Modelo Lineal Simple y Multivariado

Clase 9 — 14.310x: Data Analysis for Social Scientists

B.Sc. Cristian Lazo Quispe

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

Programa SDS (Statistics and Data Science, MITx) · **BREIT** (Brescia Institute of Technology), iniciativa de Aporta —Grupo Breca— en alianza con MIT IDSS.

clazoq@uni.pe · linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

¿Qué veremos hoy?

1. El modelo lineal: idea e intuición
2. Mínimos cuadrados (OLS)
3. Residuos, R^2 y σ^2
4. Supuestos (CLRM) y Gauss–Markov
5. Modelo multivariado (matricial)
6. Variables dummy y multicolinealidad
7. Estimar OLS en R
8. Manos a la obra

Contenido

De la media de Y a la media de Y *dado* X

La idea central (de univariado a condicional)

Antes estimábamos un solo número (la media de Y). Ahora preguntamos cómo cambia esa media *según* otra variable X: queremos la **media condicional** $\mathbb{E}[Y \mid X = x]$. Si la pensamos como una *recta*, tenemos el modelo lineal.

Definición 1 (Modelo lineal simple (bivariado))

Para n observaciones (X_i, Y_i),

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

Y *es la variable* ***dependiente (dependent)***; X, *la* ***explicativa (regressor)***; ε, *el* ***error*** *no observado;* β_0, β_1 *son los* ***coeficientes*** *a estimar.*

¿Por qué nos importa? (predicción, causalidad, entender el mundo)

Tomando esperanzas y usando $\mathbb{E}[\varepsilon_i] = 0$: $\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$. La recta *es* la mejor predicción lineal de Y a partir de X.

Demo: leer una recta de regresión

Problema

Una recta ajustada relaciona *años de educación* (X) con el *salario* mensual en soles (Y):

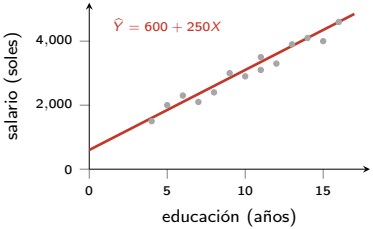
$$\widehat{Y} = 600 + 250\,X.$$

Se pide: (a) salario predicho con 11 años de educación; (b) interpretar $\hat{\beta}_1 = 250$ y $\hat{\beta}_0 = 600$.

Solución

(a) $\widehat{Y} = 600 + 250(11) = 600 + 2750 = \boxed{3350}$ soles.

(b) $\hat{\beta}_1 = 250$: cada **año adicional** de educación se asocia con $+250$ soles, en promedio.
$\hat{\beta}_0 = 600$: salario predicho con $X = 0$ años (extrapolación; cuidado fuera del rango de los datos).



La recta resume la nube: por cada año más de educación, $\sim$ 250 soles más de salario.

Contenido

¿Qué recta elegimos? La de mínimos cuadrados

El criterio de mínimos cuadrados (least squares)

Entre todas las rectas posibles, OLS elige la que hace *más pequeña* la suma de los errores al cuadrado (distancias verticales a la recta):

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_i\big)^2.$$

Solución en forma cerrada (no hace falta minimizar a mano cada vez)

Derivando e igualando a cero se obtienen fórmulas *elegantes*:

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

Lectura intuitiva

La pendiente es **covarianza sobre varianza**: cuánto co-varían X e Y, normalizado por cuánto varía X. La recta *siempre* pasa por el punto medio $(\bar{X}, \bar{Y})$.

Demo: calcular $\hat{\beta}_1$ y $\hat{\beta}_0$ a mano

Problema

Notas de un examen (Y, sobre 20) según *horas de estudio* (X) de 5 alumnos. **Se pide:** la recta OLS $\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

X (horas)	1	2	3	4	5
Y (nota)	8	11	12	15	14

Solución

Medias: $\bar{X} = 3$, $\bar{Y} = \frac{60}{5} = 12$.

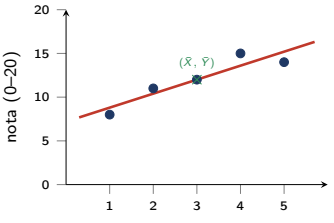
$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (-2)(-4) + (-1)(-1) + 0 + 1 \cdot 3 + 2 \cdot 2 = 8 + 1 + 0 + 3 + 4 = 16$.

$\sum (X_i - \bar{X})^2 = 4 + 1 + 0 + 1 + 4 = 10$.

$$\hat{\beta}_1 = \frac{16}{10} = 1.6, \qquad \hat{\beta}_0 = 12 - 1.6(3) = 12 - 4.8 = 7.2.$$

$$\boxed{\widehat{Y} = 7.2 + 1.6\,X.}$$

Cada hora de estudio $\approx +1.6$ puntos.



La recta OLS pasa por $(\bar{X}, \bar{Y}) = (3, 12)$ y minimiza la suma de cuadrados.

Contenido

¿Qué tan buena es la recta? Descomponer la varianza

Residuos y valores ajustados

El **valor ajustado** es $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$; el **residuo (residual)** es $\hat{\varepsilon}_i = Y_i - \widehat{Y}_i$ (lo que la recta no explica). Analizamos la varianza de Y partiéndola en dos:

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{\text{SST (total)}} = \underbrace{\sum_i (\widehat{Y}_i - \bar{Y})^2}_{\text{SSE (explicada)}} + \underbrace{\sum_i (Y_i - \widehat{Y}_i)^2}_{\text{SSR (residual)}}.$$

Coeficiente de determinación R^2 y varianza del error

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSE}}{\text{SST}} \in [0,1], \qquad s^2 = \frac{\text{SSR}}{n-2} \ \ (\text{estima } \sigma^2).$$

R^2 es la fracción de la variación de Y *explicada* por X. Dividimos entre $n-2$ porque estimamos 2 parámetros (β_0, β_1).

Aviso: R^2 alto $\neq$ causalidad

Un R^2 grande dice que la recta *predice* bien, no que X *cause* Y. Puede haber un confusor detrás (¡recuerda la Clase 8!).

Demo: calcular R^2 y s^2

Problema

Con la recta del demo anterior, $\widehat{Y} = 7.2 + 1.6X$, sobre los mismos 5 alumnos. **Se pide:** SSR, SST, R^2 y s^2.

X	Y	$\widehat{Y}$	$\hat{\varepsilon} = Y - \widehat{Y}$	$\hat{\varepsilon}^2$
1	8	8.8	−0.8	0.64
2	11	10.4	0.6	0.36
3	12	12.0	0.0	0.00
4	15	13.6	1.4	1.96
5	14	15.2	−1.2	1.44

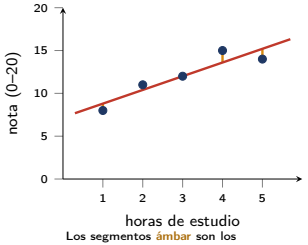
Solución

$\text{SSR} = 0.64 + 0.36 + 0 + 1.96 + 1.44 = 4.4.$

$\text{SST} = \sum (Y_i - 12)^2 = 16 + 1 + 0 + 9 + 4 = 30.$

$$R^2 = 1 - \frac{4.4}{30} = 1 - 0.147 = \boxed{0.853}, \qquad s^2 = \frac{4.4}{5-2} = \frac{4.4}{3} \approx 1.47.$$

La recta explica $\sim 85\,\%$ de la variación de las notas.



Los segmentos ámbar son los residuos $\hat{\varepsilon}_i$. OLS minimiza su suma de cuadrados.

Contenido

¿Cuándo confiar en OLS? Los supuestos del modelo clásico

Modelo clásico de regresión lineal (CLRM), en palabras

1. **Linealidad + media cero del error:** $\mathbb{E}[\varepsilon_i] = 0$; la recta es la media condicional correcta.
2. **Exogeneidad:** X y ε no están correlacionados (X no trae información "oculta" del error).
3. **Identificación:** X *varía* ($\mathrm{Var}(X) > 0$); sin variación no hay pendiente que estimar.
4. **Homocedasticidad (homoskedasticity):** la varianza del error es *constante*, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
5. **Sin autocorrelación:** errores de distintas observaciones no se relacionan, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

Teorema de Gauss–Markov: OLS es BLUE

Bajo estos supuestos, OLS es el **mejor estimador lineal insesgado** (*Best Linear Unbiased Estimator*): de todos los estimadores lineales e insesgados, OLS tiene la *menor varianza*. Si además $\varepsilon \sim N(0, \sigma^2)$, los $\hat{\beta}$ son normales y coinciden con el MLE.

Demo visual: homocedasticidad vs. heterocedasticidad

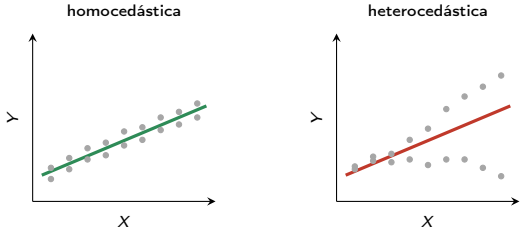
Problema (lectura de gráficos)

¿Cuál de las dos nubes *cumple* el supuesto de homocedasticidad? **Se pide:** identificarla y decir por qué.

Solución

La de la **izquierda**: la dispersión de los puntos alrededor de la recta es *pareja* en todo el rango de $X \Rightarrow$ varianza constante.

La de la derecha *abre como un embudo*: la dispersión crece con $X \Rightarrow$ **heterocedasticidad** (se viola el supuesto iv).



Izquierda: dispersión pareja (OK).

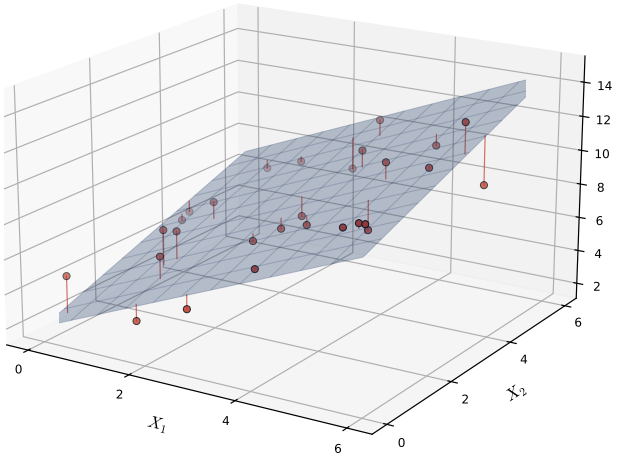
Derecha: "embudo" que crece con X (viola iv).

Contenido

De la recta al plano (y al hiperplano)

- Con **1** predictor: el modelo es una **recta**.
- Con **2** predictores X_1, X_2: es un **plano**.
- Con k predictores: un **hiperplano** en $k+1$ dimensiones.

OLS elige el plano que minimiza la suma de los residuos al cuadrado (segmentos verticales).



Varias variables a la vez: notación matricial

El modelo lineal general

Con k explicativas, $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$. Apilando las n observaciones, todo se vuelve *compacto*:

$$\underbrace{Y}_{n\times 1} = \underbrace{X}_{n\times(k+1)} \underbrace{\beta}_{(k+1)\times 1} + \underbrace{\varepsilon}_{n\times 1},$$

donde la primera columna de X es de unos (para el intercepto β_0).

Estimador OLS en forma matricial

Minimizando $(Y - X\beta)^\top (Y - X\beta)$ se obtiene la fórmula que generaliza "covarianza sobre varianza":

$$\boxed{\hat{\beta} = (X^\top X)^{-1} X^\top Y}$$ (requiere que $X^\top X$ sea invertible).

Interpretación "ceteris paribus" (manteniendo lo demás fijo)

En el modelo multivariado, $\hat{\beta}_j$ es el efecto de subir X_j en una unidad **manteniendo constantes** las demás variables. Eso es lo que permite "controlar" por confusores.

Demo: "controlar" por una variable cambia la pendiente

Problema

Regresamos *salario* sobre *educación*. Pero la educación se relaciona con la *experiencia*. Dos modelos sobre los mismos datos:

$$(1) \quad \widehat{Y} = 600 + 250\,\text{Educ},$$

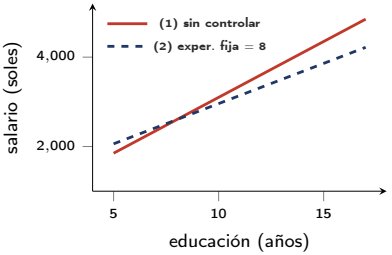
$$(2) \quad \widehat{Y} = 400 + 180\,\text{Educ} + 95\,\text{Exper}.$$

Se pide: interpretar 180 y por qué difiere de 250.

Solución

En (2), $\hat{\beta}_{\text{Educ}} = 180$: el efecto de un año más de educación **manteniendo fija la experiencia** (*ceteris paribus*).

En (1) faltaba la experiencia: parte de su efecto se "colaba" en la educación (sesgo por variable omitida), inflando la pendiente a 250. Al controlar por experiencia, el efecto *puro* de la educación baja a 180.



La recta sin controlar (roja) tiene mayor pendiente; al fijar la experiencia, baja.

Contenido

Variables categóricas: dummies y la categoría base

Variable dummy (indicadora)

Una **dummy (dummy/indicator)** toma valores 0 o 1 para marcar una característica (mujer/hombre, urbano/rural, tratado/control). En $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$, $\hat{\beta}_1$ es la *diferencia de medias (difference in means)* entre los dos grupos.

Más de dos categorías ⇒ una dummy menos (categoría base)

Con C categorías (p. ej. región: costa, sierra, selva) incluimos $C - 1$ dummies; la omitida es la **categoría base (baseline)** contra la que se comparan las demás. Cada coeficiente es la diferencia respecto de esa base.

La "trampa de las dummies" (dummy variable trap)

Si incluyes *todas* las dummies **más** el intercepto, ellas suman una columna de unos: **colinealidad perfecta**. $X^\top X$ no es invertible y OLS falla. Solución: omitir una categoría (o el intercepto).

Demo: leer coeficientes de dummies con categoría base

Problema

Gasto mensual del hogar (Y, soles) según región, con **costa** como base:

$$\widehat{Y} = 1800 - 300\, D_{\text{sierra}} - 450\, D_{\text{selva}}.$$

Se pide: gasto predicho en cada región e interpretación de -300.

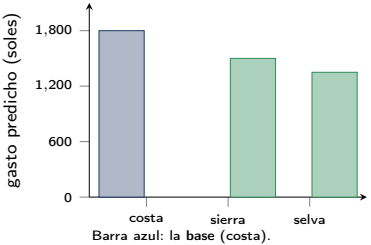
Solución

Costa ($D_{\text{sierra}} = D_{\text{selva}} = 0$): $\widehat{Y} = 1800$.

Sierra ($D_{\text{sierra}} = 1$): $\widehat{Y} = 1800 - 300 = 1500$.

Selva ($D_{\text{selva}} = 1$): $\widehat{Y} = 1800 - 450 = 1350$.

-300 *no* es "el gasto de la sierra": es cuánto **menos** gasta la sierra *comparada con la costa* (la base). Incluir también una dummy de costa rompería el modelo (trampa de las dummies).



Barra azul: la base (costa).

Las otras se leen como diferencias respecto a ella.

Momento "ajá": el dummy es una *resta* de medias, no la media

Solo el dummy: `lm(salario ~ factor(region))`

$$\widehat{\text{salario}} = 3734 - 116\, D_{\text{Selva}} - 83\, D_{\text{Sierra}}$$

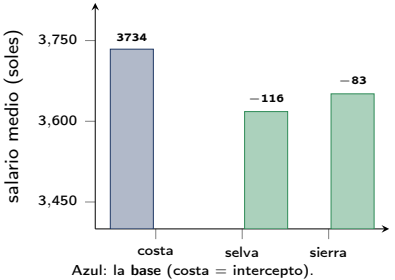
coinciden *exactamente* con las medias por grupo: $\bar{Y}_{\text{Costa}} = 3734$, $\bar{Y}_{\text{Selva}} = 3618$, $\bar{Y}_{\text{Sierra}} = 3651$.

¿Por qué?

Intercepto $= \bar{Y}_{\text{base}}$; cada coef $=$ **diferencia**: $-116 = 3618 - 3734$. Se recupera la media sumando: $3734 - 116 = 3618$.

Al agregar `educacion` ya no cuadra

El coef de Selva pasa de -116 a -89: ahora es la brecha *a igual educación* (*ceteris paribus*). Cambia porque la educación media difiere (Selva 11.9, Costa 12.3, Sierra 12.6 años).



Azul: la base (costa = intercepto).

Verdes: su altura es la media; la *caída* respecto a la azul es el coeficiente.

...y con interacción: el DiD es una *resta de restas*

Tabla 2×2 de medias (programa social)

	antes	después
control	9.73	12.11
tratado	11.05	18.32

`lm(y ~ tratado*post)` decodifica las 4 celdas

- $\hat{\beta}_0 = 9.73$: celda control-antes (base)
- $\hat{\beta}_{\text{tratado}} = 1.32$: brecha inicial
- $\hat{\beta}_{\text{post}} = 2.37$: tendencia del control
- $\hat{\beta}_{\text{inter}} = 4.89$: el **DiD**

El DiD como resta de restas

$$\underbrace{(18.32 - 11.05)}_{\text{cambio tratado}} - \underbrace{(12.11 - 9.73)}_{\text{cambio control}} = 4.89$$

La interacción aísla cuánto se movió el tratado *por encima* de la tendencia común. Igual que el dummy, es una **combinación de medias** —nunca "la media" a secas— y necesita el supuesto de *tendencias paralelas*.

Contenido

De la fórmula al código: `lm` en R

Idea

Nunca calculamos $(X^\top X)^{-1}X^\top Y$ a mano: la función `lm` ajusta el modelo, da coeficientes, errores estándar, R^2 y pruebas. Las dummies salen solas con `factor`.

En R

```
# salario ~ educacion + experiencia + region (dummy)
modelo <- lm(salario ~ educacion + experiencia + factor(region),
             data = encuesta)
summary(modelo)          # coeficientes, SE, t, p y R^2
coef(modelo)             # solo los beta estimados
predict(modelo)          # valores ajustados Y_hat
resid(modelo)            # residuos  Y - Y_hat
```

Lectura

`factor(region)` crea las dummies y *omite una categoría* como base automáticamente (evita la trampa). Cada coeficiente de región se lee como diferencia respecto de esa base, *ceteris paribus*.

Práctica 1 en R: Fisher exacto enumerando combinaciones (MIT)

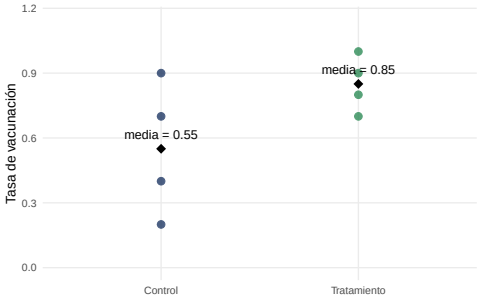
Problema

Programa en 4 de 8 regiones (`programa.csv`: tasa de vacunación). El **ATE** estimado (*difference in means*) $= 0.85 - 0.55 = 0.30$ es el `observed_test`. Enumera las $\binom{8}{4} = 70$ asignaciones.

Completa los `____` (notación MIT)

```
library(perm)
perms <- chooseMatrix(8, ____)     # 70x8 de 0/1
treatment_avg <- perms %*% ____ / n_treat
control_avg   <- (1 - perms) %*% A / n_treat
test_statistic <- abs(treatment_avg - control_avg)
pvalue <- mean(test_statistic >= ____)
```

Solución: `practica_R/p1_permutacion_sol.R`



`observed_test`$= 0.30$ (ATE), **no** el p-valor. Solo 14/70 lo **superan**: `pvalue`$= 0.20$, **no** se rechaza la *sharp null*.

Práctica 2 en R: diferencias en diferencias (diff-in-diff)

Problema

Programa social, panel antes/después (`did_panel.csv`: `tratado`, `post`, `y`). Estilo Qian (MIT): interacción `tratado*post`.

Completa los ____

```
datos$tratadopost <- datos$tratado * ____
m <- lm(y ~ tratado + post +
        ____, data = datos)
coef(m)["tratadopost"]  # DiD
```

Solución: `practica_R/p2_diff_in_diff_sol.R`

La línea punteada roja es el contrafactual; la brecha final es el DiD ≈ 4.9.

Práctica 3 en R: OLS "a mano" (lo que hace `lm` por dentro)

Problema

Calcula los coeficientes OLS con álgebra de matrices, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, y comprueba que `lm` da lo mismo. `lm` no es magia.

Completa los ____

```
Y <- datos$salario
X <- cbind(1, datos$educacion,
           ____)        # + experiencia
beta <- ____(t(X) %*% X) %*% (t(X) %*% Y)
beta                    # = coef(lm(...))
```

Solución: `practica_R/p3_ols_manual_sol.R`

Sale **lo mismo** que `lm`:

$$\hat{\beta} = \begin{pmatrix} 45.60 \\ 197.18 \\ 67.80 \end{pmatrix} \begin{matrix} \textbf{(intercepto)} \\ \textbf{(educación)} \\ \textbf{(experiencia)} \end{matrix}$$

La columna de 1s = el **intercepto**. `lm` resuelve estas mismas *normal equations* (con QR).

Práctica 4 en R: de una a dos variables (¿qué mide cada coef?)

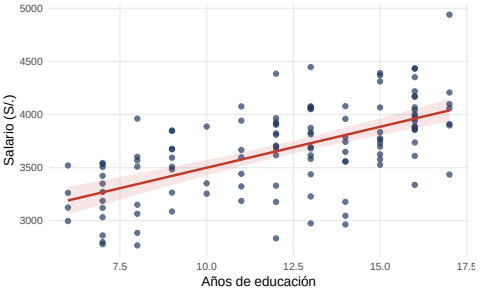
Problema

`salarios.csv`. Ajusta M1 (solo educación) y M2 ($+$ experiencia). Interpreta el coeficiente de educación en cada uno y por qué cambia.

Completa los ____

```
m1 <- lm(salario ~ ____, data=datos)
m2 <- lm(salario ~ educacion +
           ____, data=datos)
summary(m1); summary(m2)
```

Solución: `practica_R/p4_multivariable_sol.R`



$\hat{\beta}_{educ}$: 77 (M1, crudo) $\rightarrow$ 197 (M2, *ceteris paribus*). Sube por OVB ($r_{educ,exp} = -0.83$).

Práctica 5 en R: dummy, modelo mixto y F-test

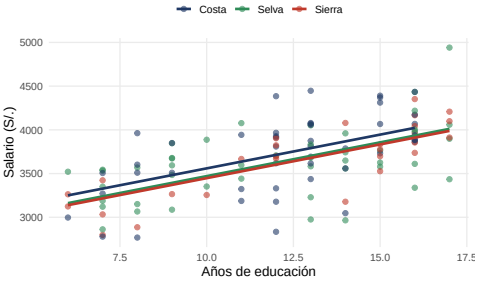
Problema

`region` como dummy (Costa = base). M3 (+ region), M4 (mixto). ¿Qué mide cada coef? ¿Región es *conjuntamente* significativa?

Completa los ____

```
m3 <- lm(salario ~ educacion +
         ____(region), data=datos)
m4 <- lm(salario ~ educacion +
   experiencia + factor(region), data=datos)
anova(m_sin, ____)   # F-test region
```

Solución: `practica_R/p5_dummy_sol.R`



Sierra = diferencia vs Costa (≈ -120). F-test región: $p = 0.032 \Rightarrow$ conjuntamente significativa.

Cómo leer el output de R: ¿qué mide cada cosa?

Qué ves	Qué es	Cómo se lee
`Estimate` ($\hat{\beta}$)	el efecto estimado	tamaño y signo (ej. $+197$: sube 197 por unidad de X)
`Std. Error`	incertidumbre de $\hat{\beta}$	más chico = estimación más precisa
`t value`	$\hat{\beta}/\mathrm{SE}$	a cuántos errores estándar de 0 está el coef
`Pr(>|t|)`	el **p-value** del coef	significancia: $< 0.05 \Rightarrow \hat{\beta} \neq 0$
R^2	varianza de Y explicada	de 0 a 1; ajuste global (*no* es una prueba)
`anova` (F)	prueba *conjunta*	¿un *grupo* de variables aporta al modelo?

Clave

El $\hat{\beta}$ (Estimate) dice el **tamaño** del efecto; el **p-value** dice solo si es **distinguible de** 0. Son cosas distintas: un efecto puede ser grande pero no significativo (poca data), o chico pero significativo (mucha data).

¿Cuándo se rechaza? H_0 vs H_1 de cada prueba

Prueba	H_0 (nula)	H_1 (si rechazas, $p < 0.05$)
t-test de un coef	$\beta = 0$: la variable no tiene efecto	$\beta \neq 0$: la variable **sí** importa
F-test (grupo)	*todos* esos $\beta = 0$: el grupo no aporta	**al menos uno** $\neq 0$: el grupo **sí** aporta
Fisher / permut.	*sharp null*: no afecta a *nadie*	el tratamiento **sí** tiene efecto

Regla de decisión

Rechazas H_0 si el **p-value** < 0.05 (lo observado sería muy raro si H_0 fuera cierta).

El matiz del F-test

Rechazar el F-test **no** dice que *todas* las variables del grupo importen, sino que **al menos una** lo hace.

Ejemplo en R: t-test de cada variable vs F-test del grupo

Output del modelo mixto (P5):

```
Coefficients:         Estimate  Pr(>|t|)
educacion              197.5    <2e-16 ***
factor(region)Selva    -85.3    0.047 *
factor(region)Sierra  -119.9    0.016 *

anova(m_sin, m4):      F=3.55   p=0.032 *
```

Cada fila = **un** t-**test** ($H_0 : \beta = 0$): educación $p < 0.001$, Selva 0.047, Sierra 0.016 $\Rightarrow$ las tres rechazan.

El anova = **un** F-**test** ($H_0 : \beta_{\textbf{Selva}} = \beta_{\textbf{Sierra}} = 0$): $p = 0.032 \Rightarrow$ la *región* importa como bloque.

Caso B: si Selva y Sierra dieran $p = 0.30$ pero el F-test $p = 0.02$, igual concluirías que la región importa (H_1: al menos una $\neq 0$; evidencia repartida).

Contenido

¿Y ahora?

- El modelo lineal estima la **media condicional** $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$: la recta es la mejor predicción lineal de Y.
- **OLS** elige la recta que minimiza $\sum (Y_i - \beta_0 - \beta_1 X_i)^2$; recuerda $\hat{\beta}_1 = \text{Cov} / \text{Var}$ y $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.
- Mide el ajuste con $R^2 = 1 - \text{SSR}/\text{SST}$ y estima σ^2 con $s^2 = \text{SSR}/(n-2)$; pero R^2 **alto** $\neq$ **causalidad**.
- Bajo los supuestos del **CLRM**, Gauss–Markov garantiza que OLS es **BLUE**. Vigila la homocedasticidad.
- En el modelo **multivariado** ($\hat{\beta} = (X^\top X)^{-1} X^\top Y$) cada coeficiente es un efecto *ceteris paribus*; úsalo para controlar confusores.
- Con **dummies** elige una categoría base y evita la *trampa* (colinealidad perfecta). En R: `lm`, `factor`, `summary`, `predict`.

¡A practicar!