas           1.200     ...       ...       ...      <- b1 = 1.2
# Multiple R-squared: 0.9                                <- R^2 = 0.90

coef(m1)        # b0 = 1.4, b1 = 1.2
fitted(m1)      # valores ajustados Y_hat
resid(m1)       # residuos (suman 0)
predict(m1, newdata = data.frame(horas = 6))   # prediccion en X = 6
```

La columna `Estimate` son los $\hat{\beta}$; `Std. Error` es su precisión (error estándar); `t value` y `Pr(>|t|)` son la prueba de $H_0 : \beta_j = 0$ (Clase 9). El `Multiple R-squared` es nuestro $R^2$. Aquí el intercepto es 1.4, la pendiente 1.2 y $R^2 = 0.90$, exactamente lo que calculamos a mano.

Para el modelo **multivariado** con una variable categórica, basta sumar regresores con `+` y envolver la categoríca en `factor()` (R omite una categoría base *automáticamente*, evitando la trampa de las dummies):

**Regresión multivariada con una variable categórica**

```
# gasto explicado por ingreso y region (Costa/Sierra/Selva)
m2 <- lm(gasto ~ ingreso + factor(region), data = hogares)
summary(m2)
```

```
# R toma "Costa" como base (la 1a en orden alfabetico) y reporta:
#   factor(region)Sierra  -> diferencia de Sierra respecto a Costa
#   factor(region)Selva   -> diferencia de Selva  respecto a Costa
# cada coeficiente es "ceteris paribus": manteniendo el resto constante.

# para fijar otra base explicitamente:
hogares$region <- relevel(factor(hogares$region), ref = "Costa")
```

Cuando corres `lm`, R construye la matriz de diseño $X$ (con su columna de unos para el intercepto y las dummies de `factor()`) y calcula $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ internamente. Si por error incluyes dos variables perfectamente colineales, R *te avisa* (deja un coeficiente como `NA`): es la trampa de las dummies en acción.

# Resumen

## Resumen de la Clase 9

- **La idea:** el modelo lineal describe la *esperanza condicional* $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$ como una recta. Es el "caballo de batalla" de la estadística aplicada.
- **Modelo simple:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. El intercepto $\beta_0$ es $Y$ cuando $X = 0$; la pendiente $\beta_1$ es el cambio en $Y$ por unidad de $X$.
- **OLS:** minimiza $\sum (Y_i - \beta_0 - \beta_1 X_i)^2$. Fórmulas cerradas: $\hat{\beta}_1 = \widehat{\mathrm{Cov}}(X, Y) / \widehat{\mathrm{Var}}(X)$ y $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.
- **Ajuste:** $\hat{Y}_i$ (ajustados), $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ (residuos, suman 0); descomposición SST = SSE + SSR y $R^2 = 1 - \mathrm{SSR}/\mathrm{SST}$ = proporción de varianza explicada ($R^2$ alto $\neq$ causalidad).
- **Inferencia base:** $s^2 = \mathrm{SSR}/(n-2)$ estima $\sigma^2$; bajo los supuestos CLRM (i)–(v), Gauss–Markov garantiza que OLS es **BLUE**.
- **Multivariado:** $Y = X\beta + \varepsilon$ con $\hat{\beta} = (X^\top X)^{-1} X^\top Y$; cada $\beta_j$ se interpreta *ceteris paribus*. Cuidado con la *multicolinealidad* (necesitamos $X^\top X$ invertible) y la *trampa de las dummies* (omitir una categoría base).
- **En R:** `lm(y ~ x1 + x2 + factor(region))` y `summary()` lo hacen todo.

**Mensaje central:** el modelo lineal convierte la pregunta "¿cómo se relaciona $X$ con $Y$?" en dos números interpretables ($\hat{\beta}_0, \hat{\beta}_1$) con una medida de ajuste ($R^2$) —y la regresión múltiple nos deja aislar el papel de cada variable manteniendo el resto constante. En la Clase 9 le pondremos *inferencia* (errores estándar, IC y pruebas sobre los coeficientes).

## Fuentes y atribución



**Bibliografía.**


- Duflo, E. & Ellison, S. F. *14.310x: Data Analysis for Social Scientists*. MIT OpenCourseWare (slides, lecture notes y problem sets). ocw.mit.edu.
- Larsen, R. J. & Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 6.ª ed., Pearson, 2018.
- DeGroot, M. H. & Schervish, M. J. *Probability and Statistics*, 4.ª ed., Pearson, 2012.
- Blitzstein, J. & Hwang, J. *Introduction to Probability* (Harvard Stat 110); W. Chen, *Probability Cheatsheet*.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics*. openintro.org.
- Materiales abiertos de la comunidad del curso en GitHub: `gevorg-hunanyan/probability`; `hanmochen/Probability-Theory-2020`; `mynameisjanus/14310xDataAnalysis`.