

2 Los valores de dos variables X e Y se distribuyen según la tabla adjunta. Determinar el coeficiente de correlación y la recta de regresión de Y sobre X. Comentar cómo de fiables son las predicciones basadas en esa recta.

| | | | | |
|---|---|---|---|---|
| | Y | 0 | 2 | 4 |
| X | | | | |
| 1 | | 2 | 1 | 3 |
| 2 | | 1 | 4 | 2 |
| 3 | | 2 | 5 | 0 |

Completemos la tabla:

| X \ Y | 0 | 2 | 4 | $n_{i\cdot}$ | $n_{i\cdot}x_i$ | x_i^2 | $n_{i\cdot}x_i^2$ | $\sum_j n_{ij}x_iy_j$ |
|-----------------------|------------------|-------------------|-------------------|--------------|-----------------|---------|-------------------|-----------------------|
| 1 | 2 ₍₀₎ | 1 ₍₂₎ | 3 ₍₁₂₎ | 6 | 6 | 1 | 6 | 14 |
| 2 | 1 ₍₀₎ | 4 ₍₁₆₎ | 2 ₍₁₆₎ | 7 | 14 | 4 | 28 | 32 |
| 3 | 2 ₍₀₎ | 5 ₍₃₀₎ | 0 ₍₀₎ | 7 | 21 | 9 | 63 | 30 |
| $n_{\cdot j}$ | 5 | 10 | 5 | N=20 | 41 | | 97 | 76 |
| $n_{\cdot j}y_j$ | 0 | 20 | 20 | 40 | | | | |
| y_j^2 | 0 | 4 | 16 | | | | | |
| $n_{\cdot j}y_j^2$ | 0 | 40 | 80 | 120 | | | | |
| $\sum_i n_{ij}x_iy_j$ | 0 | 48 | 28 | 76 | | | | |

$\sum_i \sum_j n_{ij}x_iy_j$

Las medias, varianzas y desviaciones típicas marginales son

$$\bar{X} = \frac{\sum_{i=1}^3 n_{i\cdot}x_i}{N} = \frac{41}{20} = 2'05$$

$$\sigma_X^2 = \overline{(X^2)} - (\bar{X})^2 = \frac{\sum n_{i\cdot}x_i^2}{N} - (\bar{X})^2 = \frac{97}{20} - (2'05)^2 = 0'6475$$

$$\sigma_X = +\sqrt{\text{Var}(X)} = +\sqrt{0'6475} = 0'8047$$

$$\bar{Y} = \frac{\sum_{j=1}^3 n_{\cdot j}y_j}{N} = \frac{40}{20} = 2$$

$$\sigma_Y^2 = \overline{(Y^2)} - (\bar{Y})^2 = \frac{\sum n_{\cdot j}y_j^2}{N} - (\bar{Y})^2 = \frac{120}{20} - (2)^2 = 2$$

$$\sigma_Y = +\sqrt{\text{Var}(Y)} = +\sqrt{2} = 1'4142$$

Covarianza

$$\sigma_{XY} = \overline{(X \cdot Y)} - \bar{X} \cdot \bar{Y} = \frac{\sum_{i=1}^3 \sum_{j=1}^3 n_{ij} x_i y_j}{N} - \bar{X} \cdot \bar{Y} = \frac{76}{20} - (2'05) \cdot (2) = -0'3 < 0$$

Puesto que $\sigma_{XY} = Cov(X,Y) < 0$, las variables X e Y están correlacionadas negativamente (hay correlación inversa): a un aumento de X corresponde una disminución de Y.

Coefficiente de correlación (de PEARSON)

$$\rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{-0'3}{(0'8047) \cdot (0'4142)} = -0'26$$

Puesto que ρ es próximo a 0, la correlación (inversa) es muy débil. Las variables X e Y son casi incorreladas.

Recta de regresión (mínimo-cuadrática) de Y sobre X

$$r_{Y/X}: \boxed{y - \bar{Y} = \frac{\sigma_{XY}}{\sigma_X^2} (x - \bar{X})}$$

$$r_{Y/X}: y - 2 = \frac{-0'3}{0'6475} (x - 2'05)$$

$$r_{Y/X}: y = (-0'4633)x + (2'9498)$$

Se emplea para predecir el valor de Y para un valor dado de X. Para un individuo tal que X tome el valor x_0 predice un valor de Y de

$$x_0 \mapsto \hat{y}_0 = (-0'4633)x_0 + (2'9498)$$

Puesto que el coeficiente de correlación es próximo a 0, la predicción es poco fiable.

