Apuntes - Estadística bidimensional y regresión lineal

Curso: 4 Eso

En este tema aprenderás: dos variables a la vez, tablas de contingencia, nube de puntos y correlación, covarianza y coeficiente de correlación, recta de regresión, predicciones y correlación no es causalidad. Cada idea va en un cuadro de color, con ejemplos resueltos paso a paso.

Actividad de introducción: ¿estudiar más sirve para sacar mejor nota?

Un profesor de un instituto de Salamanca anota para seis alumnos las horas que estudian a la semana y la nota que obtienen: $(1,3), (2,4), (3,5), (4,6), (5,8), (6,9)$. Si colocas esos pares como puntos en unos ejes, observarás que forman una nube que sube de izquierda a derecha: a más horas, mejor nota. Pero ¿cómo de fuerte es esa relación? ¿Se puede predecir la nota de un alumno que estudia $4{,}5$ horas? Hasta ahora has estudiado una sola variable cada vez. Ahora estudiarás **dos variables a la vez**: cómo se organizan los datos en una tabla de doble entrada, cómo se mide la relación entre ellas con la covarianza y el coeficiente de correlación y cómo se usa la recta de regresión para hacer predicciones razonables.

Dos variables a la vez

Propiedad

En la **estadística bidimensional** se observan dos variables X e Y sobre los mismos individuos: horas de estudio y nota, edad de un coche y precio, estatura y peso. Los datos son pares (x_i, y_i). Si las variables son cualitativas o están agrupadas se usa una **tabla de contingencia** (de doble entrada); si son numéricas se representan en una **nube de puntos** (diagrama de dispersión).

Tablas de contingencia

En una encuesta a 100 alumnos se anotó el deporte y el sexo:

	Chicos	Chicas	Total
Fútbol	30	10	40
Baloncesto	15	25	40
Ninguno	5	15	20
Total	50	50	100

Definición: distribuciones marginales

Los totales son las **distribuciones marginales**. Una **distribución condicionada** se obtiene fijando una categoría: entre las chicas, $\frac{25}{50} = 50\,\%$ practican baloncesto, frente a $\frac{15}{50} = 30\,\%$ de los chicos. Como los porcentajes cambian, las variables están **relacionadas**: si fueran independientes, el porcentaje sería el mismo en cada grupo. La frecuencia relativa conjunta de «chica y baloncesto» es $\frac{25}{100} = 0{,}25$.

Nube de puntos y correlación

Observación

La nube de puntos muestra si existe relación entre las variables y qué forma tiene. La **correlación** es lineal si los puntos se agrupan alrededor de una recta.



Propiedad: nube de puntos y correlación

- **Positiva:** al aumentar x, y tiende a aumentar.
- **Negativa:** al aumentar x, y tiende a disminuir.
- **Nula:** los puntos no siguen ningún patrón lineal.
- Es más **fuerte** cuanto más cerca de la recta están los puntos.

Covarianza y coeficiente de correlación

Propiedad

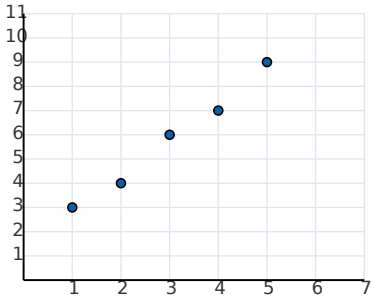
Con las medias $\bar{x}$, $\bar{y}$ y las desviaciones típicas σ_x, σ_y, se define la **covarianza**

$$\sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$$

y el **coeficiente de correlación lineal** $r = \dfrac{\sigma_{xy}}{\sigma_x\,\sigma_y}$. Siempre se cumple $-1 \le r \le 1$. El signo de r es el de la covarianza; si $|r|$ es próximo a 1 la relación lineal es fuerte; si es próximo a 0, débil o nula.

Ejemplo 1

Horas de estudio $x = 1, 2, 3, 4, 5$ y nota $y = 3, 4, 6, 7, 9$. $\bar{x} = 3$, $\bar{y} = 5{,}8$. $\sum xy = 102$, luego $\sigma_{xy} = 20{,}4 - 3 \cdot 5{,}8 = 3$. $\sigma_x^2 = 11 - 9 = 2$ ($\sigma_x \approx 1{,}41$) y $\sigma_y^2 = 38{,}2 - 33{,}64 = 4{,}56$ ($\sigma_y \approx 2{,}14$). Por tanto $r = \frac{3}{1{,}41 \cdot 2{,}14} \approx 0{,}99$: correlación positiva muy fuerte.



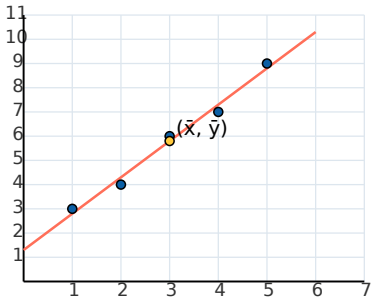
Recta de regresión

Definición: recta de regresión de y sobre x

La **recta de regresión de y sobre x** es la recta que mejor se ajusta a la nube (minimiza los cuadrados de las distancias verticales). Pasa por el punto $(\bar{x}, \bar{y})$ y su pendiente es $b = \dfrac{\sigma_{xy}}{\sigma_x^2}$:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}\,(x - \bar{x}).$$

En el ejemplo: $b = \frac{3}{2} = 1{,}5$ y $y - 5{,}8 = 1{,}5(x - 3)$, es decir, $y = 1{,}5x + 1{,}3$. La pendiente se interpreta: por cada hora más de estudio, la nota sube de media 1,5 puntos.



Cómo se hace: obtener la recta de regresión de y sobre x

1. Calcula $\bar{x}$, $\bar{y}$, σ_x^2 y σ_{xy}.
2. Halla la pendiente $m = \dfrac{\sigma_{xy}}{\sigma_x^2}$.
3. Escribe $y - \bar{y} = m(x - \bar{x})$ y despeja y.
4. Calcula r para saber si la recta es fiable antes de predecir.

Predicciones

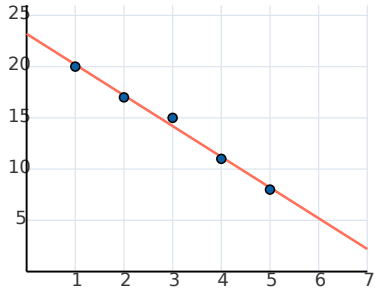
Propiedad

La recta permite estimar y para un valor de x. Con $x = 2{,}5$ h: $y = 1{,}5 \cdot 2{,}5 + 1{,}3 = 5{,}05$. Es una **interpolación** (el valor está dentro del rango de datos) y suele ser fiable si $|r|$

es grande. Para $x = 8$ h daría $13{,}3$, una nota imposible: es una **extrapolación** (fuera del rango), poco fiable. La fiabilidad aumenta con $|r|$ y con el número de datos.

Ejemplo 2: correlación negativa

Antigüedad del coche (años) $x = 1, \ldots, 5$ y precio (miles de euros) $y = 20, 17, 15, 11, 8$: $\bar{x} = 3$, $\bar{y} = 14{,}2$, $\sigma_{xy} = -6$, $\sigma_x^2 = 2$, así que $b = -3$ y la recta es $y = -3x + 23{,}2$ y $r \approx -0{,}996$. Cada año de antigüedad resta unos 3 000 € de media.



Ejemplo 3: recta de regresión completa

Con los datos $x = 1, 2, 3, 4$ e $y = 3, 5, 6, 10$, halla la recta de regresión y predice y para $x = 5$.

$\bar{x} = 2{,}5$, $\bar{y} = 6$, $\sum xy = 71$, $\sum x^2 = 30$. Entonces $\sigma_{xy} = \frac{71}{4} - 15 = 2{,}75$ y $\sigma_x^2 = \frac{30}{4} - 6{,}25 = 1{,}25$. La pendiente es $m = \frac{2{,}75}{1{,}25} = 2{,}2$ y la recta es $y - 6 = 2{,}2(x - 2{,}5)$, es decir, $y = 2{,}2x + 0{,}5$. Para $x = 5$, $y = 11{,}5$. Como $r \approx 0{,}96$, la correlación es fuerte y la predicción es fiable (interpolar o extrapolar poco).

Correlación no es causalidad

Observación

Que dos variables estén correlacionadas no significa que una cause la otra. Las ventas de helados y las visitas a la playa suben a la vez en verano porque ambas dependen de la temperatura (**variable oculta**). Para afirmar una causa hace falta un estudio específico.

¡Atención! Errores frecuentes

- Confundir una correlación fuerte con una causa.
- Usar la recta fuera del rango de datos sin avisar de que es una extrapolación.
- Interpretar r como pendiente: r mide la fuerza de la relación, no su inclinación.
- Olvidar que la recta de regresión de y sobre x no es la misma que la de x sobre y .

Cómo leer una correlación

Propiedad: el coeficiente de correlación lineal

El coeficiente $r = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ cumple $-1 \leq r \leq 1$. Si r está cerca de 1 hay una relación lineal creciente fuerte; cerca de -1, decreciente fuerte; cerca de 0, no hay relación lineal. Como orientación, $|r| > 0{,}8$ suele considerarse correlación fuerte, entre $0{,}5$ y $0{,}8$ moderada y por debajo de $0{,}5$ débil. El signo de r coincide con el de la covarianza y con el de la pendiente de la recta de regresión.

Ejemplo 2

Dos variables tienen $\sigma_x = 3$, $\sigma_y = 5$ y $\sigma_{xy} = -12$. Calcula r e interprétalo.

$r = \dfrac{-12}{3 \cdot 5} = -0{,}8$: correlación negativa y fuerte. Cuando una variable aumenta, la otra disminuye con bastante regularidad.

Método: estudio completo de dos variables

1. Ordena los datos en una tabla con columnas x_i, y_i, x_i^2, y_i^2 y $x_i y_i$.
2. Calcula las sumas y las medias $\bar{x}$ e $\bar{y}$.
3. Calcula las varianzas σ_x^2 y σ_y^2 y la covarianza σ_{xy}.
4. Calcula r y decide si la relación lineal es fuerte.
5. Si lo es, escribe la recta de regresión $y - \bar{y} = \dfrac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$.
6. Predice solo dentro del rango de los datos o con mucha prudencia fuera de él.

¡Atención! Errores y límites de la regresión

- **Correlación no es causalidad**: las ventas de helados y los ahogamientos aumentan en verano, pero los helados no causan ahogamientos; la causa común es el calor.
- $r = 0$ solo dice que no hay relación **lineal**: una parábola puede tener $r = 0$.
- La extrapolación lejos del rango de datos puede dar resultados absurdos, como precios negativos.
- Un dato atípico puede cambiar mucho r y la recta: revisa siempre la nube de puntos.
- La recta de y sobre x no es la misma que la de x sobre y, aunque ambas pasan por $(\bar{x}, \bar{y})$.

Ejemplo 3

Una recta de regresión es $y = 2{,}5x + 4$ y una muestra tiene x entre 10 y 30.

Predice y para $x = 20$ y para $x = 100$ y comenta.

Para $x = 20$: $y = 50 + 4 = 54$, una predicción fiable porque 20 está dentro del rango. Para $x = 100$: $y = 254$, que es una extrapolación muy lejos del rango y no se puede garantizar que la relación siga siendo lineal.

¿Sabías que?

El inglés Francis Galton, primo de Darwin, estudió en el siglo XIX la relación entre la estatura de los padres y de los hijos y observó que los hijos de padres muy altos solían ser altos, pero algo menos que sus padres: lo llamó «regresión a la media», y de ahí viene el nombre de la recta de regresión. Su discípulo Karl Pearson definió después el coeficiente de correlación, que lleva su nombre.

Organizar los cálculos en una tabla

Método: tabla de cálculo para la regresión

1. Escribe los datos en dos columnas, x_i e y_i, y añade tres columnas más: x_i^2, y_i^2 y $x_i y_i$.
2. Suma cada columna. Con las cinco sumas y el número de datos n puedes calcularlo todo.
3. Calcula las medias, las varianzas (media de los cuadrados menos el cuadrado de la media) y la covarianza (media de los productos menos el producto de las medias).
4. Obtén r y decide si tiene sentido usar la recta: solo si $|r|$ es próximo a 1.
5. Escribe la recta de regresión y úsala para predecir solo dentro del rango de los datos.

Ejemplo 4

Con los datos $(1, 3)$, $(2, 5)$, $(3, 7)$, calcula la covarianza.

Las medias son $\bar{x} = \dfrac{6}{3} = 2$ e $\bar{y} = \dfrac{15}{3} = 5$. Los productos $x_i y_i$ son 3, 10 y 21, que suman 34. La covarianza es $\sigma_{xy} = \dfrac{34}{3} - 2 \cdot 5 \approx 11{,}33 - 10 = 1{,}33$. Es positiva, lo que confirma lo que se ve en los datos: cuando x aumenta, y también aumenta. De hecho, los tres puntos están sobre la recta $y = 2x + 1$, así que la correlación es perfecta.

El consejo de Otto

Antes de calcular nada, dibuja la nube. Si los puntos forman una curva o hay un dato que se escapa, la recta de regresión no es la mejor herramienta.