Intro

Wednesday, January 26, 2022 6:56 PM

* Reduc° de datos o simplifica° estructural
* Ordenar y agrupar datos
* Investiga° dependencia entre vars.
* Técnicas de predic° → regresiones: lineal, multivar...
* Construc° P. H (pruebas hip.)

↳ Nota° matricial

	Var 1	Var p
Item 1	x_{11}	x_{1p}
Item n	x_{n1}	x_{np}

$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{bmatrix}$ (obs. × var.)

→ Medidas q' resumen los datos: media, mediana, var.

- Media muestral

$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1, \dots, p$

- Varianza muestral $\quad s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$

$s_k^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \quad k = 1, \dots, p$

$S_{kk} = s_k^2$

$S = \begin{bmatrix} s_{11} & \cdots \\ s_{12} & \cdots \\ \vdots & s_{pp} \end{bmatrix}$ ↳ simétrica.

varianza
covar(z, z)

- Desv. estándar : $\sqrt{s_{kk}} = s_k$

Ej: n obs. de 2 vars. $\begin{bmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{bmatrix}$

- Covarianza : $s_{12} = \frac{1}{n}\sum (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$
  - ↳ Asocia° lín. entre las vars.
  - ↳ Promedio del producto entre las desviaciones de sus respectivas medias.
  - ↳ Si hay tanto grandes como pequeños vals en ambas vars. se presentan de forma conjunta, s_{12} es positiva.
  - ↳ Si hay vals grandes en una y pequeños en otra, s_{12} es neg.
  - ↳ Si no hay asocia° entre ambas, s_{12} es aprox 0.

- Coef. correla° muestral: Normaliza la cov para no depender de las unidades.

$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum (x_{ji} - \bar{x}_i)^2}\sqrt{\sum (x_{jk} - \bar{x}_k)^2}} \rightarrow [-1, 1]$

↳ $r_{ii} = 1$

↳ Asocia° lineal entre 2 vars., q' no depende de unidades.

- $r = 0$: No hay asocia°
- $r = -1$: Correla° neg → una aumenta y la otra disminuye (en promedio).

$p = (x_1, \dots, x_p)$

- Distancia Euclideana $\quad d(0, P) = \sqrt{x_1^2 + \dots + x_p^2}$

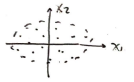
Todos los P a la misma dist. cuadrada c^2 de 0 están sobre una hiperesfera

$d(P, Q) = \sqrt{(x_1 - y_1)^2 + \dots + (x_p - y_p)^2}$

↳ Podemos tener problemas por la dif. de unidades entre componentes, las cuales contribuyen igualm. al cálculo.

Cuando las coord. representan medidas q' pueden presentar variaciones aleatorias de dif. magnitudes es conveniente ponderar con mayor peso a las vars. con menos varia°.

↳ Desarrollar una distancia estadística q' considere las difs. en varia° y la presencia de correlac°.



→ Hay n pares de medidas en x_1 y x_2.
→ Tienden a 0 y varianzas incl. ps.
→ Los P en x_1 tienen + variabilidad q' los de x_2 → x_2 tiene más ponderº.

→ $x_1^* = x_1/\sqrt{s_{11}}$
→ $x_2^* = x_2/\sqrt{s_{22}}$

$d(0, P) = \sqrt{(x_1^*)^2 + (x_2^*)^2}$

- Si la variabilidad en ambas coordenadas es igual, usar la distancia euclideana.
- Todos los P (x_1, x_2) con dist cuadrada c^2 satisfacen q' $c^2 = \frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}$ están en una elipse centrada en 0.

- Distancia estadística entre 2 puntos. → Asume indep en las vars.

$d(P, 0) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \dots + \frac{(x_p - y_p)^2}{s_{pp}}}$

$d(0, P) = \sqrt{\frac{\hat{x}_1^2}{s_{11}} + \frac{\hat{x}_2^2}{s_{22}}}$

No requiere la indep.

- $\hat{x}_1 = x_1 \cos\theta + x_2 \operatorname{sen}\theta$
- $\hat{x}_2 = -x_1 \operatorname{sen}\theta + x_2 \cos\theta$

$P = (x_1, x_2)$; $Q = (y_1, y_2)$

$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$

$d = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$

a_{ij} son fun°es de θ.

Ahora, con p vars.

$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + \dots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + \dots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)}$