Regresión y correlación lineal II

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA

Regresión parabólica.

$$\hat{y} = a + b_1 x + b_2 x^2$$

Usando el método de mínimos cuadrados, minimizamos la expresión:

$$H = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - a - b_1 x_i - b_2 x_i^2 \right)^2$$

para ello derivamos parcialmente e igualamos a cero:

$$\left.\begin{aligned}
\frac{\partial H}{\partial a} &= -2 \sum_{i=1}^{n} \left(y_i - a - b_1 x_i - b_2 x_i^2 \right) = 0 \\
\frac{\partial H}{\partial b_1} &= -2 \sum_{i=1}^{n} \left(y_i - a - b_1 x_i - b_2 x_i^2 \right) x_i = 0 \\
\frac{\partial H}{\partial b_2} &= -2 \sum_{i=1}^{n} \left(y_i - a - b_1 x_i - b_2 x_i^2 \right) x_i^2 = 0
\end{aligned}\right\} \quad Ecuaciones \quad Normales$$

Resolviendo el sistema, se obtienen los parámetros: *a*, b_1 y b_2.

Podemos simplificar realizando el cambio $Z=X^2$, con lo que:

$$\hat{y} = a + b_1 x + b_2 z$$

Y de nuevo aplicando mínimos cuadrados:

$$\left.\begin{aligned}
\frac{\partial H}{\partial a} &= -2\sum_{i=1}^{n}\left(y_i - a - b_1 x_i - b_2 z_i\right) = 0 \\
\frac{\partial H}{\partial b_1} &= -2\sum_{i=1}^{n}\left(y_i - a - b_1 x_i - b_2 z_i\right) x_i = 0 \\
\frac{\partial H}{\partial b_2} &= -2\sum_{i=1}^{n}\left(y_i - a - b_1 x_i - b_2 z_i\right) z_i = 0
\end{aligned}\right\} \quad \textit{Ecuaciones} \quad \textit{Normales}$$

resolviendo el sistema, se obtienen los valores de los parámetros.

Conclusiones:

→ $\bar{\varepsilon} = 0$

→ $a = \bar{y} - b_1\bar{x} - b_2\bar{z} \qquad \left(\bar{y} = a + b_1\bar{x} + b_2\bar{z}\right)$

$\longrightarrow$ $COV(\varepsilon X) = 0$ $\longrightarrow$ $COV(\varepsilon Z) = 0$

$\longrightarrow$ $b_i = -\dfrac{A_{1,i+1}}{A_{1,1}}$ para $i = 1,2$. Donde $A_{1,1}$ es el adjunto al elemento $a_{1,1}$ de la matriz de covarianzas Σ.

$$\Sigma = \begin{pmatrix} S_y^2 & S_{yx} & S_{yz} \\ S_{xy} & S_x^2 & S_{xz} \\ S_{zy} & S_{zx} & S_z^2 \end{pmatrix}$$

y $A_{1,i+1}$ el adjunto al elemento $a_{1,i+1}$

Al menor complementario del elemento $a_{1,1}$ de Σ, es decir:

$$\Sigma_{\vec{x}} = \begin{pmatrix} S_x^2 & S_{xz} \\ S_{zx} & S_z^2 \end{pmatrix}$$

se le denomina matriz de covarianzas de las $\vec{x}$ (v. indep.)

Este proceso se puede generalizar al caso **lineal múltiple**, dado el hiperplano de regresión:

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Regresión exponencial y potencial.

Exponencial: $\hat{y} = ak^{bx} \ ,k > 0$

Si tomamos logaritmos: $\log_k \hat{y} = \log_k a + bx \log_k k \quad \Rightarrow \quad \hat{y}^* = a^* + b^* x$

donde:

$$\left.\begin{aligned} \hat{y}^* &= \log_k \hat{y} \\ a^* &= \log_k a \\ b^* = b \log_k k &= b \end{aligned}\right\}$$

con lo que se ha transformado el problema en uno de tipo lineal. Para obtener finalmente *a*, basta tomar antilogaritmos.

Potencial: $\hat{y} = ax^b$

tomando logaritmos de nuevo: $\log \hat{y} = \log a + b \log x \quad \Rightarrow \quad \hat{y}^* = a^* + bx^*$

donde:

$$\left.\begin{aligned} \hat{y}^* &= \log \hat{y} \\ a^* &= \log a \\ x^* &= \log x \end{aligned}\right\}$$

de nuevo se ha transformado el problema en uno de tipo lineal.

Correlación múltiple: $$R_{yx_1x_2\ldots} = R_y = \frac{S_{y\hat{y}}}{S_y S_{\hat{y}}} \in [-1,1]$$

Se verifica que: $$R_y^2 \geq R_{yx_i}^2 \;\forall\, i = 1,2,\ldots$$

Varianza residual.

$$V(\varepsilon) = S_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

Por otra parte:

$$S_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \ldots = S_y^2\left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2}\right) = S_y^2\left(1 - R_{xy}^2\right)$$

Coeficiente de determinación: $$R^2 = R_{xy}^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2 = 1 - \frac{S_\varepsilon^2}{S_y^2} \in [\,0\,,1\,]$$

Siendo: $$S_\varepsilon^2 = \frac{\det \Sigma}{\det \Sigma_{\vec{x}}}$$

"Tanto por 1 (ó %) de la varianza de *Y* explicada por el modelo."

Nos informa de hasta qué punto el modelo se ajusta a los datos.

Correlación parcial.

Objetivo: Eliminar los efectos distorcionantes de terceras variables sobre las relaciones lineales entre variables.

Fundamento: Dados los modelos:

$$x = a + bz + \varepsilon_x \quad ; \quad y = a' + b'z + \varepsilon_y$$

ε_x representa la parte de *X* que no es capaz de explicar *Z* (luego *Z* no interviene), y de forma análoga ε_y .

Por tanto se define el **coeficiente de correlación entre *X* e *Y*, parcial *Z*** como:

$$\rho_{xy,z} = \rho_{\varepsilon_x \varepsilon_y}$$

Operativamente se puede obtener como:

$$\rho_{xy,z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}} = -\frac{A_{12}}{\sqrt{A_{11}A_{22}}}$$

Conclusiones:

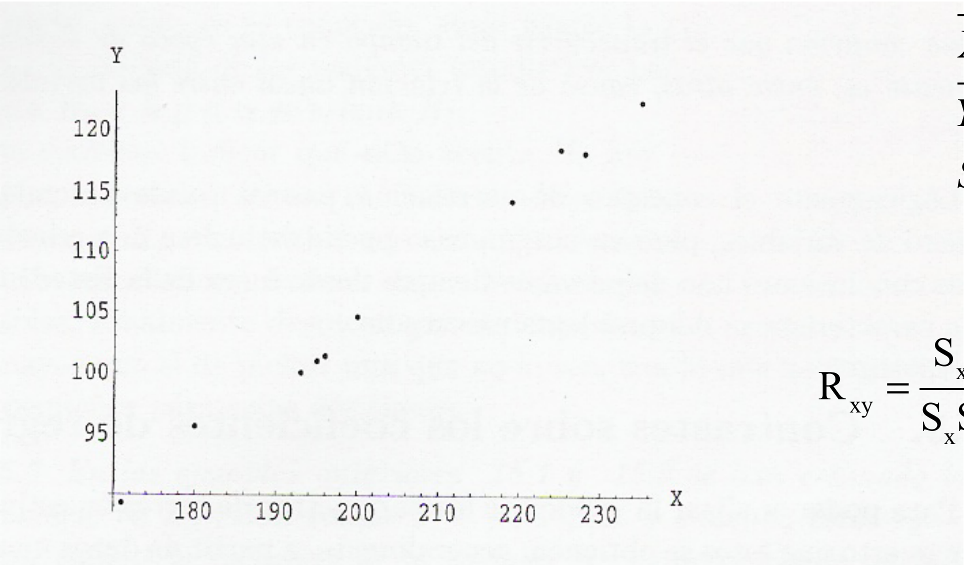
$$\rho^2_{xy,z} > \rho^2_{xy} \quad \Rightarrow \quad \begin{cases} \text{La variable } Z \text{ oculta o amortigua la} \\ \text{dependencia entre } X \text{ e } Y \end{cases}$$

$$\rho^2_{xy,z} \cong 0 \quad \Rightarrow \quad \begin{cases} \text{La interdependencia entre } X \text{ e } Y \text{ se debe} \\ \text{casi exclusivamente al efecto de } Z \end{cases}$$

$$\rho^2_{xy,z} < \rho^2_{xy} \quad \Rightarrow \quad \begin{cases} \text{La interdependencia entre } X \text{ e } Y \text{ se debe} \\ \text{parcialmente a la influencia de } Z \end{cases}$$

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA

X	171	180	193	195	196	200	219	225	228	235
Y	89.3	95.6	100	101	101.4	104.6	114.3	118.5	118.2	122.4



$$\overline{X} = 204.2 \; ; \; \overline{S}_x = 21.47246$$

$$\overline{Y} = 106.53 \; ; \; \overline{S}_y = 11.11206$$

$$S_{xy} = 237.91556$$

$$R_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{237.91556}{(21.47246)(11.11206)} = \underline{0.9971}$$

X	171	180	193	195	196	200	219	225	228	235
Y	89.3	95.6	100	101	101.4	104.6	114.3	118.5	118.2	122.4
Z	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	78	79	80	81	82	83	84	85	86	87

X ≡ “Nº de coches aparcados en una determinada facultad”
Y ≡ “Producción industrial de Japón”
Z ≡ “Periodo de años 1978-1987”

El vector de medias y la matriz de covarianzas son:

$$\vec{\mu}_{xyz} = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix} = \begin{pmatrix} 204.20 \\ 106.21 \\ 5.50 \end{pmatrix}$$

$$\sum_{XYZ} = \begin{pmatrix} S_{xx} = S_x^2 & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} = S_y^2 & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} = S_z^2 \end{pmatrix} = \begin{pmatrix} 461.0666 & 237.9155 & 63.7777 \\ 237.9155 & 123.4778 & 32.8722 \\ 63.7777 & 32.8722 & 9.16666 \end{pmatrix}$$

Coeficiente de correlación parcial:

$$R_{xy,z} = -\frac{A_{12}}{\sqrt{A_{11} \times A_{22}}}$$

$$A_{12} = -\begin{vmatrix} 237.9155 & 32.8722 \\ 63.7777 & 9.16666 \end{vmatrix} = -84.3772$$

$$A_{22} = \begin{vmatrix} 461.0666 & 63.7777 \\ 63.7777 & 9.16666 \end{vmatrix} = 158.8457$$

$$A_{11} = \begin{vmatrix} 123.4778 & 32.8722 \\ 32.8722 & 9.16666 \end{vmatrix} = 51.29747$$

$$R_{xy,z} = -\frac{A_{12}}{\sqrt{A_{11} \times A_{22}}} = -\frac{-84.3772}{\sqrt{51.29747 \times 158.8457}} = \underline{0.93473|}$$

$$\left.\begin{array}{l} R^2_{xy} = 0.9971^2 = 0.9942 \\ R^2_{xy,z} = 0.93473^2 = 0.87372 \end{array}\right\} ; R^2_{xy,z} < R^2_{xy} \Rightarrow$$

La interdependencia entre X e Y se debe parcialmente a la influencia de Z.

Aún eliminando los efectos de Z, la relación entre X e Y sigue siendo muy alta desde un punto de vista estadístico, pero esto no significa que tenga sentido fuera de este ámbito, como es obvio para el ejemplo que nos ocupa.

Continuaremos con el ejemplo determinando el modelo: $\hat{x} = a + b_1 y + b_2 z$

(Recordemos que la matriz de covarianzas $\sum_{XYZ}$ hace referencia a este modelo.)

$$\sum_{XYZ} = \begin{pmatrix} 461.0666 & 237.9155 & 63.7777 \\ 237.9155 & 123.4778 & 32.8722 \\ 63.7777 & 32.8722 & 9.16666 \end{pmatrix} \quad \Rightarrow \quad \det \sum_{XYZ} = 112.2411$$

$$\sum_{\vec{X}} = \begin{pmatrix} 237.9155 & 123.4778 \\ 63.7777 & 32.8722 \end{pmatrix} \quad \Rightarrow \quad \det \sum_{\vec{X}} = 51.2974$$

$$b_i = -\frac{A_{1,i+1}}{A_{1,1}} \quad \text{para } i = 1,2$$

$$A_{13} = -\begin{vmatrix} 237.9155 & 123.4778 \\ 63.7777 & 32.8722 \end{vmatrix} = -54.3241$$

$$\Rightarrow \begin{cases} b_1 = -\dfrac{A_{12}}{A_{11}} = -\dfrac{-84.3772}{51.29747} = 1.6448 \\[2ex] b_2 = -\dfrac{A_{13}}{A_{11}} = -\dfrac{-54.3241}{51.29747} = 1.059 \end{cases}$$

$$a = \overline{y} - b_1\overline{x} - b_2\overline{z} = 23.3145 \quad \Rightarrow \quad \hat{x} = 23.3145 + 1.6448\, y + 1.059\, z$$

Para y=130 y Z=11 $\quad \hat{x} = 23.3145 + 1.6448\,(130) + 1.059\,(11) = 248.7875 \equiv 248$

$$S_\varepsilon^2 = \frac{\det \sum}{\det \sum_{\vec{x}}} = 2.188 \qquad R^2 = R_{xy}^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2 = 1 - \frac{S_\varepsilon^2}{S_y^2} = 0.9822$$

Regresión y correlación lineal II

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba
DEPARTAMENTO DE ESTADÍSTICA