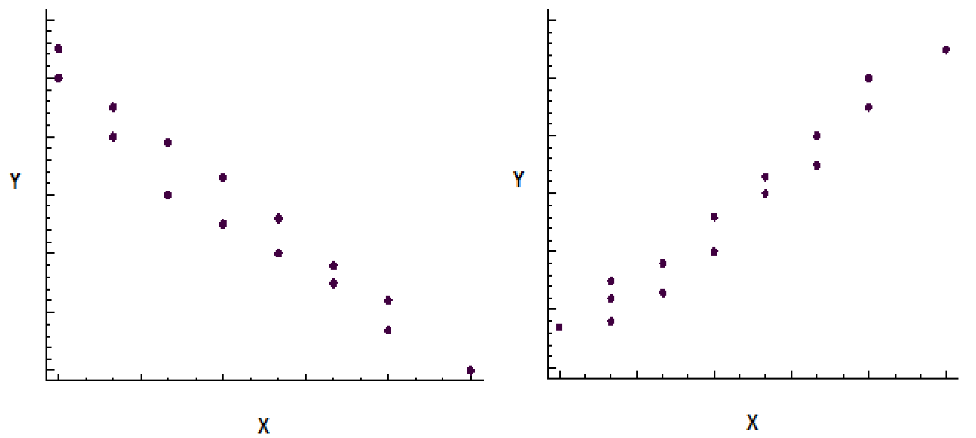
Tema 9. Regresión y correlación

Dada una muestra de n pares de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$ de la variable bidimensional (X, Y), el objetivo va a ser encontrar una curva, lo más sencilla posible, que exprese la relación entre las variables X e Y.

Por ejemplo, supongamos que la nube de puntos que se obtiene es de una de las dos formas indicadas en la figura 1, parece razonable pensar que existe una relación lineal entre los valores de X e Y.

Figura 1:



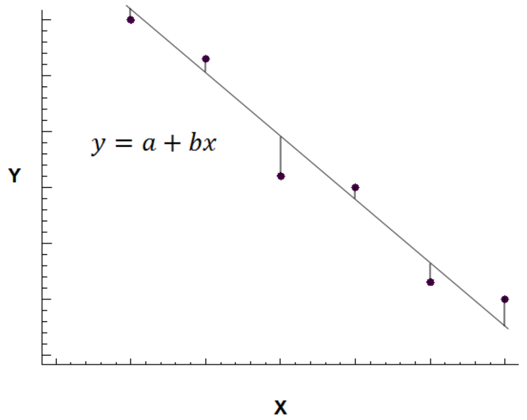
La recta de regresión

Recta de regresión de Y sobre X. Definición. Se llama recta de regresión de Y sobre X, a la recta $y = a + bx$ que minimiza el error cuadrático medio (ECM) definido como:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{1}$$

Lo que se intenta es encontrar la recta que mejor representa a la nube de puntos, en el sentido de minimizar la media de los cuadrados de las distancias verticales de cada punto de la nube a la recta (ver figura 2).

Figura 2:



A continuación se obtiene la recta de regresión, es decir, la recta que minimiza el error cuadrático medio que es función de las variables a y b.

En primer lugar desarrollando el cuadrado se tiene que

1 n

Derivando con respecto a cada variable e igualando a cero se obtiene el siguiente sistema de ecuaciones:

$$\begin{aligned} \frac{\partial(ECM)}{\partial a} &= \frac{1}{n}\left(2na - 2\sum_{i=1}^{n} y_i + 2b\sum_{i=1}^{n} x_i\right) = 0 \\ \frac{\partial(ECM)}{\partial b} &= \frac{1}{n}\left(2b\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i\right) = 0 \end{aligned}$$

cuya solución es

$$a = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x} \qquad \text{y} \qquad b = \frac{s_{xy}}{s_x^2} \tag{2}$$

donde:

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ es la media de la muestra $y_1, \ldots, y_n$

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ son, respectivamente, la media y la varianza de la muestra $x_1, \ldots, x_n$

y $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ es la covarianza muestral entre las observaciones de X e Y.

Se puede comprobar que esta solución corresponde a un mínimo de la función ECM. Por tanto, **la recta de regresión de Y sobre X** es:

$$y = a + bx = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x} + \frac{s_{xy}}{s_x^2}x$$

o también:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

A el coeficiente $b = \frac{s_{xy}}{s_x^2}$ se le llama **coeficiente de regresión de Y sobre X**.

Según como sea la nube de puntos, la recta de regresión la representará mejor o peor, lo que se medirá mediante el error cuadrático medio cometido.

Sustituyendo en (1) los valores de a y b obtenidos en (2), se tiene que la varianza residual es:

$$\text{Varianza residual} = s_y^2 \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right)$$

siendo $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$ la varianza de la muestra $y_1, \ldots, y_n$.

El cociente que aparece en esta expresión se define a continuación.

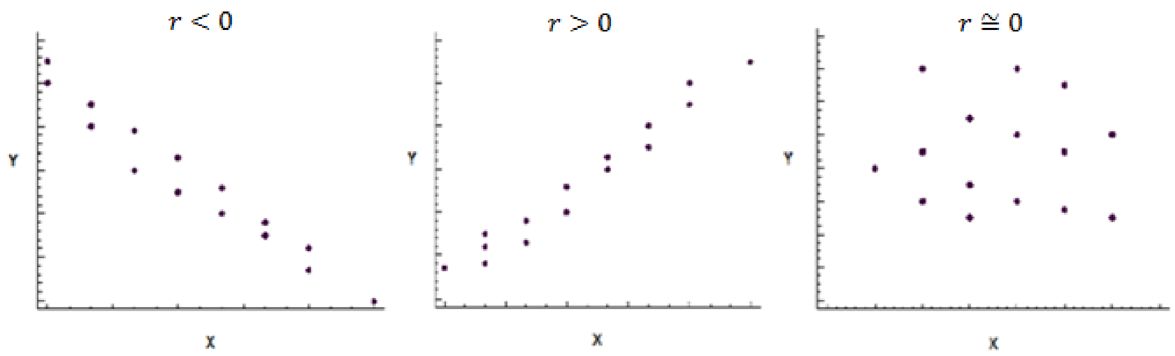
Coeficiente de correlación muestral entre X e Y. Definición. Se define el coeficiente de correlación muestral entre X e Y, (r), como:

$$r = \frac{s_{xy}}{s_x\, s_y}$$

Por lo que también se puede escribir la varianza residual como sigue:

$$\text{Varianza residual} = s_y^2 \left(1 - r^2\right)$$

Figura 3:



puesto que la varianza residual, al ser una suma de cuadrados, no puede ser negativa.

- Hay una cierta asociación entre el valor de r y la orientación de la nube de puntos. $r > 0$ corresponde a una recta creciente, $r < 0$ corresponde a una recta decreciente y si $r = 0$ se obtiene la recta $y = \overline{y}$ (Véase la figura 3).

- Los casos $r = \pm 1$ corresponden a una varianza residual nula, lo que indica que los puntos se encuentran exactamente sobre la recta calculada.

Así, el ajuste de regresión será mejor cuanto más próximo se encuentre r a 1 o a -1 y peor cuanto más próximo se encuentre r a cero.

De manera similar, minimizando las distancias cuadráticas horizontales se obtiene **la recta de regresión de X sobre Y** como sigue:

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$

Las dos rectas tiene en común el punto $(\overline{x}, \overline{y})$.

A el coeficiente $b = \frac{s_{xy}}{s_y^2}$ se le llama **coeficiente de regresión de X sobre Y**.

Si b_{yx} y b_{xy} son los coeficientes de regresión de las rectas de regresión de Y sobre X y de X sobre Y, respectivamente, se cumple que:

- $b_{yx} b_{xy} = r^2$

- $b_{yx} = r\frac{s_y}{s_x}$ y $b_{xy} = r\frac{s_x}{s_y}$

La recta de regresión permite estimar valores de la variable dependiente y, que

Así, a cada valor x_i, de los observados, le corresponde el valor observado y_i y el valor estimado $\hat{y}_i$, siendo precisamente la diferencia cuadrática $(y_i - \hat{y}_i)^2$ la contribución del punto (x_i, y_i) a la varianza residual.

Ejemplo: Dada la siguiente ditribución bidimensional, calcular las dos rectas de regresión y el coeficiente de correlación lineal.

X:	0,7	1	2	3	3	4	5	6	7	8
Y:	2,2	2,2	2,5	2,7	2,8	3	3,3	3,4	4	4

Recta de regresión de Y sobre X: $y = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x} + \frac{s_{xy}}{s_x^2}x$.

Recta de regresión de X sobre Y: $x = \overline{x} - \frac{s_{xy}}{s_y^2}\overline{y} + \frac{s_{xy}}{s_y^2}y$.

Por tanto, necesitamos obtener: $\overline{x}$, $\overline{y}$, s_x^2, s_y^2 y s_{xy}:

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	0,7	2,2	0,49	4,84	1,54
	1	2,2	1	4,84	2,2
	2	2,5	4	6,25	5
	3	2,7	9	7,29	8,1
	3	2,8	9	7,84	8,4
	4	3	16	9	12
	5	3,3	25	10,89	16,5
	6	3,4	36	11,56	20,4
	7	4	49	16	28
	8	4	64	16	32
Total	39,7	30,1	213,49	94,51	134,14

La recta de regresión de Y sobre X:

$$y = 3,01 - \frac{1,464}{5,588}\,3,97 + \frac{1,464}{5,588}\,x \Rightarrow y = 1,97 + 0,262\,x$$

La recta de regresión de X sobre Y:

$$x = 3,97 - \frac{1,464}{0,391}\,3,01 + \frac{1,464}{0,391}\,y \Rightarrow x = -7,3 + 3,74\,y$$

El coeficiente de correlación: $r = \frac{s_{xy}}{s_x\, s_y} = \frac{1,464}{\sqrt{5,588}\,\sqrt{0,391}} = 0,99.$

Ejercicios

1. La tabla siguiente recoge las pulsaciones/minuto y la temperatura de 10 enfermos.

p/m:	70	65	80	60	75	85	70	65	80	85
Temperaturas:	36,5	36,5	37	36	37	37,5	37	36	37,5	37

Obtener las rectas de regresión y estimar la temperatura que tendría un enfermo con 72 p/m.

2. En un estudio dietético se pretende ver si existe o no relación entre entre la cantidad total inyectada durante un mes de una determinada droga y el aumento de peso provocado en una persona (se toman pacientes de unas características de edad y altura similares). En una muestra de 10 personas se obtuvieron los siguientes resultados:

¿Existe relación entre estas dos variables?. Si existe, obtener el modelo matemático que representa esta relación.

3. De un estudio en cateterismo de 50 pacientes coronarios, se han obtenido los siguientes valores:

$$\overline{v} = 1,2 \text{ seg}^{-1} \quad y \quad s_v^2 = 0,25 \text{ seg}^{-2}$$

$$\overline{p} = 10 \text{ mmHg} \quad y \quad s_p^2 = 4 \text{ mmHg}^2$$

$$s_{vp} = 0,8 \text{ seg}^{-1}\text{mmHg}$$

Donde, v es la velocidad de acortamiento circunferencial a nivel del ecuador del ventrículo izquierdo y p la presión diastólica ventricular.

Analizar la correlación entre estas variables. Si fuera significativa, obtener las ecuaciones de regresión correspondientes.

4. Las rectas de regresión que dan la relación entre el perímetro torácico y el peso de un grupo de 200 individuos, vienen dadas por las siguientes ecuaciones:

$$\begin{aligned} y &= 0,52x + 21,71 \\ x &= 0,75y + 40,97 \end{aligned}$$

Donde y es el peso medido en kilos y x el perímetro torácico medido en centímetros.

Obtener el valor del coeficiente de correlación entre estas dos variables.

5. La concentración en sangre de un fármaco y la sobrepresión media arterial que el mismo origina están relacionadas por la expresión

$$p + 5 = 0,6\,c$$

6. Se han determinado la longitud y el peso de una glándula obtenida en autopsias de 10 animales. Los resultados son los siguientes (expresados en cm y gr, respectivamente):

(6 - 7), (3 - 4), (2 - 3), (1 - 1), (3 - 4), (4 - 6), (3 - 5), (2 - 2), (1 - 2) y (5 - 6)

Estimar la longitud de una glándula cuyo peso sea 4,5 gr.