

Estadística 24911

Ejercicios de regresión y correlación

Manuel Sánchez Mateos
Móvil: 650 288 220
Fax: 952 264 746
msm_77@hotmail.com
www.telefonica.net/web/msm77

Proyecto: ESTADÍSTICA

Referencia: 24911

Sección: REGRESIÓN Y CORRELACIÓN

1. UN EXPERIMENTO HA PRODUCIDO LA SIGUIENTE TABLA DE DATOS:

	X	2	3	4	5	6
Y						
0,944	1					
0,992	1					
1,345			1			
1,368			2			
1,392			2			
1,398			1			
1,746				2		
1,768				6		
1,886				1		
1,91				1		
2,256					2	
2,471					3	
2,688						1
2,981						1

DETERMINAR

1.1 DISTRIBUCIÓN MARGINAL X VS M_x 1.2 \bar{x} 1.3 $\sigma_{x,m}$ 1.4 $\sigma_{x,m-1}$ 1.5 DISTRIBUCIÓN MARGINAL Y VS M_y 1.6 \bar{y} 1.7 $\sigma_{y,m}$ 1.8 $\sigma_{y,m-1}$

1.9 DISTRIBUCIONES CONDICIONALES

 $X | Y = 1,768$, $X | Y = 2,688$ VS M_x

1.10 DISTRIBUCIONES CONDICIONALES

 $Y | X = 3$, $Y | X = 6$ VS M_y

1.11 COVARIANZA DE X CON Y

1.12 COEFICIENTE CORRELACIÓN LINEAL DE Y CON X

1.13 RECTA DE REGRESIÓN DE Y VS $X = \hat{Y}(x)$ 1.14 VALORES DE $\hat{Y}(2)$, $\hat{Y}(3)$, $\hat{Y}(4)$, $\hat{Y}(5)$, $\hat{Y}(6)$, $\bar{\hat{Y}}$

1.15 VARIANZA EXPLICADA DE Y , VARIANZA NO EXPLICADA DE Y

1.16 COEFICIENTE DE DETERMINACIÓN DE Y CON X

1.17 ERROR TÍPICO DE LA REGRESIÓN

1.1 DISTRIBUCIÓN MARGINAL X VS M_x

RESULTA DE SUMAR TODAS LAS FRECUENCIAS DE X (M_x) INDEPENDIENTEMENTE DEL VALOR DE Y.

MARGINAL X

X	2	3	4	5	6
M _x	2	6	10	5	2

1.2 MEDIA DE X

$$\bar{x} = \sum x_i f_{xi} = \sum x_i \frac{m_{xi}}{\sum m_{xi}} = \frac{\sum x_i m_{xi}}{N_x} = \frac{2 \cdot 2 + 3 \cdot 6 + 4 \cdot 10 + 5 \cdot 5 + 6 \cdot 2}{2 + 6 + 10 + 5 + 2}$$

$$\bar{x} = 3,96$$

1.3 VARIANZA DE X

$$\sigma_{x,m}^2 = \sum (x_i - \bar{x})^2 f_{xi} = \frac{\sum (x_i - \bar{x})^2 m_{xi}}{N_x} = \frac{\sum x_i^2 m_{xi}}{N_x} - \bar{x}^2$$

$$\sigma_{x,m}^2 = \frac{419}{25} - 3,96^2 \Rightarrow \sigma_{x,m} = 1,038460399 \quad (1.1)$$

1.4 UN MEJOR ESTIMADOR DE LA VARIANZA (POBLACIONAL) DE X ES:

$$\sigma_{x,m-1}^2 = \frac{\sum (x_i - \bar{x})^2 m_{xi}}{N_x - 1} = \frac{\sum x_i^2 m_{xi}}{N_x - 1} - \frac{N_x}{N_x - 1} \bar{x}^2$$

$$\sigma_{x,m-1}^2 = \frac{419}{25-1} - \frac{25}{25-1} 3,96^2 \Rightarrow \sigma_{x,m-1} = 1,059874206$$

1.5 DISTRIBUCIÓN MARGINAL Y VS M_y

RESULTA DE SUMAR TODAS LAS FRECUENCIAS DE Y (M_y) INDEPENDIENTEMENTE DEL VALOR DE X.

MARGINAL Y

Y	0,944	0,992	1,345	1,368	1,392	1,398	1,746	1,768	1,886	1,91	2,256
M _y	1	1	1	2	2	1	2	6	1	1	2
Y	2,471	2,688	2,981								
M _y	3	1	1								

Proyecto: ESTADÍSTICA

Referencia: 24911

Sección: 1. REGRESIÓN Y CORRELACIÓN

1.6 MEDIA DE Y

$$\bar{y} = \frac{\sum Y_i m_{yi}}{\sum m_{yi}} = \frac{\sum Y_i m_{yi}}{N_y} = \frac{45,689}{25} = 1,82756$$

1.7 VARIANZA DE Y

$$\sigma_{y,m}^2 = \frac{\sum Y_i^2 m_{yi}}{N_y} - \bar{y}^2 = \frac{89,922177}{25} - 1,82756^2 \Rightarrow \sigma_{y,m} = 0,5068644063 \quad (1.2)$$

1.8 UN MEJOR ESTIMADOR DE LA VARIANZA (POBLACIONAL) DE Y ES:

$$\sigma_{y,m-1}^2 = \frac{\sum Y_i^2 m_{yi}}{N_y - 1} - \frac{N_y}{N_y - 1} \bar{y}^2 = \frac{89,922177}{25 - 1} - \frac{25}{25 - 1} 1,82756^2 \therefore$$

$$\sigma_{y,m-1} = 0,5173163184$$

1.9 DISTRIBUCIONES CONDICIONALES DE X

X	2	3	4	5	6
$m_x 1,768$	0	0	6	0	0
$m_x 2,688$	0	0	0	0	1

1.10 DISTRIBUCIONES CONDICIONALES DE Y

Y	0,944 ...	1,345	1,368	1,392	1,398 ...	2,688	2,981		
$m_y 3$	0	...	1	2	2	1	...	0	0
$m_y 6$	0	1	1	

1.11 COVARIANZA DE X CON Y

$$\text{COV}(X, Y) = \sum (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) m_{ij}}{\sum m_{ij}}$$

$$\text{COV}(X, Y) = \frac{\sum x_i y_i m_{ij}}{N} - \bar{x} \bar{y} = \frac{193,884}{25} - 3,96 \cdot 1,82756$$

$$\text{COV}(X, Y) = 0,5182224 \quad \text{OBSERVAR: } N_x = N_y = N$$

1.12 COEFICIENTE CORRELACION LINEAL (DE PEARSON) DE Y CON X

$$r = \frac{\text{COV}(X, Y)}{\sigma_x \sigma_y} = \frac{0,5182224}{1,038460399 \cdot 0,5068644063}$$

$$r = 0,9845424519$$

1.13 RECTA DE REGRESIÓN Y VS X

$$\hat{Y} - \bar{Y} = B(X - \bar{X}) \Rightarrow \hat{Y} = BX + \bar{Y} - B\bar{X} = BX + A \quad (1.3)$$

$$B = \frac{\text{COV}(X, Y)}{\sigma_x^2} = \frac{0,5182224}{1,038460399^2} = 0,4805474776 \quad (1.4)$$

$$A = \bar{Y} - B\bar{X} = 1,82756 - 0,4805474776 \cdot 3,96 = -0,07540801136$$

1.14 VALORES DE $\hat{Y}(x)$

APLICANDO REITERADAMENTE (1.3)

X	2	3	4	5
\hat{Y}	0,8856869438	1,366234421	1,846781899	2,327329377
m	2	6	10	5

x	6
\hat{Y}	2,807876854
m	2

Sección: 1.14 MEDIA DE LA Y ESTIMADA

$$\bar{\hat{Y}} = \frac{\sum \hat{Y}_i m_i}{\sum m_i} = \frac{\sum \hat{Y}_i m_i}{N} = 1,82756 = \bar{Y}$$

LAS MEDIAS DE \hat{Y} (REGRESIÓN) E Y SIEMPRE COINCIDEN

1.15 VARIANZAS EXPLICADA E INEXPLICADA DE Y

$$\sigma_{\hat{Y},m}^2 = \text{VAR}(\hat{Y}) = \text{VAR}(BX+A) = \text{VAR}(BX) + \text{VAR}(A) + 2\text{COV}(BX, A) \quad (1.5)$$

$$\text{VAR}(BX) = \frac{\sum (BX_i - \overline{BX})^2 m_i}{N} = \frac{\sum (BX_i - B\bar{X})^2 m_i}{N} = B^2 \frac{\sum (X_i - \bar{X})^2 m_i}{N} = B^2 \sigma_x^2 \quad (1.6)$$

$$\text{VAR}(A) = \frac{\sum (A - \bar{A})^2 m_i}{N} \rightarrow A = \text{CTE} \Rightarrow A = \bar{A} \Rightarrow \text{VAR}(A) = 0 \quad (1.7)$$

$$\text{COV}(BX, A) = \frac{\sum (BX - \overline{BX})(A - \bar{A}) m_i}{N} = 0 \quad (1.8)$$

$$(1.6), (1.7), (1.8) \rightarrow (1.5) \Rightarrow \text{VAR}(\hat{Y}) = B^2 \sigma_x^2 = \sigma_{\hat{Y},m}^2 \rightarrow (1.1), (1.4):$$

$$\sigma_{\hat{Y},m}^2 = 0,4990295254^2 \quad (1.9) \text{ VARIANZA EXPLICADA DE } Y$$

SE VERIFICA QUE:

$$\sigma_{\hat{Y},m-1}^2 = \frac{N}{N-1} \sigma_{\hat{Y},m}^2 = \frac{25}{24} 0,4990295254^2$$

$$\sigma_{Y,m}^2 = \sigma_{\hat{Y},m}^2 + \sigma_{e,m}^2 \quad (1.10)$$

$$\sigma_{\hat{Y},m-1} = 0,5093198766$$

$\sigma_{e,m}^2$: VARIANZA NO EXPLICADA DE Y

$e = Y - \hat{Y}$: ERROR (O RESIDUO) EN LA ESTIMACIÓN DE Y MEDIANTE \hat{Y}

DEBIDO A QUE e SE CONSIDERA ALEATORIO SE VERIFICA:

$$\bar{e} = 0$$

$\text{COV}(\hat{Y}, e) = 0$ \hat{Y} y e SON (ESTADÍSTICAMENTE) INDEPENDIENTES

$$(1.10) \rightarrow \sigma_{e,m}^2 = \sigma_{y,m}^2 - \hat{\sigma}_{y,m}^2 \rightarrow (1.2), (1.9) \Rightarrow \sigma_{e,m}^2 = 7,881059153 \cdot 10^{-3} \quad (1.11)$$

VARIANZA NO EXPLICADA DE Y

1.16 COEFICIENTE DE DETERMINACIÓN

$$CD^2 = \frac{\hat{\sigma}_{y,m}^2}{\sigma_{y,m}^2} = \frac{\sigma_{y,m}^2 - \sigma_{e,m}^2}{\sigma_{y,m}^2} = 1 - \frac{\sigma_{e,m}^2}{\sigma_{y,m}^2} \quad (1.12)$$

EL CD ES ÚTIL PARA EVALUAR LA BONDAD DE AJUSTE DE MODELOS NO LINEALES. PERO EN EL CASO LINEAL CD TIENE QUE SER IGUAL A Y.

$$(1.2), (1.9), (1.11) \rightarrow (1.12) \Rightarrow CD^2 = \frac{0,4990295254^2}{0,5068644063^2} = 1 - \frac{7,881059153 \cdot 10^{-3}}{0,5068644063^2}$$

$$CD = 0,984542452 = Y \quad (1.13)$$

1.17 ERROR TÍPICO REGRESIÓN

$$\sigma_{e,m-2} = \left(\sigma_{e,m-2}^2 \right)^{\frac{1}{2}} = \left(\frac{N}{N-2} \sigma_{e,m}^2 \right)^{\frac{1}{2}} = \left(\frac{25}{23} 7,881059153 \cdot 10^{-3} \right)^{\frac{1}{2}} \quad (1.14)$$

$$\sigma_{e,m-2} = 0,09255467921$$

EN ESTE CASO, EL MEJOR ESTIMADOR DEL ERROR (POBLACIONAL) DE LA REGRESIÓN:

$$e = Y - \hat{Y} = Y - (Bx + A)$$

RESULTA SER (1.14) CON UN DENOMINADOR N-2 DEBIDO A QUE "SE HAN GASTADO 2 GRADOS DE LIBERTAD DE LOS N ORIGINALES" AL TENERSE QUE CALCULAR PREVIAMENTE LAS 2 CONSTANTES B y A CON ESTOS N DATOS.

ESTA REGLA TIENE VALIDEZ GENERAL Y YA HA SIDO EMPLEADA EN 1.4 DURANTE EL CÁLCULO DE $\sigma_{x,m-1}^2$ CUANDO TUVO QUE EVALUARSE ANTES \bar{X} Y EN 1.8 DURANTE EL CÁLCULO DE $\sigma_{y,m-1}^2$ CUANDO TUVO QUE EVALUARSE PREVIAMENTE A \bar{Y} .

Sección: 2. REGRESIÓN Y CORRELACIÓN

2. UN EXPERIMENTO HA PRODUCIDO LA SIGUIENTE TABLA DE VALORES:

Y	X ₁	X ₂
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10
58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
57	4	15
19	1	20
31	2	20
33	3	20
43	4	20

2.1 CALCULAR LA ECUACIÓN DEL PLANO DE REGRESIÓN

2.2 ESTIMAR EL VALOR DE Y CUANDO
 $x_1 = 2,5$, $x_2 = 12$

2.3 CALCULAR LA VARIANZA TOTAL DE Y
 σ_y^2 , LA VARIANZA EXPLICADA DE Y σ_y^2
Y LA VARIANZA NO EXPLICADA DE Y
 σ_e^2

2.4 CALCULAR EL COEFICIENTE DE DETERMINACIÓN

2.5 CALCULAR EL ERROR TÍPICO DE LA REGRESIÓN

SOLUCIÓN

2.1 EMPLEANDO EL MÉTODO DE LOS MÍNIMOS CUADRADOS (MINIMIZANDO EL VALOR DE $\sum (Y - \hat{Y})^2 = \sum (Y - (b_1 X_1 + b_2 X_2 + b_0))^2$ PUEDE DEMOSTRARSE LA ESTRUCTURA DEL SISTEMA DE ECUACIONES PARA EL CÁLCULO DE b_0 , b_1 , b_2

$$b_0 N + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y \quad (2.1)$$

$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

DE LOS DATOS SE OBTIENEN:

$$\sum X_1 = 40 \quad \sum X_2 = 200 \quad N = 16 \quad (2.2)$$

$$\sum X_1^2 = 120 \quad \sum X_1 X_2 = 500 \quad \sum X_2^2 = 3000$$

$$\sum Y = 723 \quad \sum X_1 Y = 1963 \quad \sum X_2 Y = 8210$$

$$(2.2) \rightarrow (2.1) \Rightarrow b_0 = 46,4375 \quad b_1 = 7,775 \quad b_2 = -1,655 \quad (2.3)$$

PLANO DE REGRESION

$$\hat{Y} = 7,775 X_1 - 1,655 X_2 + 46,4375 \quad (2.4)$$

2.2

$$\hat{Y}(2,5; 12) = 7,775 \cdot 2,5 - 1,655 \cdot 12 + 46,4375 = 46,015 \quad (2.5)$$

2.3 VARIANZAS DE Y, \hat{Y} y $e = Y - \hat{Y}$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{723}{16} = 45,1875 \quad (2.5)$$

$$\sigma_{Y,M}^2 = \frac{\sum Y^2}{N} - \bar{Y}^2 = \frac{35483}{16} - 45,1875^2 = 175,7773438 \quad (2.6)$$

VARIANZA TOTAL DE Y

$$N \sigma_{\hat{Y},M}^2 = \sum (\hat{Y} - \bar{Y})^2 \rightarrow \bar{\hat{Y}} = \bar{Y} \Rightarrow N \sigma_{\hat{Y},M}^2 = \sum (\hat{Y} - \bar{Y})^2 \rightarrow (2.4) \therefore$$

$$N \sigma_{\hat{Y},M}^2 = \sum (b_1 X_1 + b_2 X_2 + b_0 - \bar{Y})^2 = \sum (b_1^2 X_1^2 + b_2^2 X_2^2 + (b_0 - \bar{Y})^2 +$$

$$+ 2 b_1 b_2 X_1 X_2 + 2 b_1 (b_0 - \bar{Y}) X_1 + 2 b_2 (b_0 - \bar{Y}) X_2$$

$$N \sigma_{\hat{Y},M}^2 = b_1^2 \sum X_1^2 + b_2^2 \sum X_2^2 + N (b_0 - \bar{Y})^2 + 2 b_1 b_2 \sum X_1 X_2 + 2 b_1 (b_0 - \bar{Y}) \sum X_1 +$$

$$+ 2 b_2 (b_0 - \bar{Y}) \sum X_2 \quad (2.7)$$

(2.3), (2.2), (2.5) \rightarrow (2.7):

$$\sigma_{\hat{Y},M}^2 = \frac{1}{16} 2578,525 \Rightarrow \sigma_{\hat{Y},M}^2 = 161,1578125 \quad (2.8) \quad \text{VARIANZA EXPLICADA DE Y}$$

Sección: 2.3 VARIANZAS DE Y , \hat{Y} Y $e = Y - \hat{Y}$

$$N\sigma_{Y,m}^2 = \sum (Y - \bar{Y})^2 = \sum (\hat{Y} + e - \bar{Y})^2 = \sum (\hat{Y} - \bar{\hat{Y}} + e)^2 = \sum (\hat{Y} - \bar{\hat{Y}} + e - \bar{e})^2$$

\uparrow $(\bar{Y} = \bar{\hat{Y}})$ \uparrow $(\bar{e} = 0)$

$$N\sigma_{Y,m}^2 = \sum ((\hat{Y} - \bar{\hat{Y}})^2 + (e - \bar{e})^2 + 2(\hat{Y} - \bar{\hat{Y}})(e - \bar{e}))$$

$$N\sigma_{Y,m}^2 = N\sigma_{\hat{Y},m}^2 + N\sigma_{e,m}^2 + N2\text{cov}(\hat{Y}, e) \quad (2.9)$$

e SE CONSIDERA ALEATORIO $\Rightarrow \hat{Y}$ Y e SON (ESTADÍSTICAMENTE) INDEPENDIENTES:

$$\text{cov}(\hat{Y}, e) = 0 \quad (2.10)$$

$$(2.10) \rightarrow (2.9) \Rightarrow \sigma_{Y,m}^2 = \sigma_{\hat{Y},m}^2 + \sigma_{e,m}^2 \Rightarrow \sigma_{e,m}^2 = \sigma_{Y,m}^2 - \sigma_{\hat{Y},m}^2 \quad (2.11)$$

$$(2.6), (2.8) \rightarrow (2.11) \Rightarrow \sigma_{e,m}^2 = 14,6195313 \quad (2.12) \quad \text{VARIANZA NO EXPLICADA DE } Y$$

2.4 COEFICIENTE DE DETERMINACIÓN

$$CD^2 = \frac{\sigma_{\hat{Y},m}^2}{\sigma_{Y,m}^2} = 1 - \frac{\sigma_{e,m}^2}{\sigma_{Y,m}^2} \rightarrow (2.6), (2.8), (2.12) \Rightarrow CD^2 = 0,9168292626$$

$CD = 0,957512017$ COEFICIENTE DE DETERMINACIÓN DE LA REGRESIÓN

2.5 ERROR TÍPICO DE LA REGRESIÓN

$$\text{EL CÁLCULO DE } N\sigma_{e,m}^2 = \sum (e - \bar{e})^2 = \sum e^2 = \sum (Y - \hat{Y})^2 \quad (2.13)$$

\uparrow $(\bar{e} = 0)$

(2.13) REQUIERE DE LA PREVIA DETERMINACIÓN DE LOS 3 PARÁMETROS: b_0, b_1, b_2 .

LUEGO "QUEDAN $N-3$ GRADOS DE LIBERTAD DISPONIBLES".

$$\sigma_{e,m-3} = \left(\sigma_{e,m-3}^2 \right)^{\frac{1}{2}} = \left(\frac{N}{N-3} \sigma_{e,m}^2 \right)^{\frac{1}{2}} = \left(\frac{16}{13} 14,6195313 \right)^{\frac{1}{2}}$$

$$\sigma_{e,m-3} = 4,241847391 \quad (2.13) \quad \text{ERROR TÍPICO REGRESIÓN}$$