Taller Iniciación Estadística Aplicada a la Investigación

Regresión

Hugo J. Bello

2024/04

Ejemplo: Datos

Paciente	Peso al nacer	Longitud al nacer	CI a los 5 años
1.0	2.49	36.2	101.97
2.0	2.57	29.87	100.22
3.0	2.55	26.07	103.98
4.0	2.52	35.35	101.13
5.0	1.89	42.43	97.94
6.0	2.74	34.54	109.26
⋮	⋮	⋮	⋮
39.0	2.13	37.14	103.63
40.0	3.11	41.08	107.44

Estudio de la influencia mutua de dos variables cuantitativas

Los datos anteriores se corresponden con los de un estudio de varias medidas de bebes recién nacidos y su coeficiente de inteligencia al cabo de 5 años.

¿Cómo podemos medir la influencia de, por ejemplo la variable *Longitud al nacer* en la variable *CI a los 5 años*?
Estudiar las medias o varianzas de estas variables por separado nos haría perder perspectiva de que hay una correspondencia entre ellas. Para ello podemos usar la **covarianza** o el **coeficiente de correlación de Pearson**

Covarianza

La **covarianza** es un valor que indica el grado de variación conjunta de dos variables respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es necesario para estimar otros parámetros, como el coeficiente de correlación lineal.
Supongamos que tenemos unos datos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_N, y_N)$$

Definimos la covarianza como:

$$S_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

$$= \frac{1}{N}((x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y}))$$

donde $\overline{x}$ denota la media de la primera variable (x), e $\overline{y}$ denota la media de la segunda (y)

Ejemplo anterior: covarianza

Para los datos anteriores, tomando $x =$ Peso al nacer y $y =$ CI a los 5 años, obtenemos $S_{xy} = 1.7$
La covarianza no siempre es fácil de interpretar. Se puede afirmar:

- Si $S_{xy} > 0$ hay una correspondencia lineal positiva, esto es, cuanto mayor es x mayor es y
- Si $S_{xy} < 0$ hay una correspondencia lineal negativa, esto es, cuanto mayor es x menor es y
- SI $S_{xy} \cong 0$. No hay correspondencia entre x e y

En este caso solo vemos que hay una correspondencia positiva, pero no tenemos noción de cómo de fuerte es esa correspondencia. Para esto tenemos que usar la correlación de Pearson.

Coeficiente de correlación de Pearson

El coeficiente de **correlación de Pearson** es una medida de dependencia lineal entre dos variables estadísticas cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.
Se define como

$$\rho_{XY} = \frac{S_{xy}}{S_X S_Y}$$

donde S_{xy} de nota la covarianza, S_X denota la desviación típica de la primera variable y S_Y la desviación típica de la segunda.

Si $\rho_{XY} > 0$ y cercano a 1 hay dependencia lineal directa (**positiva**), es decir, a grandes valores de X corresponden grandes valores de Y.

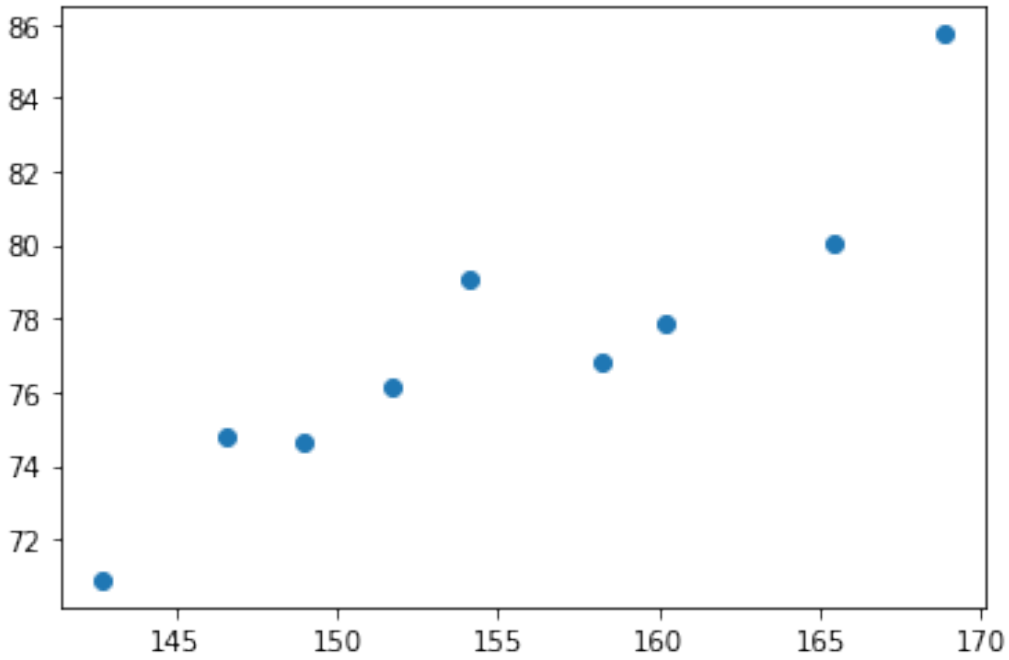


Figure: nube de puntos correlación positiva

Si ρ_{XY} es cercano a 0 se interpreta como la **no existencia de una relación lineal** entre las dos variables.

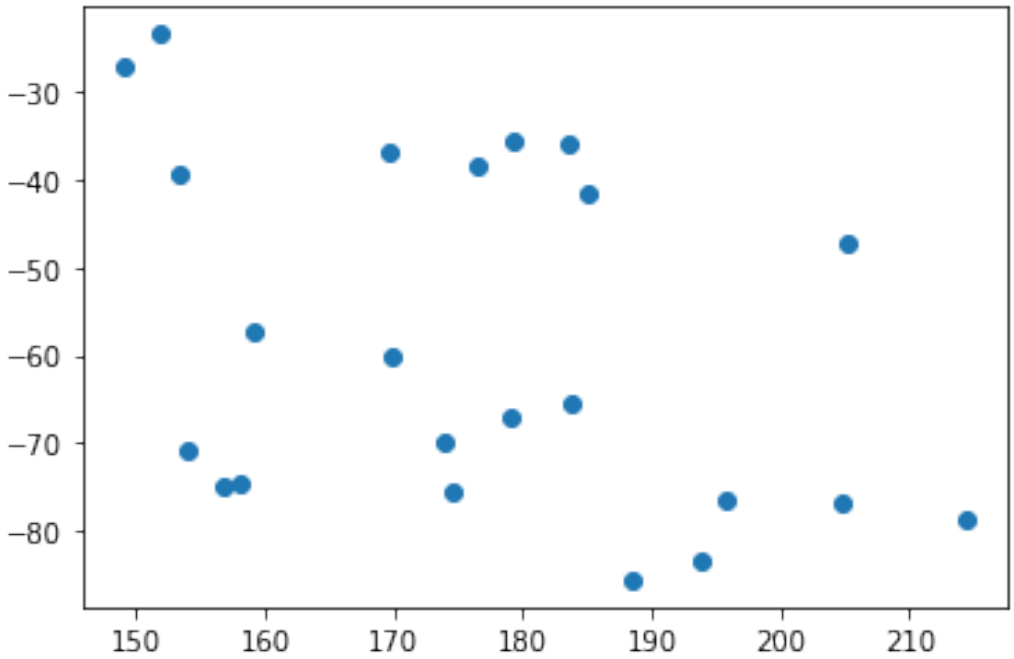


Figure: nube de puntos correlación 0

Si $\rho_{XY} < 0$ **y cercano a** -1 hay dependencia lineal inversa o **negativa**, es decir, a grandes valores de X corresponden pequeños valores de Y.

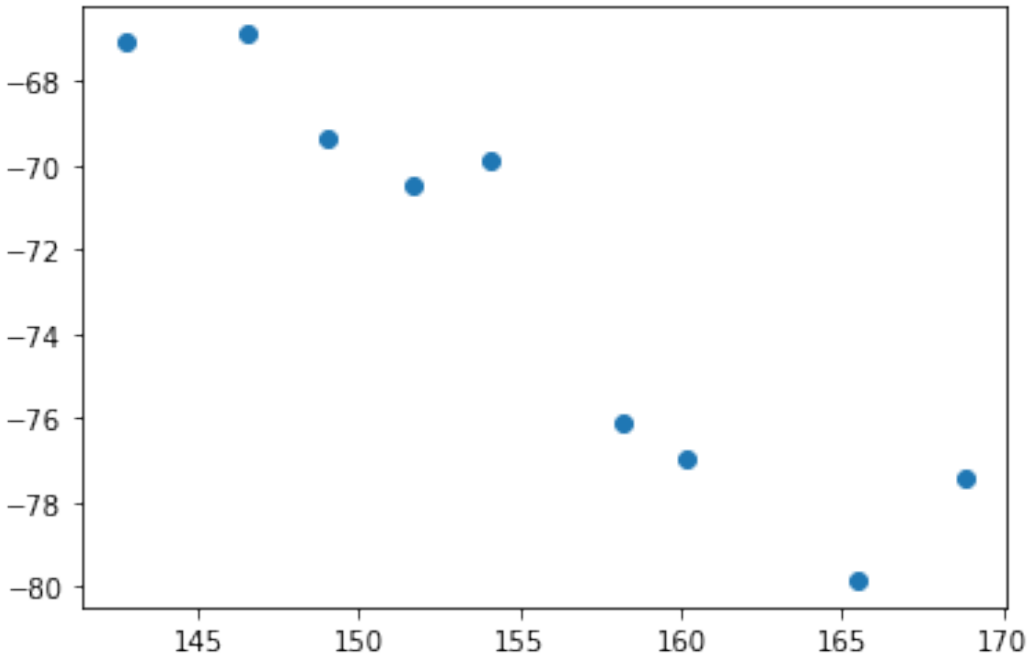


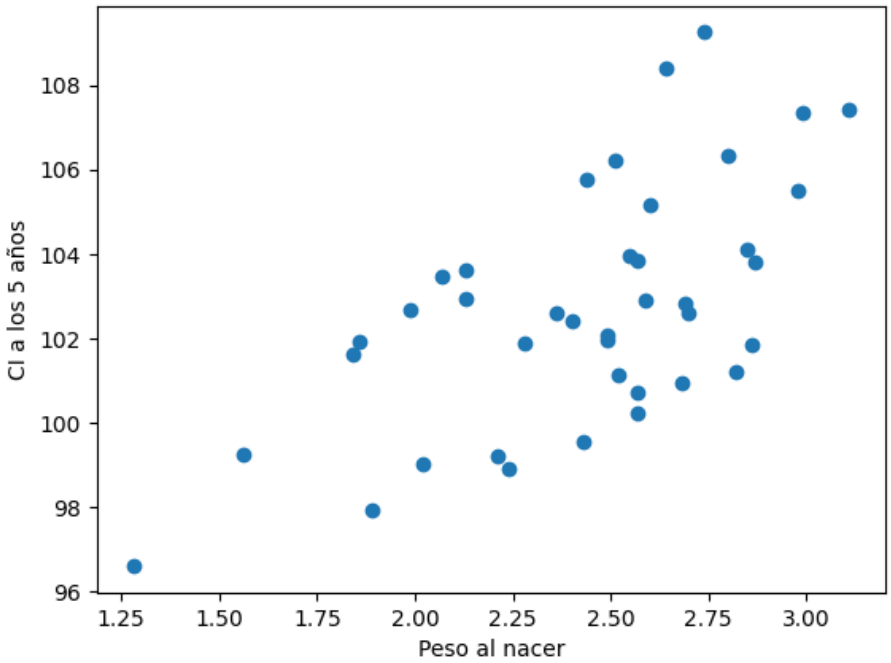
Figure: nube de puntos correlación negativa

En el ejemplo anterior

Para los datos anteriores, tomando $x =$ Peso al nacer y $y =$ CI a los 5 años.

$$\rho_{xy} = 0.62$$

Si representamos los datos obtenemos



En el ejemplo anterior

En este caso si es más fácil interpretarlo, observamos que hay una correspondencia **directa o positiva** entre ambas variables considerable. Lo siguiente que haremos será intentar utilizar un *modelo* para explicar la variable y a partir de x. Este modelo cuantificará la contribución de una variable en la otra

Modelo de regresión simple

Para unos datos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_N, y_N)$$

El **modelo de regresión simple** es un modelo lineal de la forma

$$\hat{Y} = a + bX$$

Es decir, trata de encontrar dos parámetros a, b para aproximar Y mediante X de la forma anterior. Los valores utilizados en el modelo son

$$a = \overline{y} - \frac{S_{XY}}{S_X^2}\overline{x}$$

$$b = \frac{S_{XY}}{S_X^2}$$

Modelo de regresión simple

La manera en que el modelo obtiene los parámetros a y b es mediante el **método de mínimos cuadrados**.

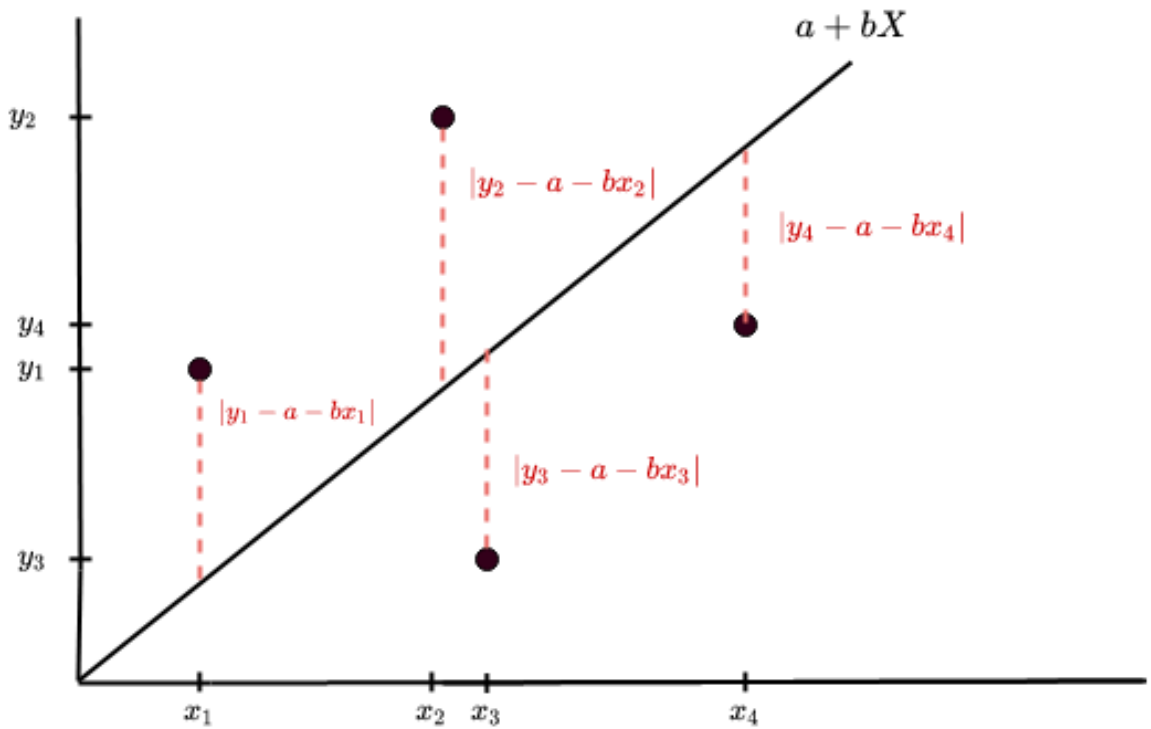


Figure: Mínimos cuadrados

En el ejemplo anterior

Aplicando el modelo de regresión simple obtenemos

$$a = 4.49$$

$$b = 91.79$$

Es decir

$$\widehat{y} = 4.49 \cdot x + 91.79$$

Escrito de otra manera

$$\text{CI a los 5 años} = 4.49 \cdot (\text{Peso al nacer}) + 91.79$$

Gracias a estos resultados podemos interpretar más precisamente cuanto afecta el peso al nacer al CI. No solo eso, podríamos usarlo para estimar el CI a los 5 años de un niño que ha nacido con cualquier valor x de peso.

En el ejemplo anterior

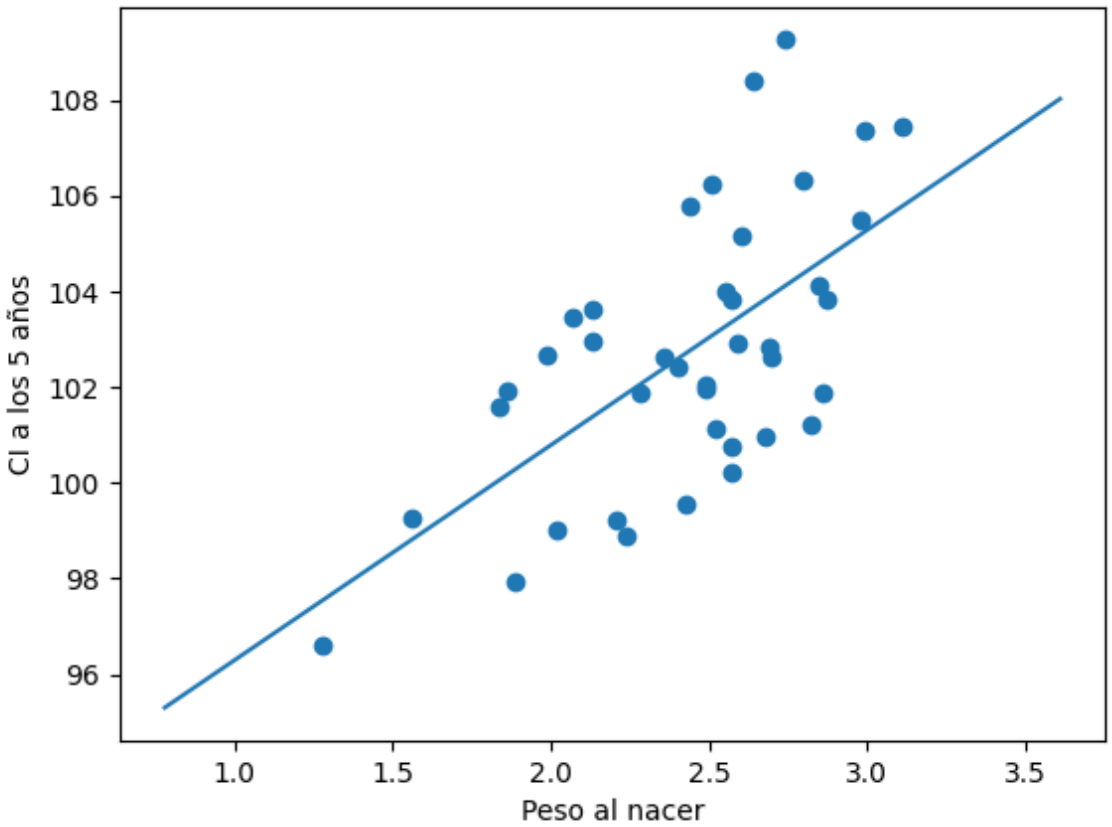


Figure: Recta de regresión

Modelo de regresión múltiple

Para unos datos con p variables $X_1, \ldots X_p$ y otra variable Y, el **modelo de regresión múltiple** es un modelo lineal de la forma

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Es decir, trata de encontrar parámetros $\beta_0, \ldots, \beta_p$ para aproximar Y mediante $X_1, \ldots, X_p$ de la forma anterior.

En el ejemplo anterior

Si tomamos X_1 =Peso al nacer, X_2 =Longitud al nacer, Y =CI a los 5 años. Aplicando el modelo de regresión simple obtenemos

$$\beta_0 = 85.36$$

$$\beta_1 = 4.68$$

$$\beta_2 = 0.16$$

Es decir

$$\widehat{Y} = 4.68 \cdot X_1 + 0.16 \cdot X_2 + 85.36$$

Escrito de otra manera

$$(\text{CI 5 años}) = 4.49 \cdot (\text{Peso al nacer}) + 0.16 \cdot (\text{Long. al nacer}) + 85.36$$