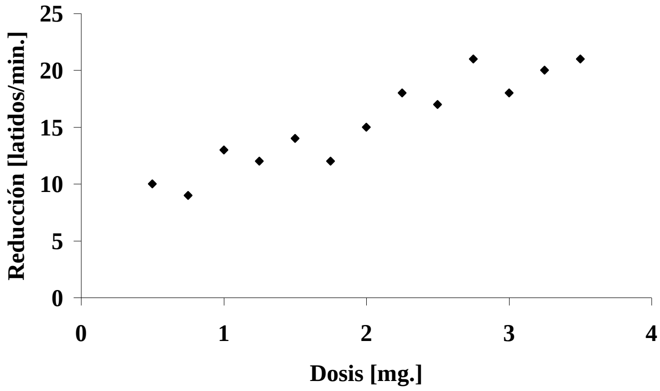
Regresión Lineal Simple

Método de mínimos cuadrados

Se realizó un experimento con el fin de estudiar el efecto de un medicamento en bajar la frecuencia cardiaca. La variable independiente es la dosis (mg.) del medicamento y la dependiente es la diferencia de la frecuencia cardiaca antes y después de la administración del medicamento (latidos / min.).

Dosis	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3	3.25	3.5
Reducción	10	9	13	12	14	12	15	18	17	21	18	20	21



Llamando x_i a la dosis e y_i a la reducción (i = 1,…,n) con n =13, se trata de ajustar una recta a los datos. Partimos de la suposición de que éstos responden a un *modelo lineal*:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

donde β_0, β_1 son parámetros desconocidos, las u_i (i = 1,…,n) son variables aleatorias (los "errores"). Entonces los parámetros se estiman mediante el método de *mínimos cuadrados*. Sean

$$r_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

(los residuos). Entonces el método consiste en hallar $\hat{\beta}_0, \hat{\beta}_1$ tales que

$$\sum_{i=1}^{n} r_i^2 = \text{mín}.$$

La solución se obtiene de la siguiente manera. Sean $\bar{x}$ e $\bar{y}$ las medias de la x y las y, y sean

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad , \quad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

y

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

Entonces la solución es

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1$$

La recta obtenida se llama *recta de regresión* de y en x. En nuestro ejemplo,

$$\hat{\beta}_0 = 7{,}473\,, \quad \hat{\beta}_1 = 3{,}956$$

El número

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

se llama *coeficiente de correlación* entre x e y, y mide la bondad del ajuste. Está entre –1 y 1; si $|R| = 1$, los puntos están exactamente sobre una recta, con pendiente del signo de R. En nuestro ejemplo, R = 0,941.

Intervalos de confianza para los parámetros

Supongamos que las variables aleatorias u_i son independientes con media 0 y desviación σ. Entonces σ se estima con s_r definido como

$$s_r^2 = \frac{S_{rr}}{n-2}, \quad \text{con} \quad S_{rr} = \sum_{i=1}^{n} S_{yy} - S_{xx}\hat{\beta}_1^2$$

En nuestro ejemplo, s_r = 1,448. Las desviaciones de los estimadores son

$$\mathrm{dt}(\hat{\beta}_0) = \sigma a_0\,, \quad \mathrm{dt}(\hat{\beta}_1) = \sigma a_1\,,$$

donde

$$a_0 = \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad a_1 = \sqrt{\frac{1}{S_{xx}}}$$

Suponiendo además las u_i normales, los intervalos de confianza de nivel α (por ejemplo α 0,95) para β_0 y β_1 son respectivamente

$$\hat{\beta}_0 \pm t_{1-\frac{\alpha}{2}} s_r a_0, \quad \hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}} s_r a_1,$$

donde t se obtiene de la tabla de Student con n-2 grados de libertad. Observe que lo que multiplica a t son las desviaciones estimadas de $\hat{\beta}_0$ y $\hat{\beta}_1$ respectivamente.

En nuestro caso es

$$a_0 = 0{,}655 \ , \ a_1 = 0{,}296,$$

y para el nivel β = 0,95, tomando de la tabla con n-2 = 11 grados de libertad $t_{1-\frac{\alpha}{2}} = 2{,}2$, los intervalos para β_0 y β_1 son respectivamente

$$7{,}473 \pm 2{,}085$$

$$\text{y}$$

$$3{,}956 \pm 0{,}944$$