Regresión y correlación lineal I

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA

Introducción

- Necesidad del estudio simultaneo de varias variables.

  - Analizar las posibles relaciones entre ellas. (Estudios de correlación).

  - Intentar establecer el modelo matemático que las relacione. (Estudios de regresión o ajuste).

- Interés por estimar una magnitud Y (variable dependiente) en función de una o varias $X_1, X_2, \cdots, X_k$ variables explicativas (variables independientes).

  - Imposibilidad de predicción exacta. Perturbación aleatoria ε.

  - Modelos lineales:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Modelos comunes

Lineales

Lineal simple: $\hat{Y} = \beta_0 + \beta_1 X$

Parabólico: $\hat{Y} = \beta_0 + \beta_2 X + \beta_1 X^2$

Cúbico: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

Polinómico: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^2 + \cdots + \beta_h X^h$

No Lineales

Exponencial: $\hat{Y} = \beta_0 k^{\beta_1 X}$

Potencial: $\hat{Y} = \beta_0 X^{\beta_1}$

Hiperbólico: $\hat{Y} = 1 \big/ \beta_0 + \beta_1 X$

Logístico: $\hat{Y} = 1 \big/ e^{-\beta_0 - \beta_1 X}$

Estimación de los coeficientes: $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$

Problema de inferencia:

Contrastes sobre los coeficientes y sobre el ajuste del modelo.

Correlación simple

$$R_{xy} = \frac{S_{xy}}{S_x S_y} \in \left[-1;1\right]$$

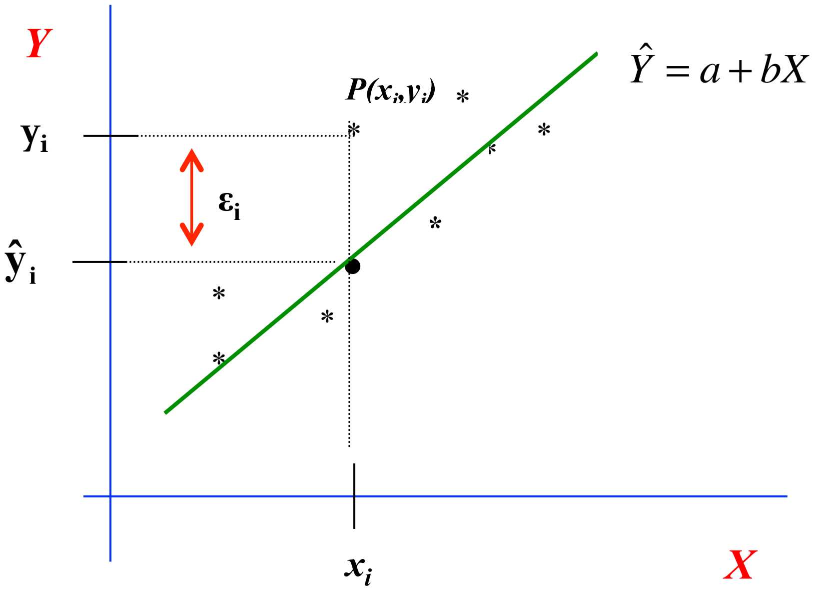
$$R_{x_j x_i} = \frac{S_{x_j x_i}}{S_{x_j} S_{x_i}} \ \forall\, i \neq j = 1,2,\ldots$$

Matriz de Correlación: $$\Gamma_{yx_1x_2\ldots} = \begin{pmatrix} 1 & R_{yx_1} & R_{yx_2} & \cdots \\ R_{yx_1} & 1 & R_{x_1x_2} & \cdots \\ R_{yx_2} & R_{x_1x_2} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA

Regresión Simple: Línea de Regresión

Recta de regresión de Y sobre X



$\hat{y}_i$ valor estimado de *Y* para *X*=x_i

ε_i residuos: $\varepsilon_i = y_i - \hat{y}_i$

Mínimos cuadrados $$H = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

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

Derivamos parcialmente e igualamos a cero:

Ecuaciones Normales

Resolviendo el sistema, se obtienen *a* y *b*.

Resultados intermedios:

- $\bar{\varepsilon} = 0$

$$\sum_{i=1}^{n}\left(y_i - a - bx_i\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n}\varepsilon_i = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = 0 \Rightarrow \bar{\varepsilon} = 0$$

- $\bar{y} = a + b\bar{x} \quad \Rightarrow \quad (\bar{x},\bar{y})$ satisface la ecuación de regresión ≅ la recta de regresión pasa por el centro de gravedad.

$$\sum_{i=1}^{n}\left(y_i - a - bx_i\right) = 0 \Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n} na - \frac{1}{n} b\sum_{i=1}^{n} x_i = 0 \Rightarrow$$

$$\Rightarrow \bar{y} - a - b\bar{x} = 0$$ Con lo que $$\boxed{a = \bar{y} - b\bar{x}}$$

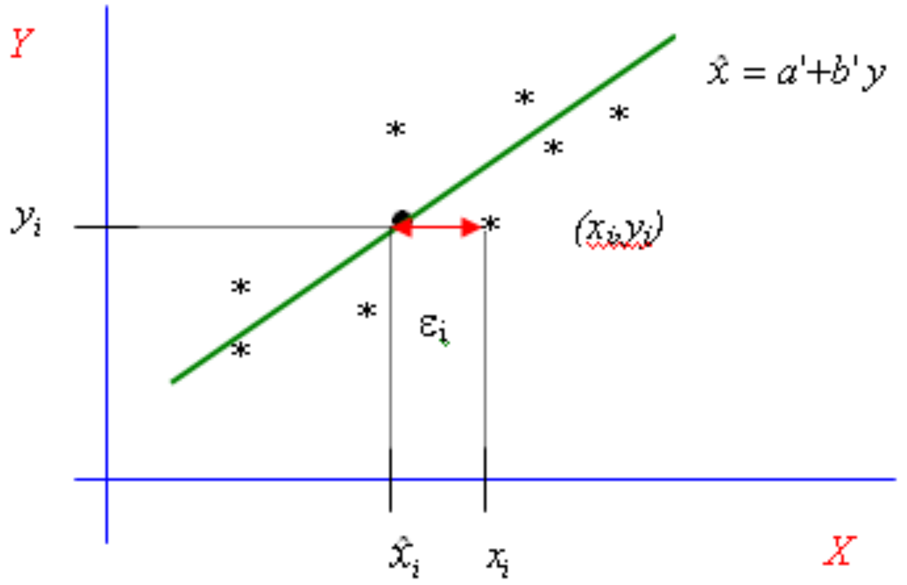
- $\sum_{i=1}^{n}\varepsilon_i x_i = 0 \quad \Rightarrow \quad COV(\varepsilon, x) = S_{\varepsilon x} = 0$ Los residuos y la variable independiente están incorrelados.

$$\sum_{i=1}^{n}\left(y_i - a - bx_i\right)x_i = 0 \Rightarrow \sum_{i=1}^{n}\varepsilon_i x_i = 0 \qquad \Rightarrow S_{\varepsilon X} = \sum_{i=1}^{n}\varepsilon_i x_i - \bar{\varepsilon}\bar{x} = 0 - 0 = 0$$

- $COV\left(\varepsilon, X\right) = S_{xy} - bS_x^2 = 0 \Rightarrow \quad \boxed{b = \frac{S_{xy}}{S_x^2}}$

Recta de regresión de X sobre Y: $\left(\hat{x} = a' + b' y\right)$

Estudio análogo.

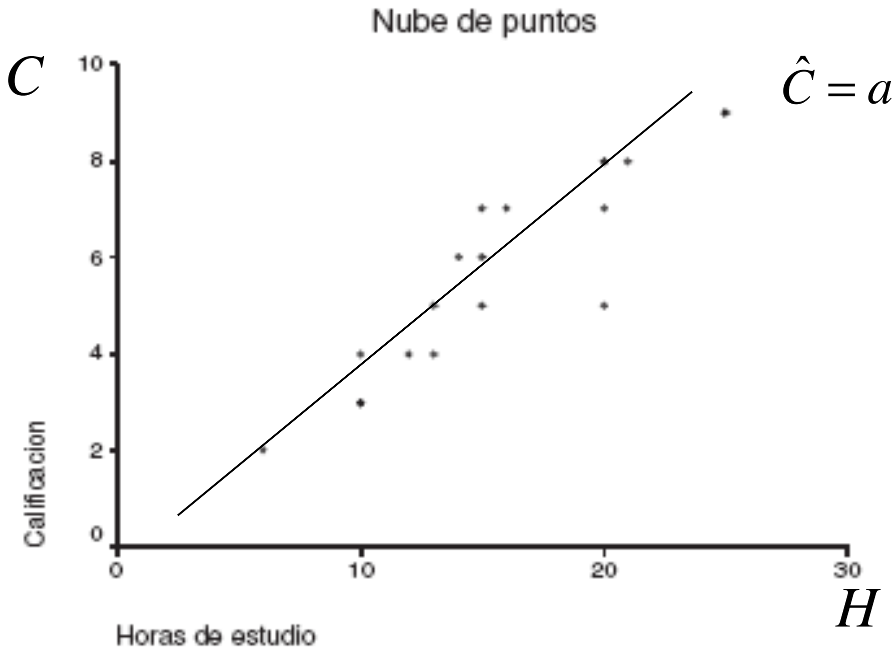


En este caso: $a' = \overline{x} - b' \overline{y}$ $\qquad b' = \dfrac{S_{xy}}{S_y^2}$

Relación entre las pendientes: $b\,b' = \dfrac{S_{xy}^2}{S_x^2 S_y^2} = R_{xy}^2$

C	$: Calificación$	5	7	3	8	4	6	9	8	3	5
H	$: Horas$	15	20	10	21	12	15	25	20	10	13

C	$: Calificación$	7	4	3	2	6	7	8	5	9	4
H	$: Horas$	15	13	10	6	14	16	20	20	25	10



$$\hat{C} = a + bH$$

$$\overline{C} = 5.65 \ ; \ S_C^2 = 4.4275$$

$$\overline{H} = 15.5 \ ; \ S_H^2 = 26.55$$

$$S_{CH} = 9.975 \qquad R_{CH} = \frac{S_{CH}}{S_C S_H} = 0.92$$

$$\begin{cases} b = \dfrac{9.975}{26.55} = 0.3757 \\ a = 5.56 - (0.3757)(15.5) = -0.2633 \end{cases}$$

El modelo obtenido es: $\hat{C} = -0.2633 + 0.3757\,H$

$$\hat{C}_{(H=23)} = -0.2633 + (0.3757)(23) = 8.38$$

$$R^2 = R_{CH}^2 = \frac{S_{CH}^2}{S_C^2 S_H^2} = 0.8464$$

Regresión y correlación lineal I

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA