Tema 2: Estadística descriptiva: Dos variables.

Representación de tablas bidimensionales.

C\E	0	1	$n_{i\bullet}$
0	2	4	6
1	4	8	12
2	2	0	2
$n_{\bullet j}$	8	12	20 = N

Distribución de C

Distribución de E

n_{11} : 2 veces aparece el dato $\overset{(E_1, C_1)}{(x_1, y_1)}$ en mi muestra

Distribuciones marginales. Distribución de E y distribución de C.

$f_{ij} = \frac{n_{ij}}{N}$

N: tamaño de la muestra

$N = \sum_{i=1}^{2} \sum_{j=1}^{3} n_{ij}$

Tablas de frecuencia de C y E por separadas

C	$n_{i\bullet}$
0	6
1	12
2	2

$\Sigma = 20$

E	$n_{\bullet j}$
0	8
1	12

$\Sigma = 20$

INDEPENDENCIA: Relación entre variables

INDEPENDENCIA

x\y	C_1	C_2	C_3	C_4	
A	4	6	10	2	22
B	2	3	5	1	11
	6	9	15	3	33

X|Y = C_1	ni	f_i
A	4	4/6 = 2/3
B	2	2/6 = 1/3
	6	

X|Y = C_2	ni	f_i
A	6	6/9 = 2/3
B	3	3/9 = 1/3
	9	

X|Y = C_3	ni	f_i
A	10	10/15 = 2/3
B	5	5/15 = 1/3
	15	

X|Y = C_4	n_i	f_i
A	2	2/3
B	1	1/3
	3	

Def 1 → X es independiente de Y si $f_i^{\,j} =$

Def 2 → Y es independiente de X si $f_i^{\,j} = f_{\bullet j}$ $\forall ij$

Def 3 → X es independiente de Y (y viceversa) si $f_{ij} = f_{i\bullet} \times f_{\bullet j}$

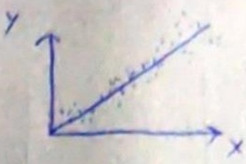
Propiedad: X es independiente de Y ⟺ Y es independiente de X.

DEPENDENCIA FUNCIONAL

X = "número de días en un año"
Y = "número de horas en un año"

$\Big\}\; Y = 24X$ (Una en función de la otra) (Error = 0)

DEPENDENCIA ESTADÍSTICA



x = "número de monitores que hay en el aula de informática"

y = "número de computadores que hay en el aula de informática"

$$y = x + \varepsilon \; (\varepsilon = \text{error})$$

DEPENDENCIA FUNCIONAL

x = "nº de días en un año" // y = nº de horas en un año" $\quad y = 24x$ (Error = 0)

REGRESIÓN Y CORRELACIÓN

TIPOS DE AJUSTES

Método general para el ajuste de mínimos cuadrados

Supongamos que tenemos $\{(x_i, y_i)\}_{i=1}^{N}$ para dos variables cuantitativas x e y

Tenemos una tabla simple de datos ¿Cómo puedo obtener una tabla simple de datos a partir de una tabla bidimensional?

x\y	1	2
3	0	1
4	9	4

tabla bidimensional $\{(x_i, y_j)\}_{j=1,2}^{i=1,2}$ $\rightsquigarrow$

x	y	n_i
3	1	0
3	2	1
4	1	9
4	2	8

tabla simple $\{(x_i, y_i)\}_{i=1}^{4}$

Lo que buscamos es una función φ que mejor "explique" los datos: Es decir, dado un dato (x_i, y_i) voy a tratar de aproximar $y_i \approx y_i^*$, donde: $y_i^* = \varphi(x_i)$ {Introduciendo notación}

- $y_i^* = \varphi(x_i)$ es el valor de "y" estimado por la regresión de x
- $e_i = y_i - y_i^*$ es el error cometido por el ajuste por el dato i-ésimo $((x_i, y_i))$

¿Cuál será la idea para encontrar φ?

→ Minimizar e_i para todo $i = 1, 2, \dots, N$

Definimos la variable $E = y_i - y_i^*$ $\quad (e_i = y_i - y_i^*)$

Lo que haré es minimizar el "error cuadrático medio" de E

$$ECM_E = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i^*)^2$$

→ Como minimizar el error anterior: ORGANIZO LOS DATOS: $\{(x_i, y_i)\}_{i=1}^{N}$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad ; \quad y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_N^* \end{pmatrix} = \begin{pmatrix} \varphi(x_1) \\ \varphi(x_2) \\ \vdots \\ \varphi(x_n) \end{pmatrix} = \varphi(\overline{x})$$

Vamos a suponer que φ es de la siguiente forma:

$$\varphi(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \dots + a_s \varphi_s(x)$$

•) Ajuste lineal $y = a_0 + a_1 x$

Se corresponde a $s=1$

$$\begin{cases} \varphi_0(x) = 1 \\ \varphi_1(x) = x \end{cases}$$

•) Ajuste parabólico $y = a_0 + a_1 x + a_2 x^2$

$s=2$

$$\begin{cases} \varphi_0(x) = 1 \\ \varphi_1(x) = x \\ \varphi_2(x) = x^2 \end{cases}$$

•) Ajuste trigonométrico $y = a_0 \cdot \cos(x) + a_1 \cdot \operatorname{sen}(x)$

$s=1$

$$\begin{cases} \varphi_0(x) = \cos(x) \\ \varphi_1(x) = \operatorname{sen}(x) \end{cases}$$

→ Este problema de regresión estará resuelto cuando encuentre $a_0, a_1, \dots, a_s$ que minimizan el ECM anterior definido

$$y^* = \begin{pmatrix} \varphi_0 a_0(x_1) + \varphi_1 a_1(x_1) + \varphi_2 a_2(x_1) \dots + \varphi_s a_s(x_1) \\ \varphi_0 a_0(x_2) + \varphi_1 a_1(x_2) + \dots + \varphi_s a_s(x_2) \\ \vdots \qquad \vdots \qquad \vdots \\ \varphi_0 a_0(x_n) + \varphi_1 a_1(x_n) + \dots + \varphi_s a_s(x_n) \end{pmatrix}$$

$$y^* = \underbrace{\begin{pmatrix} \varphi_0 a_0(x_1) + \varphi_1 a_1(x_1) + \dots + \varphi_s a_s(x_1) \\ \varphi_0 a_0(x_2) + \varphi_1 a_1(x_2) + \dots + \varphi_s a_s(x_2) \\ \vdots \\ \varphi_0 a_0(x_n) + \varphi_1 a_1(x_n) + \dots + \varphi_s a_s(x_n) \end{pmatrix}}_{M} \cdot \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_s \end{pmatrix}}_{A} = M \cdot A$$

Vamos a minimizar el ECM

$$\min_A \left\{ \sum_{i=1}^{n} e_i^2 \right\} = \min_A \left\{ \sum_{i=1}^{n} (y_i - y_i^*)^2 \right\} = \min_A \left\{ (y-y^*)^t \cdot (y-y^*) \right\} = \min_A \left\{ y^t y - y^t MA - (MA)^t y + (MA)^t (MA) \right\}$$

El mínimo se alcanza:

Sea $f(A) = y^t y - y^t MA - (MA)^t y + (MA)^t (MA)$

$\frac{\partial f}{\partial A} = 0 \rightarrow$ Eso ocurre cuando: $\boxed{(M^t M) A = M^t y}$ Ecuaciones normales

CASO PARTICULAR

• Ajuste lineal → $y = a_0 + a_1 x$ la llamamos recta de regresión $Y|X$

Encontrar A resolviendo el sistema lineal $(M^t M) A = M^t Y$ $\left\{ \begin{matrix} \varphi_0(x) = 1 \\ \varphi_1(x) = x \end{matrix} \right.$

$$M = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad M^t = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \quad M^t M = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} N & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix}$$

Sistema lineal

$$M^t Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix} \qquad \begin{pmatrix} N & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix}$$

Ejemplo diapositiva 40

$(M^t M) A = M^t Y$

$$M = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_s(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_s(x_2) \\ & & & \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_s(x_n) \end{pmatrix}$$

X	2002	2003	2004	2005	...	2011
Y	913215	926708	963513	981403		1239183

Recta de regresión Y/X

1) Tabla de entrada

2) $N = 10$

3) Ecuaciones normales $(M^t M) A = M^t Y$

4) Construyo M

$$M = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{10} \end{pmatrix}, \quad A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

5) Calculo $M^t M$

$$M^t M = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_{10} \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{10} \end{pmatrix} = \begin{pmatrix} 10 & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix} = \begin{pmatrix} 10 & 20065 \\ 20065 & 40260505 \end{pmatrix}$$

6) Calculo $M^t Y$

$$M^t Y = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_{10} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_{10} \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma y_i x_i \end{pmatrix} = \begin{pmatrix} 10736140 \\ 2'154529648 \cdot 10^{9} \end{pmatrix}$$

7) Resolvemos el sistema lineal

$$\begin{pmatrix} 10 & 20065 \\ 20065 & 40260505 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 10736140 \\ 2'154529648 \cdot 10^{9} \end{pmatrix}$$

8) Obtenemos la recta

$$Y(x) = a_0 + a_1 x$$

Este modelo matemático explica la variable Y en función de la variable X

→ ¿Podría encontrar otro "modelo matemático" lineal que explique x en función de y?

Sí, a eso la llamamos recta de regresión de $x|y$

$$\boxed{x(y) = a_0' + a_1' y}$$

El método general ya visto, puede aplicarse también

•) Modelo $\boxed{x = a_0' + a_1' y}$

•) $(M^t M) A = M^t X$ → Pongo lo que quiero explicar

•) Construyo M. Como en este caso el modelo es $x = a_0 + a_1 y$, $\left.\begin{matrix} \phi_0(y) = 1 \\ \phi_1(y) = y \end{matrix}\right\}$ $M = \begin{pmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_N \end{pmatrix}$

$$M^t M = \begin{pmatrix} 1 & \dots & 1 \\ y_1 & \dots & y_N \end{pmatrix} \begin{pmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_N \end{pmatrix} = \begin{pmatrix} N & \Sigma y_i \\ \Sigma y_i & \Sigma y_i^2 \end{pmatrix} \qquad M^t X = \begin{pmatrix} 1 & \dots & 1 \\ y_1 & \dots & y_N \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} \Sigma x_i \\ \Sigma x_i y_i \end{pmatrix}$$

•) Resolver

$$\begin{pmatrix} N & \Sigma y_i \\ \Sigma y_i & \Sigma y_i^2 \end{pmatrix} \begin{pmatrix} \overset{a_0'}{\cancel{\Sigma x_i}} \\ \overset{a_1'}{\cancel{\Sigma x_i y_i}} \end{pmatrix} = \begin{pmatrix} \Sigma x_i \\ \Sigma x_i y_i \end{pmatrix}$$

→ Recta $y|x \equiv \dfrac{y - \bar{y}}{\sigma_y} = r \dfrac{(x - \bar{x})}{\sigma_x}$

→ Recta $x|y \equiv \dfrac{x - \bar{x}}{\sigma_x} = r \dfrac{(y - \bar{y})}{\sigma_y}$

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

"r" → "Coeficiente de Correlación de Pearson", mide la magnitud y dirección de las rectas, es decir, da una idea de como es la pendiente de las rectas (tanto $y|x$ como $x|y$)

- ¿Qué recta explica mejor los datos?

La recta que menor error cuadrático medio tenga con respecto a los datos

ECM de $y|x$

$$y = \bar{y} + \frac{r \sigma_y}{\sigma_x}(x - \bar{x})$$

El error cuadrático medio es: $\frac{1}{N}\sum_i (Y_i - \bar{Y}) = \dots = \frac{1}{N}\sum_i \left(y_i - \left(\bar{y} + \frac{\sigma_y}{\sigma_x} r (x - \bar{x})\right)\right)^2 = \dots = \sigma_y^2 (1 - r^2)$

Del mismo modo, el ECM de $x|y$ es $\sigma_x^2(1 - r^2)$

→ $\left\{\begin{array}{l} \text{Será mejor } y|x \text{ si } \sigma_y^2(1-r^2) < \sigma_x^2(1-r^2) \\ \text{Será mejor } x|y \text{ si } \sigma_x^2(1-r^2) < \sigma_y^2(1-r^2) \end{array}\right.$

Observaciones

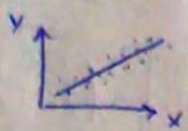
- $r \in [-1, 1]$

$$\underbrace{\frac{1}{N} \sum (y_i - \bar{y})^2}_{\geq 0} = \sigma_y^2 (1 - r^2) \geq 0$$

$$1 - r^2 \geq 0 \; ; \; r^2 \leq 1 \rightarrow r \in [-1, 1]$$

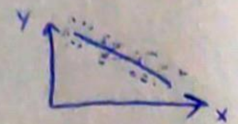
- Si $r > 0 \rightarrow$ Correlación lineal directa

$\boxed{r \in (0, 1)}$



- Si $r < 0 \rightarrow$ Correlación lineal inversa

$\boxed{r \in (-1, 0)}$



- Si $\boxed{r = 0}$

$\rightarrow$ recta $Y|X \Rightarrow y = \bar{y}$

$\rightarrow$ recta $X|Y \Rightarrow x = \bar{x}$

$$\left(\frac{y - \bar{y}}{\sigma_y} = r \, \frac{(x - \bar{x})}{\sigma_x} \right)$$

Variables incorreladas

- Si $\boxed{r = 1 \text{ ó } r = -1} \Rightarrow$ El error cuadrático es 0 $\Rightarrow$ Correlación lineal perfecta.

$\rightarrow$ **Definición**: Dada una nube de puntos llamamos vector residuo a:

$$E : e_i = y_i - y_i^* \qquad \left(y_i^* \text{ valor estimado, es decir } y_i^* = \varphi(x_i) \right)$$

$\rightarrow$ **Definición**: Llamaremos varianza residual a:

$$V_r = \frac{1}{N} \sum_i (e_i - \bar{e})^2 = \frac{1}{N} \sum e_i^2 - E^2$$

$\rightarrow$ **Definición**: Llamaremos coeficiente de determinación a:

$$R^2 = 1 - \frac{V_r}{V_y}$$

$\rightarrow y$ es la variable que quiero explicar

Observaciones:

- El coeficiente R^2 coincide con el coeficiente de correlación de Pearson al cuadrado para el caso en que tenga las rectas $X|Y$ ó Y/X : $\boxed{R^2 = r^2}$

- Si $R^2 = 1 \rightarrow$ Ajuste perfecto
- Si $R^2 = 0 \rightarrow$ Variables incorreladas

6.2 Ajuste de un plano

x	y	z
0	0	0
1	1	2
0	1	1

$N=3$

Ajuste mediante un plano

$$z = \varphi(x,y) = a_0 + a_1 x + a_2 y$$

$$M^t M = \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} = \begin{pmatrix} N & \Sigma x_i & \Sigma y_i \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \end{pmatrix}$$

$$M^t z = \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \Sigma z_i \\ \Sigma z_i x_i \\ \Sigma z_i y_i \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 3 \end{pmatrix}$$

Sol. al sistema lineal

$a_0 = 0 \quad a_1 = 1 \quad a_2 = 1$

$$z = \varphi(x,y) = 0 + x + y = x + y \qquad z^* = x + y = \varphi(x,y)$$

$$R^2 = 1 - \frac{V_r}{V_z} \quad \begin{matrix} \leftarrow \text{varianza de la} \\ \leftarrow \text{variable que quiero aplicar} \end{matrix} \Longrightarrow R^2 = 1 - 0 = 1$$

$$V_r = \frac{1}{N}\left(\Sigma(\varepsilon - \bar{\varepsilon})^2\right); \quad \varepsilon = z - z^* ; \quad \varepsilon = \underbrace{z}_{\text{variable}} - \varphi(x,y) = z - x - y;$$

$$\bar{\varepsilon} = (\overline{z - x - y}) = \bar{z} - \bar{x} - \bar{y} = \frac{3}{3} - \frac{1}{3} - \frac{2}{3} = 0$$

$$V_r = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2 = \frac{1}{3}(e_1^2 + e_2^2 + e_3^2) = \frac{1}{3}\left((z_1 - (x_1+y_1))^2 + (z_2 - (x_2+y_2))^2 + (z_3 - (x_3+y_3))^2\right) = 0;$$

Como $R^2 = 1$ el ajuste es perfecto

Observados: z_i

Estimados: $z_i = \varphi(x_i, y_i) \qquad (x_i, y_i, \varphi(x_i, y_i))$