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

Ejercicios (5 niveles)

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

Niveles de dificultad *(difficulty levels)*

Cada problema está rotulado por color según su nivel, de lo más accesible (Nivel 1) a lo más difícil (Nivel 5). Empieza por donde te sientas cómodo y avanza a tu ritmo; los niveles 4–5 son extensiones para profundizar, no un requisito.

- **Nivel 1** – **Fundamentos**: calentamiento muy guiado; para quien parte de casi cero.
- **Nivel 2** – **Núcleo**: el nivel objetivo del curso/examen.
- **Nivel 3** – **Aplicación**: combina varias ideas en contexto realista.
- **Nivel 4** – **Reto**: difícil, integra varios temas.
- **Nivel 5** – **Reto+**: muy difícil / fuera del scope típico.

Indicaciones

Resuelve los ejercicios de tu nivel; si terminas, sube al siguiente (ver la leyenda de arriba). Recuerda el *modelo lineal simple* $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$: los estimadores de mínimos cuadrados (OLS) son

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

El residuo es $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ con $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$; la bondad de ajuste es $R^2 = 1 - \mathrm{SSR}/\mathrm{SST}$, donde $\mathrm{SST} = \sum_i (Y_i - \bar{Y})^2$ y $\mathrm{SSR} = \sum_i \hat{\varepsilon}_i^2$. En forma matricial, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ y $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. Bajo los supuestos del modelo clásico (CLRM), OLS es el estimador lineal insesgado de mínima varianza (Gauss–Markov). Interpreta siempre cada coeficiente *manteniendo constantes* los demás regresores.

1. Problemas

Nivel 1 – Fundamentos (warm-up)

Para quien parte de casi cero: cálculos directos y guiados.

Ejercicio 1.1 (Leer una recta de regresión dada)**.** Un analista ajusta la relación entre las horas de estudio semanal (X) y la nota final (sobre 20) y obtiene la *recta de regresión* (regression line)

$$\hat{Y} = 8 + 1.5\, X.$$

(a) ¿Cuál es el *intercepto* (intercept) $\hat{\beta}_0$ y cuál la *pendiente* (slope) $\hat{\beta}_1$?

(b) Predice la nota de alguien que estudia $X = 4$ horas.

(c) Predice la nota de alguien que estudia $X = 6$ horas.

Ejercicio 1.2 (Interpretar la pendiente en palabras)**.** Para una muestra de distritos, una regresión del gasto mensual en libros Y (en soles) sobre el ingreso familiar X (en cientos de soles) da

$$\hat{Y} = 20 + 3\,X.$$

(a) Interpreta el valor $\hat{\beta}_1 = 3$ *en palabras* (incluye las unidades).

(b) Interpreta el intercepto $\hat{\beta}_0 = 20$: ¿qué representa cuando $X = 0$?

Ejercicio 1.3 (Valor ajustado y residuo)**.** Con la recta $\hat{Y} = 8 + 1.5\,X$ (nota vs. horas de estudio), una alumna estudió $X = 6$ horas y sacó una nota *real* de $Y = 15$.

(a) Calcula el *valor ajustado* (fitted value) $\hat{Y}$ para esta alumna.

(b) Calcula el *residuo* (residual) $\hat{\varepsilon} = Y - \hat{Y}$.

(c) ¿La recta *sobrestima* o *subestima* su nota?

Ejercicio 1.4 (Promedios, $\widehat{\text{Cov}}$ y $\widehat{\text{Var}}$ muestrales)**.** Para 5 familias se registra el ingreso X (en cientos de soles) y el ahorro Y (en cientos de soles):

$$(X, Y): \ (2,5),\ (4,8),\ (6,9),\ (8,13),\ (10,15).$$

(a) Calcula los promedios $\bar{X}$ y $\bar{Y}$.

(b) Calcula $\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$ y $\sum_i (X_i - \bar{X})^2$.

Ejercicio 1.5 (Calcular $\hat{\beta}_1$ y $\hat{\beta}_0$ a mano (guiado))**.** Usa los mismos datos del problema anterior: $(X, Y): \ (2,5),\ (4,8),\ (6,9),\ (8,13),\ (10,15)$, con $\bar{X} = 6$, $\bar{Y} = 10$, $\sum(X - \bar{X})(Y - \bar{Y}) = 50$ y $\sum(X - \bar{X})^2 = 40$.

(a) Calcula la pendiente $\hat{\beta}_1 = \dfrac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$.

(b) Calcula el intercepto $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

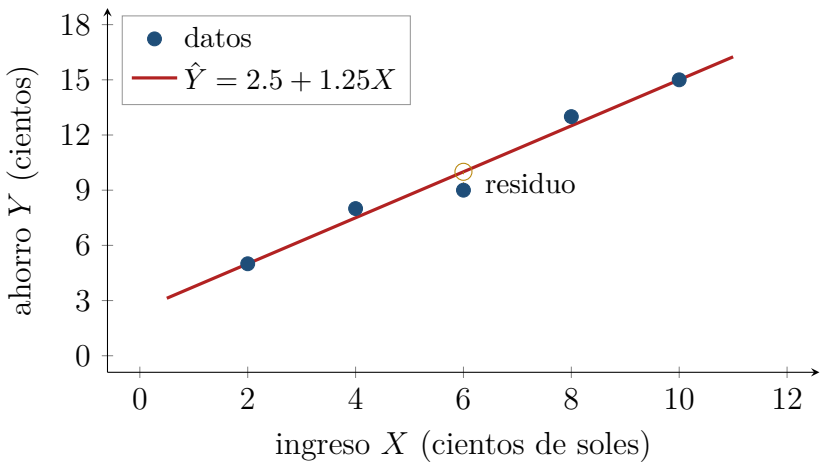
(c) Escribe la recta de regresión $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Ejercicio 1.6 (Calcular un valor ajustado y un residuo dentro del dataset)**.** Con la recta recién obtenida $\hat{Y} = 2.5 + 1.25\,X$ y la familia con $X = 6$, $Y = 9$:

(a) Calcula el valor ajustado $\hat{Y}$ para $X = 6$.

(b) Calcula el residuo $\hat{\varepsilon} = Y - \hat{Y}$.

Ejercicio 1.7 (Nube de puntos con recta ajustada: lectura visual)**.** La figura muestra 5 puntos (X, Y) y su recta de regresión $\hat{Y} = 2.5 + 1.25\,X$ (la del problema anterior).



(a) ¿La pendiente de la recta es positiva o negativa? ¿Qué significa eso para la relación ingreso–ahorro?

(b) El punto en $X = 6$ está *debajo* de la recta. ¿Su residuo es positivo o negativo?

Nivel 2 – Núcleo (core)

El nivel objetivo del curso/examen; todos deberían llegar aquí.

Ejercicio 1.8 (OLS completo desde cero (pendiente, intercepto, ajustados, residuos))**.** Para 5 trabajadores se mide la educación X (años por encima de primaria) y el salario por hora Y (en soles). Datos:

$$(X, Y): \ (1,3),\ (2,9),\ (3,10),\ (4,11),\ (5,12).$$

(a) Calcula $\bar{X}$, $\bar{Y}$, $\sum(X-\bar{X})(Y-\bar{Y})$ y $\sum(X-\bar{X})^2$.

(b) Obtén $\hat{\beta}_1$ y $\hat{\beta}_0$.

(c) Calcula los valores ajustados $\hat{Y}_i$ y los residuos $\hat{\varepsilon}_i$, y verifica que $\sum_i \hat{\varepsilon}_i = 0$.

Ejercicio 1.9 (R^2 vía SST, SSR y SSM)**.** Sigue con la regresión anterior ($\hat{Y} = 3 + 2X$, residuos $(-2, 2, 1, 0, -1)$, $\bar{Y} = 9$, $Y = (3, 9, 10, 11, 12)$).

(a) Calcula $\text{SST} = \sum_i (Y_i - \bar{Y})^2$ y $\text{SSR} = \sum_i \hat{\varepsilon}_i^2$.

(b) Calcula $R^2 = 1 - \text{SSR}/\text{SST}$ e interprétalo.

Ejercicio 1.10 (Estimar σ^2 con $s^2 = \text{SSR}/(n-2)$)**.** En la misma regresión, $\text{SSR} = 10$ y $n = 5$.

(a) ¿Por qué el denominador es $n - 2$ y no n?

(b) Calcula la estimación insesgada de la varianza del error $s^2 = \dfrac{\text{SSR}}{n-2}$ y la desviación estándar de la regresión $s = \sqrt{s^2}$.

Ejercicio 1.11 (La relación $\hat{\beta}_1 = \widehat{\text{Cov}}(X,Y)/\widehat{\text{Var}}(X)$)**.** La pendiente de OLS se puede escribir como $\hat{\beta}_1 = \dfrac{\widehat{\text{Cov}}(X,Y)}{\widehat{\text{Var}}(X)}$, donde $\widehat{\text{Cov}}(X,Y) = \dfrac{1}{n}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$ y $\widehat{\text{Var}}(X) = \dfrac{1}{n}\sum_i (X_i - \bar{X})^2$ (mismo n en ambas).

(a) Con el dataset $(X,Y) : (1,3), (2,9), (3,10), (4,11), (5,12)$ ya sabes que $\sum(X - \bar{X})(Y - \bar{Y}) = 20$ y $\sum(X - \bar{X})^2 = 10$ ($n = 5$). Calcula $\widehat{\text{Cov}}(X,Y)$ y $\widehat{\text{Var}}(X)$.

(b) Verifica que $\hat{\beta}_1 = \widehat{\text{Cov}}(X,Y)/\widehat{\text{Var}}(X)$ coincide con el $\hat{\beta}_1 = 2$ de antes. ¿Por qué el $1/n$ se cancela?

Ejercicio 1.12 (La recta de regresión pasa por $(\bar{X}, \bar{Y})$)**.** Una propiedad clave de OLS (con intercepto) es que la recta ajustada *siempre* pasa por el punto de promedios $(\bar{X}, \bar{Y})$.

(a) Usando $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, demuestra algebraicamente que al evaluar la recta en $X = \bar{X}$ se obtiene $\hat{Y} = \bar{Y}$.

(b) Verifícalo con la recta $\hat{Y} = 3 + 2X$ y los promedios $\bar{X} = 3$, $\bar{Y} = 9$.

Ejercicio 1.13 (Predicción dentro y fuera del rango (extrapolación))**.** Con la regresión salario–educación $\hat{Y} = 3 + 2X$ (X = años sobre primaria, Y = soles/hora; los datos van de $X = 1$ a $X = 5$):

(a) Predice el salario para $X = 4$ (dentro del rango).

(b) Predice el salario para $X = 20$. ¿Por qué hay que tomar esta predicción con cautela?

Ejercicio 1.14 (Interpretar un coeficiente con signo y unidades)**.** Un estudio sobre 200 viviendas regresa el precio de alquiler Y (en soles/mes) sobre la distancia al centro X (en km) y obtiene

$$\hat{Y} = 1\,800 - 90\,X, \qquad R^2 = 0.42.$$

(a) Interpreta $\hat{\beta}_1 = -90$ en palabras (signo y unidades).

(b) ¿Qué dice $R^2 = 0.42$ sobre el ajuste?

Ejercicio 1.15 (Efecto de cambiar las unidades de X)**.** La regresión $\hat{Y} = 3 + 2X$ tiene X en *años* de educación e Y en soles/hora. Un colega quiere medir la educación en *meses* ($X' = 12X$).

(a) Si $X' = 12X$, ¿cuál es la nueva pendiente $\hat{\beta}_1'$ de la regresión de Y sobre X'? (Pista: el efecto de *un mes* debe ser $1/12$ del efecto de un año.)

(b) ¿Cambia el intercepto $\hat{\beta}_0$? ¿Cambia el R^2?

Ejercicio 1.16 (Variable dummy: la pendiente es una diferencia de medias)**.** Sea D una variable *dummy* (indicadora): $D = 1$ si la persona vive en zona urbana y $D = 0$ si vive en zona rural. Se regresa el ingreso Y (soles) sobre D:

$$\hat{Y} = 900 + 400\, D.$$

(a) ¿Cuál es el ingreso promedio predicho de los **rurales** ($D = 0$) y de los **urbanos** ($D = 1$)?

(b) ¿Qué representa $\hat{\beta}_1 = 400$?

Ejercicio 1.17 (De R^2 a SSR y SSM)**.** Una regresión con $n = 30$ observaciones reporta $\text{SST} = 500$ y $R^2 = 0.64$.

(a) Calcula SSR (suma de cuadrados de residuos) y $\text{SSM} = \text{SST} - \text{SSR}$ (suma de cuadrados del modelo).

(b) Estima σ^2 con $s^2 = \text{SSR}/(n-2)$.

Ejercicio 1.18 (R^2 en regresión bivariada y el coeficiente de correlación)**.** El material de clase (MIT 14.310x, L17) recuerda que, en regresión *bivariada* (un solo regresor), $R^2 = r_{XY}^2$, el cuadrado del coeficiente de correlación muestral entre X e Y.

(a) Si la correlación muestral es $r_{XY} = 0.9$, ¿cuánto vale R^2?

(b) Si en otra regresión $R^2 = 0.49$ y la pendiente es *negativa*, ¿cuál es r_{XY} (con su signo)?

Nivel 3 – Aplicación

Combina dos o más ideas en contexto realista; consolida lo aprendido.

Ejercicio 1.19 (Regresión múltiple: "manteniendo constantes los demás")**.** Con datos de un mercado laboral peruano se estima el salario mensual Y (en soles) según educación ed (años), experiencia exp (años) y un dummy mujer (= 1 si es mujer):

$$\widehat{Y} = 800 + 250\,\text{ed} + 60\,\text{exp} - 300\,\text{mujer}.$$

(1) Interpreta el coeficiente 250 de educación *manteniendo constantes* la experiencia y el sexo.

(2) Interpreta el coeficiente -300 del dummy mujer.

(3) Predice el salario de un **hombre** con $\text{ed} = 5$, $\text{exp} = 10$ y de una **mujer** con los mismos valores. ¿Cuál es la brecha?

Ejercicio 1.20 (Variables dummy con varias categorías y la trampa de las dummies)**.** Se quiere incluir la *región* (Costa, Sierra, Selva) en una regresión de salarios. Un analista crea tres dummies: Costa, Sierra, Selva (cada una 0/1) y las mete *todas* junto con el intercepto.

(1) Explica por qué esto provoca *colinealidad perfecta* (la "trampa de las dummies", *dummy variable trap*).

(2) Propon la solución correcta. Si se omite "Costa" como **categoría base**, ¿cómo se interpreta el coeficiente de Sierra?

Ejercicio 1.21 (Regresión simple vs. múltiple: el coeficiente cambia)**.** Para estudiar el efecto de las *horas de sueño* sobre la nota, se corren dos regresiones sobre los mismos estudiantes:

$$\text{(simple) } \widehat{\text{nota}} = 8 + 1.2\,\text{sueño}; \qquad \text{(múltiple) } \widehat{\text{nota}} = 6 + 0.4\,\text{sueño} + 0.9\,\text{estudio}.$$

(1) ¿Por qué el coeficiente del sueño cae de 1.2 a 0.4 al añadir las horas de estudio?

(2) ¿Cuál de los dos coeficientes (1.2 o 0.4) se acerca más al efecto *propio* del sueño? Explica.

Ejercicio 1.22 (Predecir con varios regresores e interpretar el ajuste)**.** Una municipalidad modela el gasto mensual en transporte Y (soles) de los hogares:

$$\widehat{Y} = 50 + 0.05\,\text{ingreso} + 120\,\text{auto} - 15\,\text{distancia}, \qquad R^2 = 0.55,$$

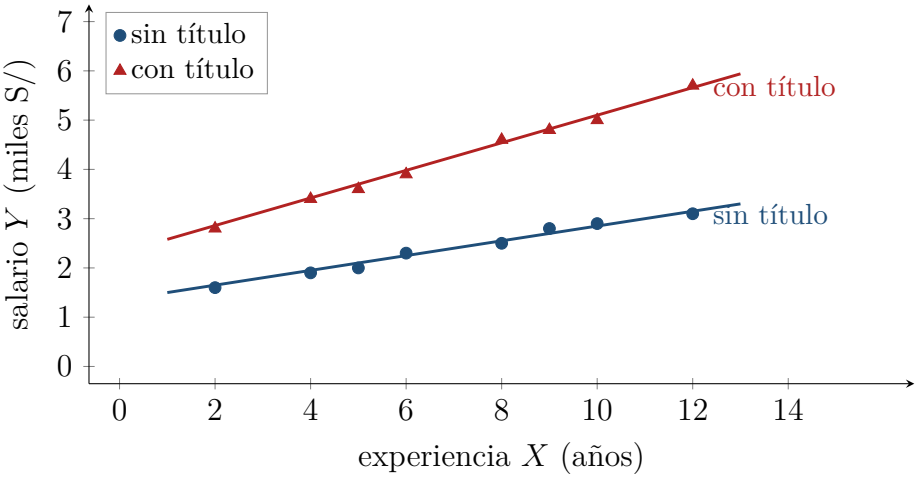
donde ingreso está en soles, auto $= 1$ si el hogar tiene auto y distancia en km al trabajo.

(1) Predice el gasto de un hogar con ingreso 3 000, auto $= 1$ y distancia $= 8$.

(2) Interpreta el coeficiente 120 del dummy auto y el -15 de distancia, ambos "manteniendo constantes los demás".

(3) ¿Qué informa $R^2 = 0.55$?

Ejercicio 1.23 (Comparar dos rectas en una nube de puntos (dummy de grupo))**.** La figura muestra el salario Y (miles de soles) vs. años de experiencia X para dos grupos: con título universitario (rojo) y sin título (azul), junto con la recta ajustada de cada grupo.



(1) ¿Qué grupo tiene mayor intercepto y cuál mayor pendiente? Interpreta ambas diferencias.

(2) Si quisieras capturar ambas diferencias (nivel y pendiente) en *una sola* regresión, ¿qué términos incluirías? (idea)

Ejercicio 1.24 (Construir la recta y el R^2 en contexto (gasto vs. ingreso))**.** Una encuesta a 5 hogares registra el ingreso X (en cientos de soles) y el gasto en alimentos Y (en cientos de soles):

$$(X, Y): \ (1,3),\ (2,9),\ (3,10),\ (4,11),\ (5,12).$$

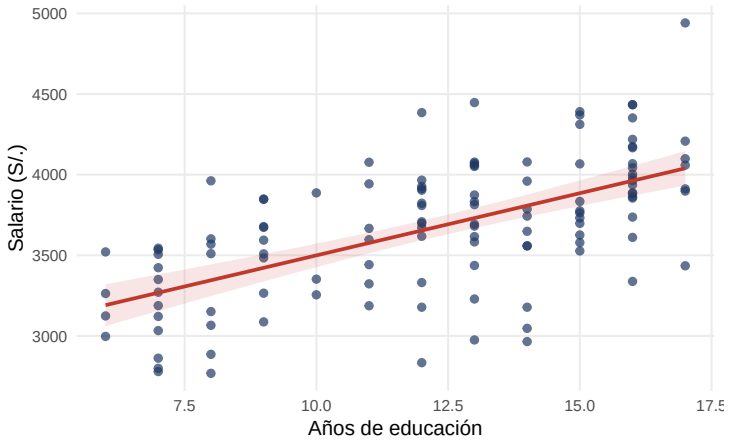
(Es el mismo patrón numérico del Núcleo, ahora en contexto.)

(1) Halla la recta de regresión $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

(2) Calcula R^2 e interprétalo.

(3) Estima la *propensión marginal a gastar* en alimentos por cada 100 soles de ingreso e interpreta el intercepto en este contexto.

Ejercicio 1.25 (Interpretar coeficientes en R: de una a dos variables)**.** El archivo `salarios.csv` tiene 120 trabajadores: `salario` (S/.), `educacion`, `experiencia` y `region`. Estimaremos el salario **añadiendo regresores de a poco** para leer *qué representa cada coeficiente*. La nube salario vs. educación con la recta simple (M1):



Completa el código

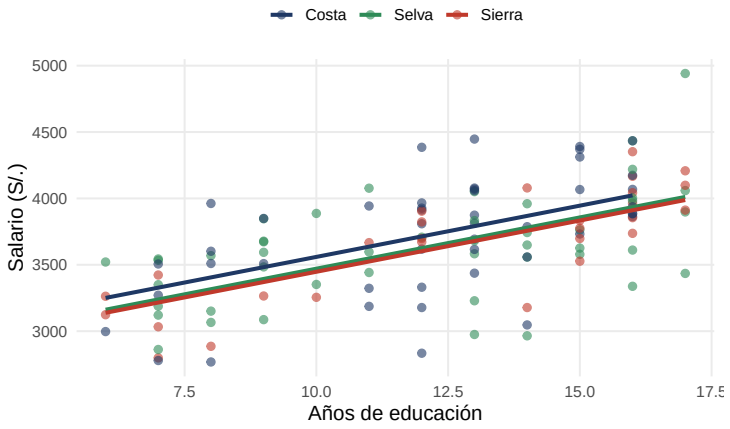
```
datos <- read.csv("salarios.csv")
m1 <- lm(salario ~ ____, data = datos)              # una variable (solo
    educacion)
summary(m1)                  # coeficiente, p-value, R^2
m2 <- lm(salario ~ educacion + ____, data = datos)   # dos variables (+
    experiencia)
summary(m2)
```

(a) Completa las dos casillas.

(b) ¿Qué representa el coeficiente de `educacion` en **M1** y en **M2**? ¿Cuál es *ceteris paribus*? ¿Son significativos (mira el p-value)?

(c) El coeficiente de `educacion` pasa de ≈ 77 (M1) a ≈ 197 (M2). ¿Por qué cambia tanto al agregar la experiencia?

Ejercicio 1.26 (Dummy y modelo mixto en R: cada coeficiente y el F-test)**.** Seguimos con `salarios.csv`. Ahora usamos la variable *categórica* `region` (Costa, Sierra, Selva) como **dummy**, y luego mezclamos todo. Con dummies las rectas por región son **paralelas** (mismo slope, distinto intercepto):



Completa el código

```
datos <- read.csv("salarios.csv")
m3 <- lm(salario ~ educacion + ____(region), data = datos)   # dummy de region
summary(m3)                  # Costa = base; cada coef = diferencia vs Costa
m4 <- lm(salario ~ educacion + experiencia + factor(region), data = datos)  #
   mixto
summary(m4)
# F-test: las dummies de region, juntas, agregan algo?
m_sin <- lm(salario ~ educacion + experiencia, data = datos)
anova(m_sin, ____)           # compara sin region vs con region
```

(a) Completa las dos casillas.

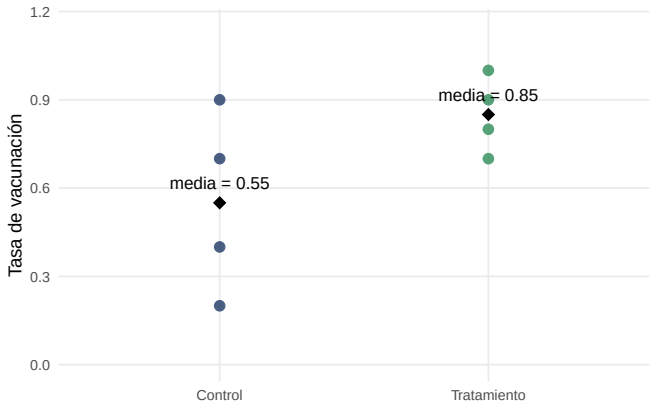
(b) ¿Qué representa el coeficiente de `Sierra`? ¿Respecto de quién se mide?

(c) ¿Las dummies de región son **conjuntamente** significativas? ¿Qué prueba usas y en qué se diferencia de mirar cada p-value por separado?

Nivel 4 – Reto

Difícil: integra varios temas en un mismo problema.

Ejercicio 1.27 (Prueba exacta de Fisher: enumerar todas las combinaciones (método MIT))**.** Un programa de salud se aplicó a 4 de 8 regiones (regiones $1, 3, 4, 8$); se mide la **tasa de vacunación** (`programa.csv`). La media es 0.85 en tratamiento y 0.55 en control, así que el **estimador del ATE** (*average treatment effect*) — la *difference in means* — vale 0.30. Ese 0.30 es el **observed test statistic** (`observed_test`); **no** es el p-valor:



¿Es 0.30 real o pudo salir por azar al elegir qué regiones tratar? Bajo la **sharp null** de Fisher (el programa no tiene efecto en *nadie*) cualquiera de las $\binom{8}{4} = 70$ asignaciones es igual de probable. Las **enumeramos todas** en una matriz de 0 y 1 y la **multiplicamos** por los outcomes (así, sin barajar al azar, obtenemos la *null distribution* exacta):

Completa el código

```
library(perm)
datos <- read.csv("programa.csv")     # region, grupo (T/C), tasa
A <- datos$tasa                       # outcome (vaccination rate)
n_treat <- sum(datos$grupo == "T")    # treated units = 4
# observed test statistic = |difference in means| (estima el ATE)
observed_test <- abs(mean(A[datos$grupo=="T"]) - mean(A[datos$grupo=="C"]))
    # 0.30

# todas las C(8,4)=70 asignaciones -> matriz de 0 y 1 (70 x 8)
perms <- chooseMatrix(8, ____)        # treated units
treatment_avg  <- perms %*% ____ / n_treat        # multiplica por A (los
    outcomes)
control_avg    <- (1 - perms) %*% A / n_treat
test_statistic <- abs(treatment_avg - control_avg)   # una por asignacion (
    sharp null)
pvalue <- mean(test_statistic >= ____)   # fraccion tan o mas extrema que
    observed_test
pvalue
```

(a) Completa las tres casillas.

(b) ¿Qué es cada fila de `perms`, y qué calcula `perms %*% A`?

(c) ¿`observed_test` es el ATE o el p-valor? Con $\alpha = 0.05$, ¿rechazas la *sharp null*? Concluye.

Ejercicio 1.28 (Diferencias en diferencias (diff-in-diff) en R)**.** Un programa social se aplica a un grupo *tratado* y no a un *control*; se observa a cada uno **antes** (`post=0`) y **después** (`post=1`). El archivo `did_panel.csv` tiene columnas `tratado` (0/1), `post` (0/1) y el resultado `y`. El estimador *diff-in-diff* resta la tendencia común (la línea roja punteada es el contrafactual del tratado):

Completa el código

```
datos <- read.csv("did_panel.csv")
# interaccion explicita (estilo Qian): 1 solo para tratados en el post
datos$tratadopost <- datos$tratado * ____   # multiplica por post
# DiD por regresion: el ATE es el coeficiente de la interaccion
m <- lm(y ~ tratado + post + ____, data = datos)   # agrega tratadopost
summary(m)
coef(m)["tratadopost"]      # estimador DiD
```

(a) Completa las dos casillas.

(b) Calcula el DiD "a mano" con las cuatro medias de celda y verifica que coincide con la interacción.

(c) ¿El efecto es significativo? ¿Qué supuesto clave necesita el DiD?

Ejercicio 1.29 (OLS "a mano" en R: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ y verificar con `lm`)**.** La idea es *ver que `lm` no es magia*: calculamos los coeficientes OLS con álgebra de matrices y confirmamos que dan lo mismo. Usamos `salarios.csv` (`salario`, `educacion`, `experiencia`).

Completa el código

```
datos <- read.csv("salarios.csv")
Y <- datos$salario
```

```
# matriz de diseno X: columna de 1s (intercepto) + regresores
X <- cbind(1, datos$educacion, ____)     # agrega experiencia
XtX <- t(X) %*% X
XtY <- t(X) %*% Y
beta_hat <- ____(XtX) %*% XtY            # invierte X'X con solve()
beta_hat
# verificar contra lm (mismos numeros):
coef(lm(salario ~ educacion + experiencia, data = datos))
```

(a) Completa las dos casillas.

(b) ¿Por qué la primera columna de $\mathbf{X}$ es de unos?

(c) ¿Coinciden `beta_hat` y `coef(lm(...))`? ¿Qué hace `lm` por dentro?

Ejercicio 1.30 (Derivar $\hat{\beta}_0, \hat{\beta}_1$ por condiciones de primer orden)**.** OLS elige $\hat{\beta}_0, \hat{\beta}_1$ minimizando la suma de cuadrados $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$.

(1) Deriva las dos *condiciones de primer orden* $\partial S/\partial\beta_0 = 0$ y $\partial S/\partial\beta_1 = 0$ (las *ecuaciones normales*).

(2) De la primera ecuación obtén $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$; de la segunda, sustituyendo $\hat{\beta}_0$, obtén $\hat{\beta}_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$.

Ejercicio 1.31 (Propiedades de los residuos: $\sum \hat{\varepsilon}_i = 0$ y $\sum X_i\hat{\varepsilon}_i = 0$)**.** A partir de las ecuaciones normales del problema anterior, demuestra dos propiedades de los residuos OLS y una consecuencia geométrica.

(1) Muestra que $\sum_i \hat{\varepsilon}_i = 0$ y que $\sum_i X_i\hat{\varepsilon}_i = 0$.

(2) Deduce que $\sum_i \hat{Y}_i\hat{\varepsilon}_i = 0$ (los ajustados son ortogonales a los residuos) y que el punto $(\bar{X}, \bar{Y})$ está en la recta.

Ejercicio 1.32 (Sesgo por variable omitida (OVB): fórmula y signo)**.** El modelo "largo" (correcto) es $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, pero por no observar Z (habilidad) se corre el modelo "corto" $Y = \alpha_0 + \alpha_1 X + u$. Sea δ la pendiente de regresar Z sobre X ($\delta = \widehat{\mathrm{Cov}}(X, Z)/\widehat{\mathrm{Var}}(X)$). Se puede mostrar que $\hat{\alpha}_1 \xrightarrow{p} \beta_1 + \beta_2\,\delta$.

(1) Identifica el *sesgo por variable omitida* y su signo si $\beta_2 > 0$ (la habilidad sube el salario) y $\delta > 0$ (X y Z covarían positivamente).

(2) Con $\beta_1 = 3$ (retorno causal real a la educación), $\beta_2 = 2$ y $\delta = 1.5$, calcula el coeficiente que reportaría la regresión corta. ¿Sobrestima o subestima?

(3) ¿En qué caso el sesgo es *cero* aunque Z importe ($\beta_2 \neq 0$)?

Ejercicio 1.33 (Invertir $\mathbf{X}^\top\mathbf{X}$ (2×2) y obtener $\hat{\boldsymbol{\beta}}$)**.** Para el modelo con intercepto $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, la matriz de diseño es $\mathbf{X}$ con primera columna de unos y segunda columna X_i. Con 4 datos $X = (1, 2, 3, 4)$, $Y = (3, 8, 11, 12)$:

(1) Calcula $\mathbf{X}^\top\mathbf{X}$ y $\mathbf{X}^\top\mathbf{Y}$.

(2) Invierte $\mathbf{X}^\top\mathbf{X}$ (matriz 2×2) y obtén $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.

(3) Verifica que $\hat{\beta}_1$ coincide con la fórmula escalar $\sum(X-\bar{X})(Y-\bar{Y})/\sum(X-\bar{X})^2$.

Ejercicio 1.34 (Insesgadez de OLS en el modelo simple)**.** En el modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ con los X_i fijos (o condicionando en ellos) y $\mathbb{E}[\varepsilon_i] = 0$, demuestra que $\hat{\beta}_1$ es *insesgado*: $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

(1) Escribe $\hat{\beta}_1 = \sum_i w_i Y_i$ con pesos $w_i = \dfrac{X_i - \bar{X}}{\sum_j (X_j - \bar{X})^2}$, y verifica $\sum_i w_i = 0$ y $\sum_i w_i X_i = 1$.

(2) Sustituye $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ y toma esperanza para concluir $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

Nivel 5 – Reto+

Muy difícil / fuera del scope típico: demostraciones y casos límite.

Ejercicio 1.35 (Derivación matricial: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$)**.** En notación matricial el modelo es $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ con $\mathbf{X}$ de $n\times(k{+}1)$ y rango columna completo. OLS minimiza $S(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

(1) Expande $S(\boldsymbol{\beta})$ y deriva respecto a $\boldsymbol{\beta}$ para obtener las *ecuaciones normales* $\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{Y}$.

(2) Despeja $\hat{\boldsymbol{\beta}}$ y di qué supuesto garantiza que la inversa exista. ¿Por qué el mínimo es global (convexidad)?

Ejercicio 1.36 (Insesgadez y varianza de $\hat{\boldsymbol{\beta}}$ en forma matricial)**.** Bajo el modelo clásico, con $\mathbf{X}$ fija (o condicionando en ella), $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ y $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\mathbf{I}_n$.

(1) Demuestra que $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ (insesgadez).

(2) Demuestra que $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

(3) Aplica al dataset $X = (1, 2, 3, 4)$, $Y = (3, 8, 11, 12)$ (Reto): allí $(\mathbf{X}^\top\mathbf{X})^{-1} = \left(\begin{smallmatrix} 3/2 & -1/2 \\ -1/2 & 1/5 \end{smallmatrix}\right)$ y la SSR resultó 4 ($n = 4$). Estima σ^2 con $s^2 = \mathrm{SSR}/(n-2)$ y da $\widehat{\mathrm{Var}}(\hat{\beta}_0)$, $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ y $\widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$.

Ejercicio 1.37 (Frisch–Waugh–Lovell: "partialling out")**.** El teorema de Frisch–Waugh–Lovell (FWL) dice que, en $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, el coeficiente $\hat{\beta}_1$ de la regresión múltiple se puede obtener así: (i) regresar X_1 sobre el resto $(1, X_2)$ y guardar los residuos $\tilde{X}_1$; (ii) regresar Y sobre el resto y guardar $\tilde{Y}$ (o usar Y directamente); (iii) regresar $\tilde{Y}$ sobre $\tilde{X}_1$.

(1) Escribe la fórmula resultante $\hat{\beta}_1 = \dfrac{\sum_i \tilde{X}_{1i} Y_i}{\sum_i \tilde{X}_{1i}^2}$ y explica en palabras qué significa "*partialling out*" (parcializar) X_2.

(2) Usa FWL para explicar *por qué* en una regresión múltiple el coeficiente de X_1 es su efecto "manteniendo constante X_2", y cómo se conecta con el sesgo por variable omitida del Nivel 4.

Ejercicio 1.38 (Multicolinealidad perfecta $\Rightarrow \mathbf{X}^\top\mathbf{X}$ singular)**.** Suponga que un regresor es combinación lineal exacta de otros, p. ej. se incluyen X_1 = estatura en cm y X_2 = estatura en metros ($X_2 = 0.01\, X_1$), además del intercepto.

(1) Argumenta por qué la matriz $\mathbf{X}$ *no* tiene rango columna completo y, por tanto, $\mathbf{X}^\top\mathbf{X}$ es **singular** (no invertible), de modo que $\hat{\boldsymbol{\beta}}$ no es único.

(2) Da una intuición de por qué el problema es de *identificación*: ¿qué familia de coeficientes (β_1, β_2) produce *exactamente* el mismo ajuste? Relaciónalo con la "trampa de las dummies" (Nivel 3) y con el supuesto de identificación de L18 ($\mathbf{X}$ de rango completo).

Fuentes y atribución

Material de repaso **basado en MIT 14.310x – Data Analysis for Social Scientists** (Esther Duflo & Sara Fisher Ellison), MIT OpenCourseWare, licencia **CC BY-NC-SA 4.0** (uso no comercial, con atribución, compartir bajo la misma licencia), y en las fuentes de la bibliografía. Los enunciados se basan en el material del curso y en diversas fuentes abiertas; cuando un problema se adapta de una fuente específica, se cita en el enunciado.

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`.