

1 INTRODUCCIÓN

"Hacer regresión no es más que ajustar los datos a una función para tener conclusiones sobre el ajuste y la relación entre las variables; una explicada y otra que explica (correlación)"

Definición

Sean x, y dos variables aleatorias sobre el mismo espacio de probabilidad, definimos la esperanza de x condicionada a y como una nueva variable aleatoria $E[x|y]$ cuyos valores son las esperanzas $E[x|y=y_0]$ con $y_0 \in \Omega$ espacio muestral del espacio de probabilidad; luego, $E[x|y]$ estará en función de y .

Si $\exists E[x]$ (condición necesaria y suficiente)

• Continuo.

$$E[x|y=y] = \int_{-\infty}^{\infty} x f_{x|y=y}(x) dx \quad \text{con } f_y(y) \neq 0 \quad \forall y \in \Omega$$

• Discreto.

$$E[x|y=y] = \sum_{x \in \mathcal{X}} x P[x=x|y=y] \quad \text{con } P[y=y] \neq 0 \quad \forall y \in \Omega$$

Donde podemos ver que $E[x|y=y]$ es fija, no es más que la esperanza de x considerando como distribución la condicionada $x|y=y$.

De forma similar se define $E[y|x=x]$

Caracterización para el caso $x=g(z)$

Definimos la esperanza de x condicionada y siendo $x=g(z)$, que es la E_x con z variable aleatoria sobre el mismo espacio probabilístico que x y como sigue:

s. $\exists E[g(z)]$: (condición necesaria y suficiente)

• Continuo.

$$E[g(z)|y=y] = \int_{-\infty}^{\infty} g(z) f_{z|y=y}(z) dz \quad \text{con } f_y(y) \neq 0 \quad \forall y \in \Omega$$

• Discreto.

$$E[g(z)|y=y] = \sum_{z \in \mathcal{Z}} g(z) P[z=z|y=y] \quad \text{con } P[y=y] \neq 0 \quad \forall y \in \Omega$$

Ejemplo discípulo

Se extraen sucesiv. sin reemplazamiento 2 bolas de una urna.

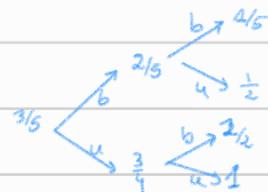
- 3 bolas blancas

- 2 negras

Variables definidas

$$x = \begin{cases} 1 & \text{s. 1^a bola blanca} \\ 0 & \text{s. 1^a bola negra} \end{cases}$$

$$y = \begin{cases} 1 & \text{s. 2^a bola blanca} \\ 0 & \text{s. 2^a bola negra} \end{cases}$$



Calcular $E[Y|X]$, $E[X|Y]$

$X \setminus Y$	0	1	$P[X=x]$
0	$\frac{3}{5} \cdot \frac{1}{2}$	$\frac{3}{5} \cdot \frac{1}{2}$	$\frac{3}{5}$
1	$\frac{2}{5} \cdot \frac{3}{4}$	$\frac{2}{5} \cdot \frac{1}{4}$	$\frac{2}{5}$
$P[Y=y]$	$\frac{3}{5}$	$\frac{2}{5}$	

$$E[X|Y=0] = 0 \cdot P[X=0|Y=0] + 1 \cdot P[X=1|Y=0] = \frac{P[X=1, Y=0]}{P[Y=0]} = \frac{\frac{2}{5} \cdot \frac{3}{4}}{\frac{3}{5}} = \frac{\frac{6}{20}}{\frac{3}{5}} = \frac{\frac{3}{10}}{\frac{3}{5}} = \frac{1}{2}$$

$$E[X|Y=1] = 0 \cdot P[X=0|Y=1] + 1 \cdot P[X=1|Y=1] = \frac{P[X=1, Y=1]}{P[Y=1]} = \frac{\frac{1}{2}}{\frac{2}{5}} = \frac{\frac{1}{10}}{\frac{2}{5}} = \frac{5}{20} = \frac{1}{4}$$

$$E[X|Y=y] = \begin{cases} \frac{1}{2} & \text{s. } y=0 \\ \frac{1}{4} & \text{s. } y=1 \end{cases}$$

De la misma forma obtenemos $E[Y|X]$ obteniendo que

$$E[Y|X=x] = \begin{cases} \frac{1}{2} & \text{s. } x=0 \\ \frac{1}{4} & \text{s. } x=1 \end{cases}$$

Ejeepb coulavo

Sea (X, Y) un vector aleatorio con $f(x, y) = 2$ si $0 < x < y < 1$ función de prob. conjunta.

Calcular $\mathcal{E}[x|y]$ y $\mathcal{E}[y|x]$

$$\int_{-x}^x f(x) = 2 \int_x^0 dy = 2(-y) \Big|_x^0 = 2(x) \quad \forall x \in [0, 1] \quad f_{x+y_0}(x) = \frac{f(x+y_0)}{f(y_0)} = \frac{2}{2y} = \frac{1}{y} \quad \text{acyclic} \quad \forall x \in [0, 1]$$

$$\int_X f(x) dx = 2 \int_0^y dx = 2y \quad \forall y \in [0, 1] \quad \int_{\{x=x_0\}} f(y) dy = \frac{f(x_0, y)}{f_x(x_0)} = \frac{2}{2x_0} = \frac{1}{x_0} \quad \text{as } x_0 \neq 0 \quad \forall y \in [x_0, 1]$$

$$E[X|Y=y] = \int_0^y x \cdot \frac{1}{y} dx = \frac{1}{y} \int_0^y x dx = \frac{1}{y} \left[\frac{x^2}{2} \right]_0^y = \frac{y}{2} \quad \forall y \in [0, 1]$$

$$E[Y|X=x] = \int_x^4 y \frac{1}{1-x} dy = \frac{1}{1-x} \left[\frac{y^2}{2} \right]_x^4 = \frac{1}{1-x} \left(\frac{16}{2} - \frac{x^2}{2} \right) = \frac{(1-x^2)}{2(1-x)} \quad \forall x \in [0, 1]$$

Ejemplo transformación variable discreta

En el exp. aleatorio del lanzamiento de 3 monedas se consideran

$x = u^o$ caras , $y = \text{diferencia en u o entre } u^o \text{ caras y cruces}$

Calcular $E[x^2|y]$

$x \setminus y$	1	3	$P[X=x]$
0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{3}{8}$	0	$\frac{3}{8}$
2	$\frac{3}{8}$	0	$\frac{3}{8}$
3	0	$\frac{1}{8}$	$\frac{1}{8}$

Esta tabla est醕 en el plantador

$x^2 \setminus y$	1	3	$P[x^2 = x^2]$
0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{3}{8}$	0	$\frac{3}{8}$
4	$\frac{3}{8}$	0	$\frac{3}{8}$
9	0	$\frac{1}{8}$	$\frac{1}{8}$
$P[x^2 = x^2]$		$\frac{6}{8}$	$\frac{2}{3}$

$$E[X^2 | Y=1] = X P[X=1 | Y=1] + 2^2 P[X=2 | Y=1] = \frac{P[X=1, Y=1]}{P[Y=1]} + 4 \cdot \frac{P[X=2, Y=1]}{P[Y=1]} = \frac{3/8}{6/8} + 4 \cdot \frac{3/8}{6/8} = \frac{5}{2} = 2.5$$

$$E[X^2 | Y=3] = 3^2 P[X=3 | Y=3] = 9 \cdot \frac{P[X=3, Y=3]}{P[Y=3]} = 9 \cdot \frac{\frac{1}{8}}{\frac{2}{8}} = 9 \cdot \frac{1}{2} = 4.5$$

$$E[x^2 | y=y] = \begin{cases} 2.5 & \text{si } y=1 \\ 4.5 & \text{si } y=3 \end{cases}$$

Group transformation variables continue

Sea (x,y) un p.p. con $f(x,y) = 2$ en $x < y < 1$. Calcular $E[x^3|y]$

Por lo calculado antes:

$$f_Y(y) = \alpha y \quad \forall y \in [0, 1]$$

$$\int g_1(y) = \frac{1}{y} \quad x \in (0, y)$$

$$E[x^3|y] = \int_{-\infty}^{\infty} x^3 f_X(x) dx = \int_0^1 x^3 \frac{1}{y} dx = \frac{1}{y} \left[\frac{x^4}{4} \right]_0^1 = \frac{y^3}{4}$$

Propiedades

1. $E[c|y] = c$

$$E[c|y] = \sum_{x \in E_x} x P[x=c|y=g] = c \cdot \frac{P[x=c, y=g]}{P[y=g]} = c \cdot \frac{P[y=g]}{P[y=g]} = c \quad \square$$

2. Linealidad Sean x, y una definidas en el mismo espacio probabilístico tales que $\exists E[x]$

$$y, a, b \in \mathbb{R} \Rightarrow \exists E[(ax+b)|y] = aE[x|y] + b$$

-Demostración-

Como $\exists E[z]$ sabemos que, si $z = ax + b \Rightarrow E[z] = E[ax] + b = aE[x] + b$ luego $\exists E[ax + b]$.

Vemos ahora la igualdad.

$$\begin{aligned} E[(ax+b)|y] &= \sum_{x \in E_x} (ax+b) P[x|y] = \sum_{x \in E_x} ax P[x|y] + \sum_{x \in E_x} b P[x|y] = a \sum_{x \in E_x} x P[x|y] + b \sum_{x \in E_x} P[x|y] = \\ &= a \sum_{x \in E_x} x P[x|y] + b = aE[x|y] + b \end{aligned} \quad \square$$

3 Sean x_1, \dots, x_n variables aleatorias tales que $\exists E[x_i] \forall i \in \Delta_n \Rightarrow \forall i_1, \dots, i_n \in \mathbb{N}$

$$\exists E[(a_1 x_1 + \dots + a_n x_n)|y] = a_1 E[x_1|y] + \dots + a_n E[x_n|y]$$

-Demostración-

Consecuencia directa de la linealidad de la esperanza condicionada

4 Si $x \geq 0$ y $\exists E[x] \Rightarrow E[x|y] \geq 0$ y $E[x|y] = 0 \Leftrightarrow P[x=0] = 1$

-Demostración-

$$E[x|y] = \sum_{x \in E_x} x P[x|y], \text{ como } P[x|y] \in [0, 1] \text{ y } x \geq 0 \quad \forall x \in E_x \Rightarrow E[x|y] \geq 0$$

$$\Rightarrow \text{Si } E[x|y] = 0 \Rightarrow 0 = \sum_{x \in E_x} x P[x|y] = x_0 P[x_0|y] + \dots + x_n P[x_n|y] \quad \text{donde } \exists i \in \Delta_n | x_i = 0$$

$$\Rightarrow x_0 P[x_0|y] + \dots + x_n P[x_n|y] = 0$$

Trivialmente si $P[x=x_j] > 0$ para alguno $j \in \mathbb{N}$ $E[x|y] \neq 0$!!

\Leftarrow Si $P[x=0] = 1$, como $1 = \sum_{x \in E_x} P[x|y]$ entonces $P[x=x_j] = 0 \quad \forall j \in \mathbb{N} \Rightarrow$

$$E[x|y] = \sum_{x \in E_x} x P[x|y] = 0 P[x^0|y] + \dots + x_n P[x^0|y] = 0$$

\square

5 Conservación del orden. Si x_1, x_2 son dos variables aleatorias tales que $\exists E[x_1], E[x_2]$.

$$y \in \mathbb{R}_n \Rightarrow E[x_1|y] \leq E[x_2|y]$$

-Demostración-

Consecuencia directa de la conservación del orden en la suma y la integral

\square

6 Si x , y son variables aleatorias independientes, para cualquier función medible g , tal que

$\exists E[g(x)]$, se tiene $E[g(x)|y] = E[g(x)]$. En particular $E[x|y] = E[x]$

-Demostración -

x, y independientes

$$E[g(x)|y] = \sum_{x \in \mathcal{E}_x} g(x) P[x|y] = \sum_{x \in \mathcal{E}_x} g(x) \frac{P[x=y]}{P[y]} = \sum_{x \in \mathcal{E}_x} g(x) \frac{P[x=x] P[y|x]}{P[y]} =$$

$$= \sum_{x \in \mathcal{E}_x} g(x) P[x=x] = E[g(x)]$$

Tomando $g = \text{Id}_{\Omega}$ siendo Ω el espacio muestral de x obtenemos la particularidad \square

7. Si x , y son variables aleatorias sobre un mismo espacio probabilístico tales que

$\exists E[g(x)]$ con g una función medible, $\Rightarrow E[E[g(x)|y]] = E[g(x)]$ En particular,

$$E[E[x|y]] = E[x]$$

-Demostración -

$$E[g(x)|y] = \sum_{x \in \mathcal{E}_x} g(x) P[x=y]$$

$$E[E[g(x)|y]] = \sum_{y \in \mathcal{E}_y} E[g(x)|y] P[y] = \sum_{y \in \mathcal{E}_y} \left(\sum_{x \in \mathcal{E}_x} g(x) P[x=y] \right) P[y] =$$

$$= \sum_{y \in \mathcal{E}_y} \left(\sum_{x \in \mathcal{E}_x} g(x) \cdot \frac{P[x=y]}{P[y]} \right) = \sum_{y \in \mathcal{E}_y} \sum_{x \in \mathcal{E}_x} g(x) P[x=y] = \sum_{x \in \mathcal{E}_x} g(x) \sum_{y \in \mathcal{E}_y} P[x=y]$$

$$= \sum_{x \in \mathcal{E}_x} g(x) P[x=x] = E[g(x)]$$

De particularidad se obtiene tomando $g = \text{Id}_{\Omega}$ siendo Ω el espacio muestral de x \square

Momentos condicionados y centrados

Sean x , y variables aleatorias definidas sobre el mismo espacio probabilístico, $y \in \mathbb{R}$.

Definimos el momento condicionado de orden n de x dado y como la variable aleatoria $E[x^n|y]$, siempre que exista dicha esperanza

• Caso discreto

$$E[x^n|y] = \sum_{x \in \mathcal{E}_x} x^n P[x=y] \quad \forall y \in \mathcal{E}_y \text{ con } P[y] > 0$$

• Caso continuo

$$E[x^n|y] = \int_{-\infty}^{\infty} x^n f_{x|y}(x) dx \quad \forall y \in \mathcal{E}_y \text{ con } f_{x|y}(y) > 0$$

Momentos condicionados centrados

Bajo la misma casuística definimos el momento condicionado centrado de orden n de X dado Y como la variable aleatoria $E[(x - E[x|y])^n | Y]$ siempre que exista dicha esperanza.

Caso discreto

$$E[(x - E[x|y])^n | Y] = \sum_{x \in \mathcal{X}_y} (x - E[x|y])^n P[X=x|Y=y] \quad \forall y \in \mathcal{Y} \text{ con } P[Y=y] > 0$$

Caso continuo

$$E[(x - E[x|y])^n | Y] = \int_{-\infty}^{\infty} (x - E[x|y])^n f_{x|y}(x) dx \quad \forall y \in \mathcal{Y} \text{ con } f_y(y) > 0$$

Resaltamos ahora el caso de la varianza condicionada, es decir, el momento centrado de orden 2 dado por

$$\text{Var}[x|y] = E[(x - E[x|y])^2 | Y]$$

que dispone de una gran importancia en el estudio de la regresión

Propiedades

Si $\exists E[x^2]$ siendo x y variables aleatorias:

a) $\exists \text{Var}[x|y] \text{ n } \text{Var}[x|y] \geq 0$. Además $\text{Var}[x|y]=0$ si todos los valores de x son el mismo

b) Punto consecuencia, $\text{Var}[x|y]$ no está acotada superiormente

c) Para poder interpretarla es necesario usar la desviación típica

$$\text{Var}[x|y] = E[x^2|y] - E[x|y]^2$$

$$\text{d)} \text{ Si } \exists \text{Var}[E[x|y]] \text{ y } \exists E[\text{Var}[x|y]] \Rightarrow \text{Var}[x] = \text{Var}[E[x|y]] + E[\text{Var}[x|y]].$$

Ejemplo

Continuación ejemplo A (primero de todos) (bolsa azul)

$$\text{Var}[x] = \text{Var}[E[x|y]] + E[\text{Var}[x|y]]$$

$x y$	0	1	$P[x=y]$
0	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{6}{10}$
1	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$
	$\frac{6}{10}$	$\frac{4}{10}$	1

$$E[x|y] = \begin{cases} \frac{1}{2} & \text{si } y=0 \\ \frac{1}{4} & \text{si } y=1 \end{cases}$$

$$\rightarrow \text{Var}[x] = E[x^2] - E[x]^2 = \frac{2}{5} - \left(\frac{2}{5}\right)^2 = \frac{2}{5} - \frac{4}{25} = \frac{10}{25} - \frac{4}{25} = \frac{6}{25} = \left(\frac{2}{5}\right)^2$$

$$\therefore \text{Var}[\mathcal{E}[x|y]] = \mathcal{E}[\mathcal{E}[x|y]^2] - \mathcal{E}[\mathcal{E}[x|y]]^2 = \mathcal{E}[\mathcal{E}[x|y]^2] - \mathcal{E}[x]^2 = \frac{7}{10} \cdot \frac{4}{25} = \frac{35}{200} - \frac{32}{200} = \frac{3}{200}$$

$$E[E[x_1y]^2] = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{2}{5} = \frac{3}{20} + \frac{1}{40} = \frac{7}{40}$$

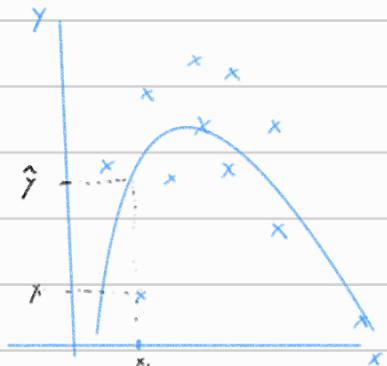
$$\rightarrow \mathbb{E}[\text{Var}[x|y]] = \mathbb{E}[\mathbb{E}[x^2|y] - \mathbb{E}[x|y]^2] = \mathbb{E}[\mathbb{E}[x^2|y]] - \mathbb{E}[\mathbb{E}[x|y]^2] = \mathbb{E}[x^2] - \mathbb{E}[\mathbb{E}[x|y]^2] = \frac{2}{5} - \frac{7}{40}$$

$$= \frac{9}{40}$$

2 Región mielina axodílica

Dada una nube de puntos A que representa la variable aleatoria X siendo X y variables aleatorias buscamos aproximar esa nube de puntos mediante una función $f: x \rightarrow y$.

$$Y \cong \varphi(x) = \hat{y}$$



A la variable y la llamarémos dependiente mientras que x será la independiente.

Para considerar que el problema de una solución óptima, debemos buscar que la diferencia entre f y su mínima; para facilitar los cálculos bajaríamos la suma del cuadrado de las diferencias.

Para ello, definimos la función periódica $E[(Y - \hat{Y}(x))^2]$ conocida como error cuadrático medio o varianza residual

Por tanto, el método de mínimos cuadrados consiste en elegir entre todos los posibles funciones Φ , la que minimiza ese error.

$$\varphi_{\text{opt}}(x) = \min_{\varphi} E[(y - \varphi(x))^2]$$

Curvas de regresión mínimos cuadráticos

En este caso, sin desvirtuarlo esas intenciones que la función f_{opt} viene dadas por:

$$C_{opt}(x) = E[y|x]$$

de donde deducimos que la varianza residual o error cuadrático medio se obtiene como:

Si $\text{Var}[Y] = \text{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\text{Var}[Y|X]]$, es decir $\text{Var}_{\text{explicada}} + \text{Var}_{\text{no explicada}}$; sabiendo que $\mathbb{E}[\text{Var}[Y|X]] = \text{ECM}(\varphi(x))$ entonces:

$$\text{ECM}(\varphi(x)) = \text{Var}[y] - \text{Var}[\varphi_{\text{opt}}(x)]$$

Propiedades

i) Si Y depende funcionalmente de X , es decir, $Y=f(x)$, la curva de regresión de Y en función de X es $Y=f(x)$

ii) Si la dependencia funcional es inversa, se cumple i) y se cumple que $X=f^{-1}(Y)$
luego $\varphi_{\text{opt}}(y)=f'(y)=x$

iii) Si X e Y son independientes \Rightarrow las curvas de regresión son rectas paralelas a los ejes
 $y=\mathbb{E}[Y]$, $x=\mathbb{E}[X]$

En este caso, estaremos una variable sin visicular la otra mediante su esperanza y varianza

En definitiva, es el siguiente resumen

	Sin observar X	Observando X	Paso posterior: $X=x_0$
\hat{Y}	$\mathbb{E}[Y]$	$\mathbb{E}[Y X]$	$\mathbb{E}[Y X=x_0]$
Ecme	$\text{Var}[y]$	$\mathbb{E}[\text{Var}[Y X]]$	$\mathbb{E}[\text{Var}[Y X=x_0]]$

Ejemplo

Se tiran 2 veces un dado con los números 1 y 2.

X = suma de los resultados

Y = máximo de los caras obtenidas

Obtener las funciones de regresión univariadas/cuadráticas de Y dado X y de X dado Y

así como los ECM asociados

$$\varphi_{\text{opt}}(x) = \mathbb{E}[Y|x]$$

$x \setminus y$	1	2	$P[X=x_i]$
2	$\frac{1}{4}$	0	$\frac{1}{4}$
3	0	$\frac{2}{9}$	$\frac{2}{9}$
4	0	$\frac{1}{4}$	$\frac{1}{4}$
$P[Y=y_i]$	$\frac{1}{4}$	$\frac{3}{4}$	

$$\mathbb{E}[x|y=1] = 2 \cdot P[x|y=1] = 2 \cdot \frac{1}{4} \cdot 4 = 2$$

$$\mathbb{E}[x|y=2] = 3 \cdot P[x|y=2] + 4 \cdot P[x|y=2] = 3 \cdot \frac{3}{4} \cdot \frac{4}{1} + 4 \cdot \frac{1}{4} \cdot \frac{4}{3} = 3+4 = 7$$

$$\Phi_{\text{opt}}(y) = \mathbb{E}[x|y] = \begin{cases} 2 & \text{si } y=1 \\ \frac{10}{3} & \text{si } y=2 \end{cases}$$

$$\mathbb{E}[y|x=1] = 1 \cdot P[y|x=1] = \frac{1}{4} \cdot \frac{4}{1} = 1$$

$$\mathbb{E}[y|x=2] = 2 \cdot \frac{3}{4} \cdot \frac{4}{2} = 2$$

$$\mathbb{E}[y|x=3] = 3 \cdot \frac{1}{4} \cdot \frac{4}{1} = 3$$

$$\mathbb{E}[\text{Cov}(\mathbb{E}[y|x])] = \mathbb{E}[\text{Var}_{\text{ex}}[y|x]]$$

$$\text{Var}_{\text{ex}}[y|x] = \mathbb{E}[y^2|x] - \mathbb{E}[y|x]^2 = \begin{cases} 1-1^2=0 & \text{si } x=1 \\ 4-2^2=0 & \text{si } x=2, 3 \end{cases}$$

$$\mathbb{E}[y^2|x] = \begin{cases} \mathbb{E}[y^2|x=2]=1 \cdot P[y|x=2]=1 \\ \mathbb{E}[y^2|x=3]=4 \\ \mathbb{E}[y^2|x=4]=4 \end{cases}$$

Por tanto $\mathbb{E}[\text{Cov}(\mathbb{E}[y|x])] = 0$ luego la dependencia de y respecto a x es fuerte, lo cual se ve fácil en la tabla, para cada valor de x hay un único valor de y con probabilidad unitaria, no será reciproco pues el análogo no cumple la misura condición.

$$\mathbb{E}[\text{Cov}(\mathbb{E}[y|x])] = \mathbb{E}[\text{Var}_{\text{ex}}[\mathbb{E}[y|x]]] = \frac{1}{3}$$

$$\text{Var}_{\text{ex}}[y|x] = \mathbb{E}[x^2|y] - \mathbb{E}[x|y]^2 = \begin{cases} 0 & \text{si } y=1 \\ \frac{2}{9} & \text{si } y=2 \end{cases}$$

$$\mathbb{E}[x^2|y] = \begin{cases} 4 & \text{si } y=1 \\ \frac{24}{9} & \text{si } y=2 \end{cases}$$

Razones de correlación

La correlación estudia la bondad del ajuste de la función de regresión encontrada mediante el método de los mínimos cuadrados, esto es, en qué medida la función de regresión explica una variable a partir de la otra.

Una primera forma podría ser estudiar $\text{Cov}(x,y)$ de manera que:

$\text{Cov}(x,y)=0 \Rightarrow x \text{ e } y \text{ están incorrelacionados}$

$\text{Cov}(x,y) > 0 \Rightarrow x \text{ e } y \text{ están correlacionados y si } x \text{ crece entonces } y \text{ crece}$

$\text{Cov}(x,y) < 0 \Rightarrow x \text{ e } y \text{ están correlacionados y si } x \text{ crece entonces } y \text{ decrece}$

Pero esto no determina el grado de correlación; para ello, usaremos las razones de correlación,

se define la razón de correlación de x sobre y , ρ_{xy}^2 , y de y sobre x , ρ_{yx}^2 , como sigue

$$\rho_{xy}^2 = \frac{\text{Var}[\mathbb{E}[x|y]]}{\text{Var}[x]} = 1 - \frac{\mathbb{E}[\text{Var}_{\text{ex}}[x|y]]}{\text{Var}[x]}$$

$$\eta^2 = \frac{\text{Var}_{Y|X}[\mathbb{E}[Y|X]]}{\text{Var}_Y[Y]} = 1 - \frac{\text{Var}_{Y|X}[Y|X]}{\text{Var}_Y[Y]}$$

Observaciones

• Es una medida adimensional que representa la proporción de varianza de la variable dependiente que queda explicada por la función de regresión.

• Es una medida de bondad del ajuste que se interpreta como sigue:

- Si $\eta^2=0 \Rightarrow$ el ajuste es pésimo.
- Si $\eta^2=1 \Rightarrow$ el ajuste es perfecto.

Propiedades

i) $\eta_{Y|X}^2, \eta_{X|Y}^2 \in [0,1]$

ii) $\eta_{Y|X}^2 = 0 \Leftrightarrow \mathbb{E}_{\text{Opf}}(x) = \mathbb{E}[x]$

$\eta_{X|Y}^2 = 0 \Leftrightarrow \mathbb{E}_{\text{Opf}}(y) = \mathbb{E}[y]$

iii) $\eta_{Y|X}^2 = \eta_{X|Y}^2 = 0 \Leftrightarrow x \text{ y } y \text{ están incorrelados.}$

iv) $\eta_{Y|X}^2 = 1 \Leftrightarrow x \text{ y } y \text{ dependen funcionalmente}$

$\eta_{X|Y}^2 = 1 \Leftrightarrow x \text{ y } y \text{ dependen funcionalmente}$

v) $\eta_{X|Y}^2 = \eta_{Y|X}^2 = 1 \Leftrightarrow x \text{ y } y \text{ dependen funcionalmente recíprocamente.}$

Ejemplo (continuación anterior)

$$\eta_{Y|X}^2 = 1 - \frac{\mathbb{E}[\text{Var}[Y|X]]}{\text{Var}_Y[Y]} = 1 - \frac{0}{\frac{13}{4}} = 1 \Rightarrow \text{hay dep. funcional}$$

$$\text{Var}_Y[Y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 = \frac{13}{4} - \frac{49}{16} = \frac{3}{16}$$

$$\mathbb{E}[y^2] = 1 \cdot \frac{1}{4} + 4 \cdot \frac{3}{4} = \frac{1}{4} + 3 = \frac{13}{4}$$

$$\mathbb{E}[y] = 1 \cdot \frac{1}{4} + 2 \cdot \frac{3}{4} = \frac{1}{4} + \frac{6}{4} = \frac{7}{4} = \frac{49}{16}$$

$$\eta_{X|Y}^2 = 1 - \frac{\mathbb{E}[\text{Var}[X|Y]]}{\text{Var}_X[X]} = 1 - \frac{1}{6} \cdot 2 = 1 - \frac{1}{3} = \frac{2}{3} \Rightarrow \text{la función ajustable explica el } 66.66\% \text{ de la variabilidad de } x \text{ a partir de } y \rightarrow \text{realmente es malo.}$$

$$\text{Var}_X[X] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{19}{2} - 9 = \frac{1}{2}$$

$$\mathbb{E}[x^2] = \frac{19}{2}$$

$$\mathbb{E}[x] = 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{1}{2} + \frac{3}{2} + 1 = 3$$

Ejemplo en continua

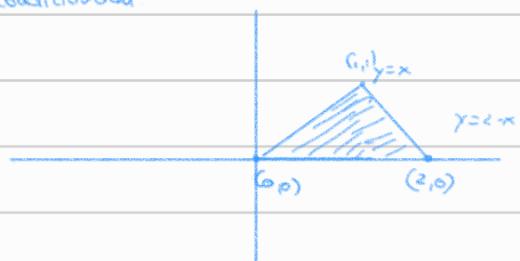
Sea (x, y) un vector aleatorio con función de densidad $f(x,y)$ con $(x,y) \in T$ siendo T el triángulo de vértices $(0,0)$, $(2,0)$ y $(1,1)$. Obtener la función de regresión óptima de la variable y a partir de x , se ECU e interpretar el ajuste.

Méjase falta la marginal para obtener la condicionada.

$$1. f_x(a)$$

$$2. f_{y|x}$$

$$3. E[y|x]$$



$$f_x(x) = \begin{cases} \int_0^x dy = x & \text{si } 0 \leq x \leq 1 \\ \int_0^{2-x} dy = 2-x & \text{si } 1 \leq x \leq 2 \end{cases}$$

Por tanto

$$f_{y|x}(x,y) = \frac{f(x,y)}{f_x(x)} = \begin{cases} \frac{1}{x} & \text{si } 0 \leq y \leq x \quad (0 \leq x \leq 1) \\ \frac{1}{2-x} & \text{si } 0 \leq y \leq 2-x \quad (1 \leq x \leq 2) \end{cases}$$

$$\text{Obtenemos } E[y|x] = \int_{\mathbb{R}} y f_{y|x}(x,y) = \begin{cases} \frac{1}{x} \int_0^x y dy = \frac{x}{2} & \text{si } 0 \leq x \leq 1 \\ \frac{1}{2-x} \int_0^{2-x} y dy = \frac{2-x}{2} & \text{si } 1 \leq x \leq 2 \end{cases}$$

Entonces la función óptima es:

$$E[y|x] = \begin{cases} \frac{x}{2} & \text{si } 0 \leq x \leq 1 \\ \frac{2-x}{2} & \text{si } 1 \leq x \leq 2 \end{cases}$$

$$ECU(E[y|x]) = E[\text{Var}[y|x]] = \int_{\mathbb{R}} \text{Var}[y|x] \cdot f_x(x) dx = \begin{cases} \int_0^1 \frac{x^2}{12} \cdot x dx = \left[\frac{x^4}{48} \right]_0^1 = \frac{1}{48} \\ \int_1^2 \frac{(2-x)^2}{12} \cdot (2-x) dx = \frac{1}{24} \end{cases}$$

$$\text{Var}[y|x] = E[y^2|x] - E[y|x]^2 = \begin{cases} \frac{x^2}{3} - \frac{x^2}{4} = \frac{x^2}{12} & \text{si } 0 \leq x \leq 1 \\ \frac{2-x}{3} \cdot \frac{2-x}{4} = \frac{(2-x)^2}{12} & \text{si } 1 \leq x \leq 2 \end{cases}$$

$$E[y^2|x] = \begin{cases} \int_0^x y^2 \cdot \frac{1}{x} dy = \frac{x^2}{3} & \text{si } 0 \leq x \leq 1 \\ \int_0^{2-x} y^2 \cdot \frac{2-x}{2} dy = \frac{(2-x)^2}{3} & \text{si } 1 \leq x \leq 2 \end{cases}$$

Obtenemos la razón de correlación:

$$\eta^2_{y|x} = 1 - \frac{ECU(E[y|x])}{\text{Var}[y]} = 1 - \frac{\frac{1}{48}}{\frac{1}{18}} = 1 - \frac{63}{64} = \frac{1}{64} \Rightarrow \text{malo ajuste}$$

$$\text{Var}[y] = E[y^2] - E[y]^2 = \frac{1}{18}$$

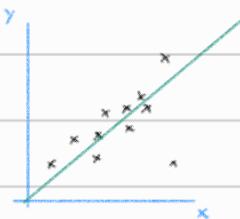
3 Rectas de regresión

En ocasiones, es de especial interés conocer la recta de puntos puros si identificamos su forma será útil para determinar la mejor curva de regresión, que será de ese tipo.

En caso interesante será el lineal, es decir, si x, y son dos v.a. relacionadas entre sí.

$$y = a + bx \quad \text{con } a, b \in \mathbb{R}$$

Gráficamente



A estos funciones $y = f(x) = a + bx$ las llamaremos rectas de regresión.

Planteamiento de la recta de regresión de y sobre x .

Siguiendo la filosofía de la regresión mínima cuadrática buscamos una función $\Phi_{\text{opt}}(x)$ que minimice el error cuadrático medio, es decir:

$$\Phi_{\text{opt}}(x) = \min_{a, b} E[(y - a - bx)^2]$$

Por tanto, sólo deberemos obtener $a, b \in \mathbb{R}$ tales que minimicen la función $L = E[(y - a - bx)^2]$.

Teniendo que $L = E[y^2] + b^2 E[x^2] + a^2 + 2abE[x] - 2bE[xy] - 2aE[y]$ donde lleva usado la linealidad de la esperanza.

Derivando respecto a y y b e igualando a cero obtenemos el sistema de ecuaciones normales

$$\left. \begin{array}{l} bE[x^2] + aE[x] - E[xy] \\ a + bE[x] = E[y] \end{array} \right\}$$

cuya solución es $b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$ y $a = E[y] - bE[x]$ obteniendo así la recta R_{xy} dada por

$$\Phi_{\text{opt}}(x) = y = a + bx, \quad b \text{ y } a \text{ los ya conocidos}$$

No obstante, queremos llegar más allá y obtener el $\text{ECM}(\varphi_{\text{opt}}(x))$ que viene dado por

$$\text{Ecu}(\varphi_{\text{opt}}(x)) = \text{Var}[y] - \frac{[\text{Cov}(x,y)]^2}{\text{Var}(x)}$$

donde $\text{Var}[y]$ es la variabilidad total y $\frac{[\text{Cov}(x,y)]^2}{\text{Var}(x)}$ es la variabilidad explicada por el ajuste; entonces despejando:

$$\text{Var}[y] = \text{Var}[\varphi_{\text{opt}}(x)] + \text{ECM}(\varphi_{\text{opt}}(x))$$

Observación

Todo lo anterior es exactamente análogo para explicar x en función de y con una regresión lineal.

Coeficiente de determinación lineal

Para poder estudiar la bondad del ajuste dentro una interpretación usaremos el coeficiente de determinación lineal definido como

$$P_{x,y}^2 = \frac{[\text{Cov}(x,y)]^2}{\text{Var}(x)\text{Var}(y)}$$

Observación

Este coeficiente puede calcularse de varias formas

$$i) P_{x,y} = P_{y,x}$$

$$ii) P_{x,y} = \frac{\text{Var}(\varphi_{\text{opt}}(x))}{\text{Var}(x)}$$

$$iii) P_{x,y} = \frac{\text{Var}(\varphi_{\text{opt}}(y))}{\text{Var}(y)}$$

Como devolvemos que de aquí no se puede deducir que $\varphi_{\text{opt}}(x) = \varphi_{\text{opt}}^{-1}(y)$; de hecho, esto es universalmente falso.

Añadimos, este coeficiente sólo es útil en el caso lineal y no dispone de interpretación sobre la bondad en otros ajustes que no sean linealizables.

Propiedades

i) $P_{ax+bx, cy+dy}^2 = P_{x,y}$

ii) $0 \leq P_{x,y}^2 \leq 1$

iii) $P_{x,y}^2 = 0 \Leftrightarrow y = E[y], x = E[x] \Leftrightarrow \text{Cov}(x,y) = 0 \Leftrightarrow \text{están independientes}$

iv) $P_{x,y}^2 = 1 \Leftrightarrow \text{existe dependencia funcional entre } x \text{ y } y$

v) $P_{x,y}^2 \leq \eta_{yx}^2 \wedge P_{x,y}^2 \leq \eta_{xy}^2$ Además, si $E[y|x] \cdot \eta_{yx}^2 + E[x|y] \cdot \eta_{xy}^2 = 1 \Rightarrow$ se cumple la igualdad

Coeficiente de correlación lineal de Pearson

Responde a que el coeficiente de determinación lineal da toda la información acerca de la bondad de nuestro ajuste lineal, la medida de correlación por excelencia que nos permite conocer la bondad del ajuste es el coeficiente de correlación lineal que aparta dos tipos de interpretación

1 Fuerza de asociación o grado de correlación

2 Sentido de correlación

- Si es negativo \Rightarrow crecimiento en sentido opuesto \Rightarrow

- Si es positivo \Rightarrow crecimiento en el mismo sentido \Rightarrow

Se define de la siguiente forma:

$$P_{x,y} = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}[x] \text{Var}[y]}} = P_{y,x}$$

Observación

El coeficiente de correlación lineal se puede obtener de la siguiente forma:

$$P_{x,y}^2 = P_{x,y} \cdot P_{y,x} = (P_{x,y})^2$$

Propiedades

i) $-1 \leq P_{x,y} \leq 1$

ii) $P_{ax+bx, cy+dy} = \pm P_{x,y}$

iii) Si x, y son independientes $\Rightarrow P_{x,y} = 0$

iv) Si $P_{x,y} = 1 \Rightarrow$ la dependencia lineal es exacta y positiva

v) Si $P_{x,y} = -1 \Rightarrow$ la dependencia lineal es exacta y negativa

vi) $1_{\mathbb{P}_{x_1, x_2}}(1=1 \Leftrightarrow)$ existe dependencia funcional lineal \Leftrightarrow las dos rectas de regresión coinciden
entresí y con la recta de dependencia, además de con las curvas de regresión

4. Covariancia y correlación

Sean x_1, x_2 dos variables aleatorias, se define su covarianza como

$$\text{Cov}(x_1, x_2) = E[(x_1 - E[x_1])(x_2 - E[x_2])] = E[x_1 x_2] - E[x_1]E[x_2]$$

Dispone de la siguiente interpretación:

- Si $\text{Cov}(x_1, x_2) > 0 \Rightarrow$ crecimiento similar, es decir, ambas tienen una relación fuerte.
- Si $\text{Cov}(x_1, x_2) < 0 \Rightarrow$ crecimiento inverso.
- Si $\text{Cov}(x_1, x_2) = 0 \Rightarrow$ están incorrelacionadas/neutralizadas.

Debe destacar que estas medidas de correlación no pueden capturar con precisión las relaciones curvilineas, por ejemplo, dos variables incorrelacionadas no tienen por qué ser independientes ya que "incorreladas" significa que no existe relación lineal pero no significa que no tenga curvilinearidad.

El recíproco del ejemplo es cierto