Regresión Lineal

Tema 2

ESTUDIO CONJUNTO DE DOS VARIABLES

Tendremos una variable bidimensional (x,y), que se referirá a dos características de un mismo individuo.

- Cada fila indica los datos de un individuo
- En cada columna se expresan los valores que toma una variable sobre los individuos

Podemos representar las observaciones en un diagrama de dispersión o nube de puntos. En él, cada individuo es un punto cuyas coordenadas son los valores de las variables.

Si las variables están correlacionadas, el gráfico mostrará algún nivel de correlación (tendencia) entre las dos variables. Si no hay ninguna correlación, el gráfico presentaría una figura sin forma, una nube de puntos dispersos.

El punto $(\bar{x},\bar{y})$ representa siempre el centro de gravedad de la nube de puntos

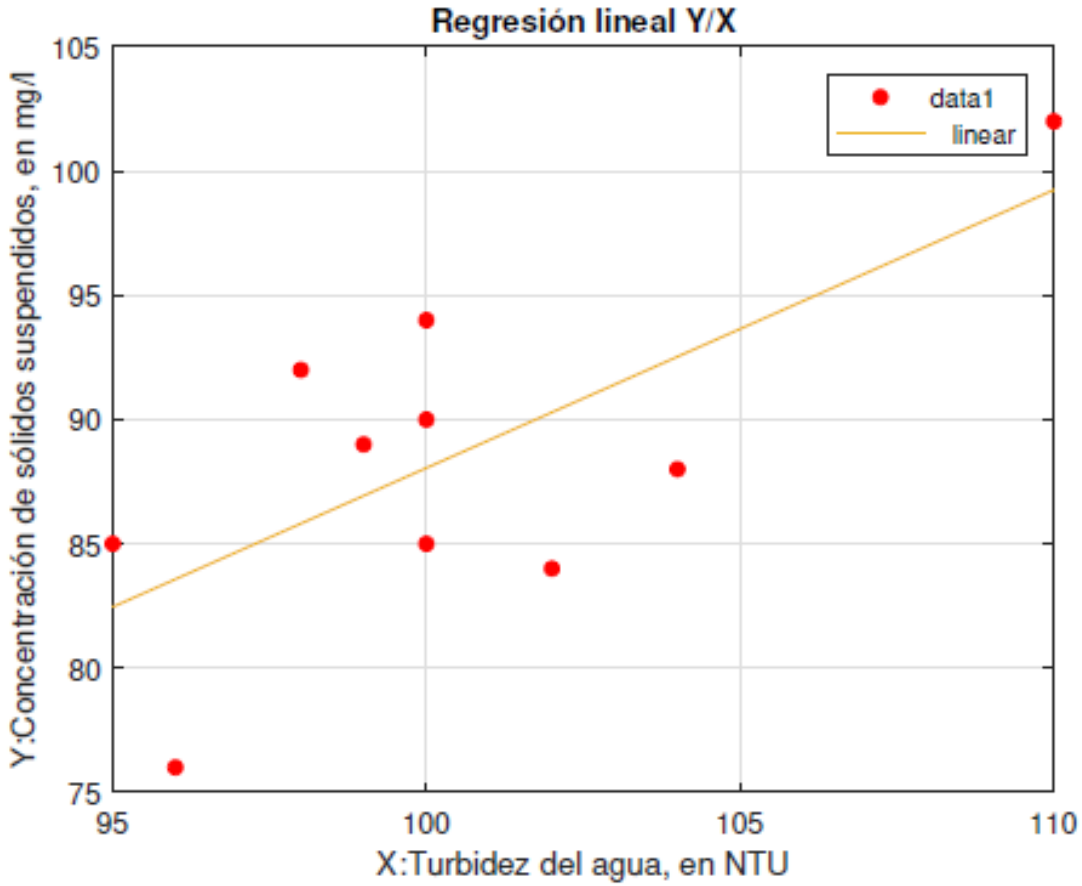
Ejemplo **Análisis de la calidad de las aguas de un embalse.**

En un análisis de las aguas de cierto embalse, se obtienen los siguientes valores de *concentración de sólidos suspendidos* (medida en mg/l) y de *turbidez* del agua (medida en Unidades Nefelométricas de Turbidez, o Nephelometric Turbidity Unit (NTU))

turbidez (NTU)	95	100	102	104	100	98	96	100	110	99
sólidos suspendidos (mg/l)	85	94	84	88	85	92	76	90	102	89

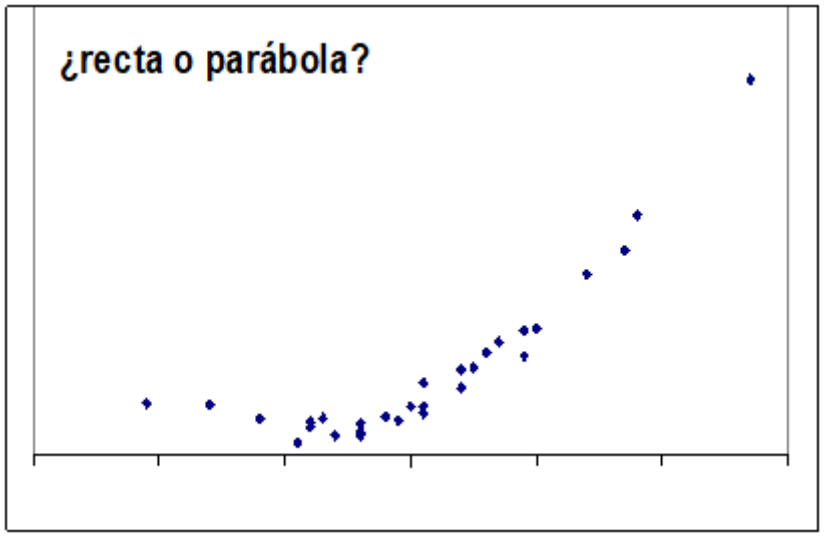
Diagramas de dispersión o nube de puntos y relación entre variables

Diagrama de Dispersión correspondiente al ejemplo anterior. Al aumentar la concentración de sólidos suspendidos en el agua, parece que aumenta la turbidez de la misma

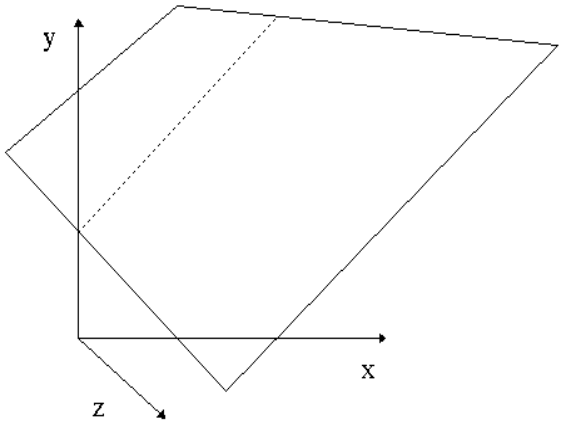


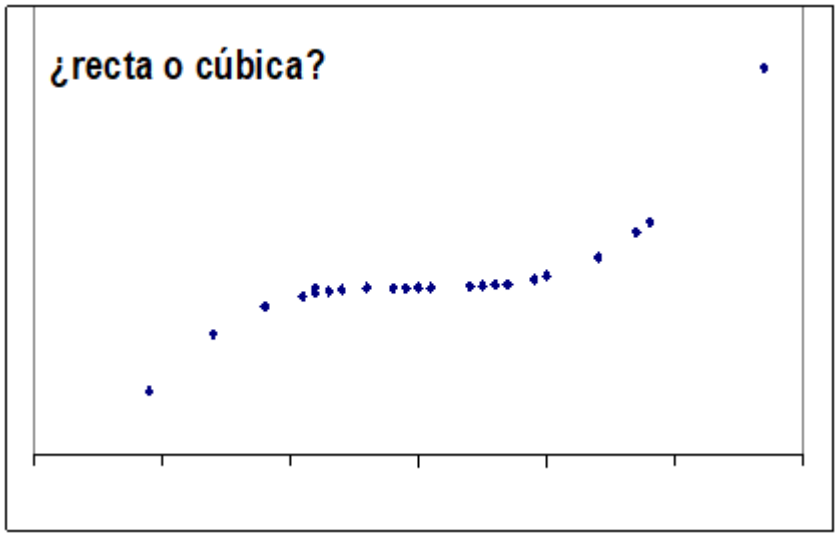
Relación entre variables

- Se pueden considerar otros tipos de modelos, en función del aspecto que presente el diagrama de dispersión (regresión no lineal)



- Incluso se puede considerar el que una variable dependa de varias (regresión múltiple).





Covarianza de dos variables X e Y

- Para saber que tipo de relación hay entre las dos variables se emplea la covarianza S_{xy}.
- Muestra el grado de variación conjunta de cada una de las variables a su media

$$S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Operando y simplificando queda:

$$S_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x}\bar{y}$$

Regresión lineal

- Recta de regresión de y/x

$$\boldsymbol{y = bx + a}$$

Es la recta que hace mínimo el error cuadrático medio:

$$\mathrm{ECM_y} = \frac{1}{N}\sum_{i=1}^{N}(y_i - bx_i - a)^2$$

Debemos calcular los coeficientes a y b que hacen que el error cuadrático medio sea mínimo. Es decir, la suma de las distancias al cuadrado entre los puntos reales (los que nos proporciona la distribución) y los teóricos o calculados a través de la recta de regresión.

Regresión lineal

$$\frac{d\mathrm{ECM_y}}{db} = \frac{2}{N}\sum_{i=1}^{N}(-x_i)(y_i - bx_i - a) = 0$$

$$\frac{d\mathrm{ECM_y}}{da} = \frac{2}{N}\sum_{i=1}^{N}(-1)\,(y_i - bx_i - a) = 0$$

Derivamos las dos expresiones y llegamos al siguiente sistema de ecuaciones:

$$b\sum x_i{}^2+\mathrm{a}\sum x_i=\sum x_i\,y_i$$

$$b\sum x_i+a\,N=\sum y_i$$

Operando y simplificando llegaríamos a la ecuación de la recta en la forma punto pendiente:

$$y-\bar{y}=\frac{s_{xy}}{s_x{}^2}(x-\bar{x})$$

Por tanto los coeficientes a y b valdrían:

$$b=\frac{s_{xy}}{s_x{}^2};\qquad a=\bar{y}-b\bar{x}$$

- A la pendiente de la recta se le llama coeficiente de regresión de la recta y/x y es igual al valor del coeficiente b, es decir:

$$b = \frac{S_{xy}}{{S_x}^2}$$

El signo de la pendiente de la recta, coincide con el signo de la covarianza

Regresión lineal

- De la misma manera, con la misma nube de puntos podríamos calcular la recta de regresión de x/y

$$x = b_{xy}y + a_{xy}$$

- En este caso predice el valor de x como función de y
- Se procede igual que para la recta de y/x, de todas las rectas que componen la nube de puntos se coge aquella que hace mínimo el error cuadrático medio:

$$\mathrm{ECM}_x = \frac{1}{N}\sum_{i=1}^{N}(x_i - by_i - a)^2$$

Hace que los residuos con respecto de x sean mínimos.

$$\frac{d\mathrm{ECM}_x}{db} = 0$$

$$\frac{d\mathrm{ECM}_x}{da} = 0$$

Procediendo al igual que en la anterior recta calculada la ecuación final queda:

$$\mathrm{x} - \bar{\mathrm{x}} = \frac{s_{xy}}{{s_y}^2}(\mathrm{y} - \bar{\mathrm{y}})$$

El coeficiente de regresión de x/y es:

$$b_{\mathrm{xy}} = \frac{s_{xy}}{{s_{\mathrm{y}}}^2}$$

La pendiente de la recta es: b_{xy}

CONCLUSIONES

- Las pendientes de las dos rectas $^{x}/_{y}$ e $^{y}/_{x}$ tienen el mismo signo (el de la covarianza)
- Las dos rectas pasan por el punto $(\bar{x},\ \bar{y})$ y se cortan en ese punto

Análisis del ajuste

- Se realiza por:
  - Coeficiente de correlación lineal de Pearson *r*
  - Estudio de las varianzas residuales

Coeficiente de correlación lineal

- Indica si los puntos presentan una tendencia a alinearse. Cuanto más alineados están los puntos de la nube entre ellos mejor es el ajuste y más parecidas son las dos rectas que podemos obtener de un diagrama de dispersión: $^{x}/_{y}$ e $^{y}/_{x}$

$$r^2 = \frac{b_{yx}}{\frac{1}{b_{xy}}} \rightarrow b_{yx} b_{xy} \leq 1$$

$$\boxed{r = \frac{S_{xy}}{S_x S_y}}$$

Coeficiente de correlación lineal

$$r = \frac{S_{xy}}{S_x S_y}$$

Dónde:

- r = coeficiente de correlación lineal
- S_{xy}=Covarianza de las variables X, Y
- S_x= Desviación típica de X
- S_y= Desviación típica de Y

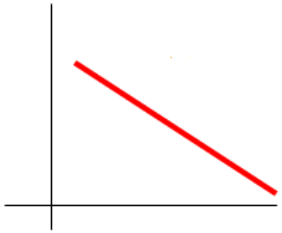
Toma el signo de la covarianza y su valor oscila entre -1 y 1

A r^2 se le denomina coeficiente de determinación

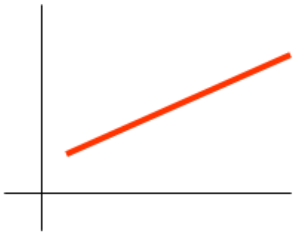
Toma valores entre 0 y 1

Correlación entre las variables

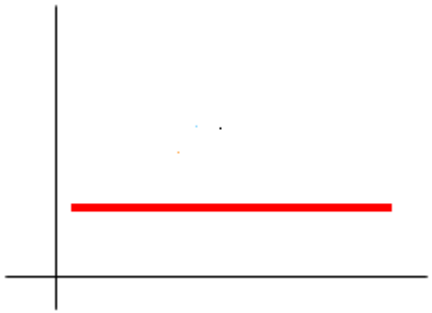
- Si $r = -1 \rightarrow$ Ajuste perfecto $^{x}/_{y} = {}^{y}/_{x}$ y pendiente negativa



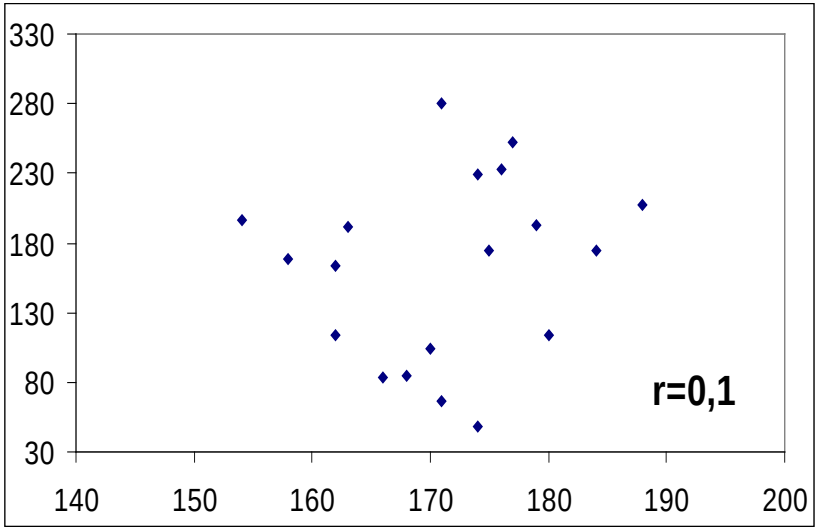
- Si $r = 1 \rightarrow$ Ajuste perfecto $^{x}/_{y} = {}^{y}/_{x}$ y pendiente positiva

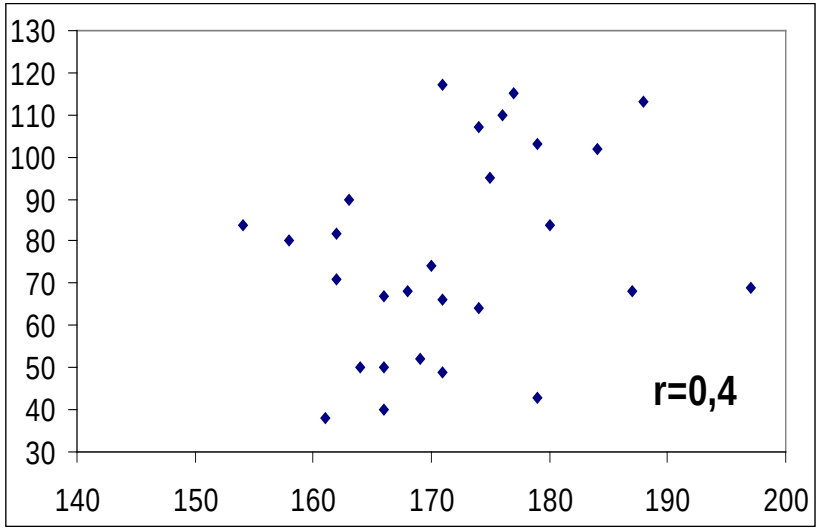


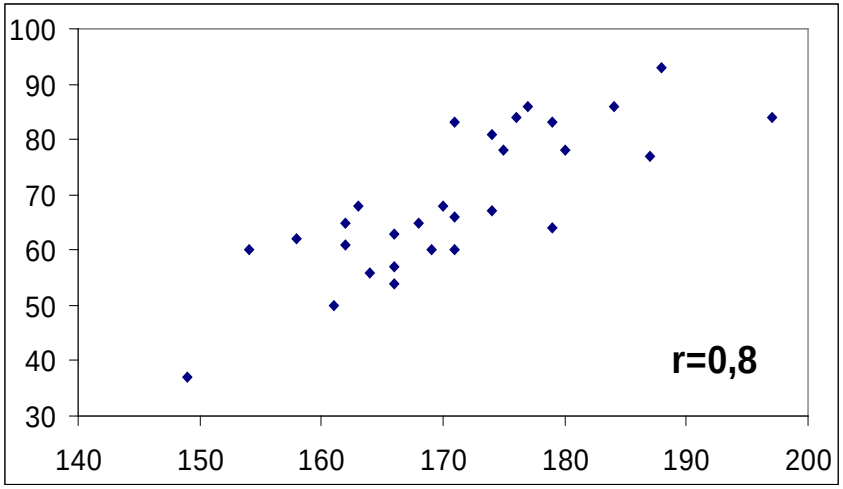
- Si $r = 0 \rightarrow$ No existe correlación entre las variables
- Las rectas $^{x}/_{y}$ e $^{y}/_{x}$ son perpendiculares

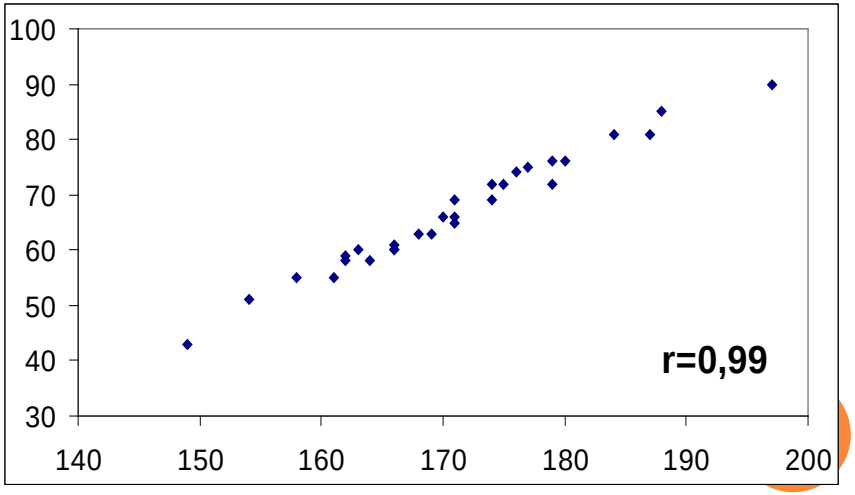


Ejemplos de coeficientes de correlación

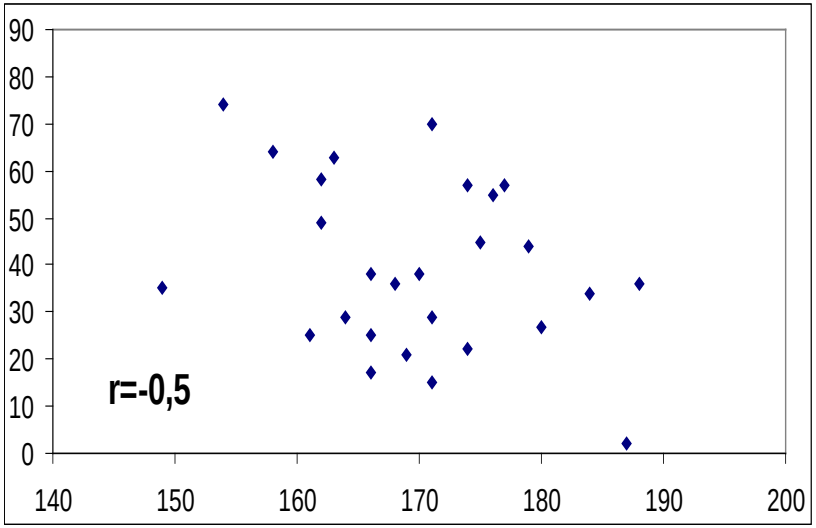


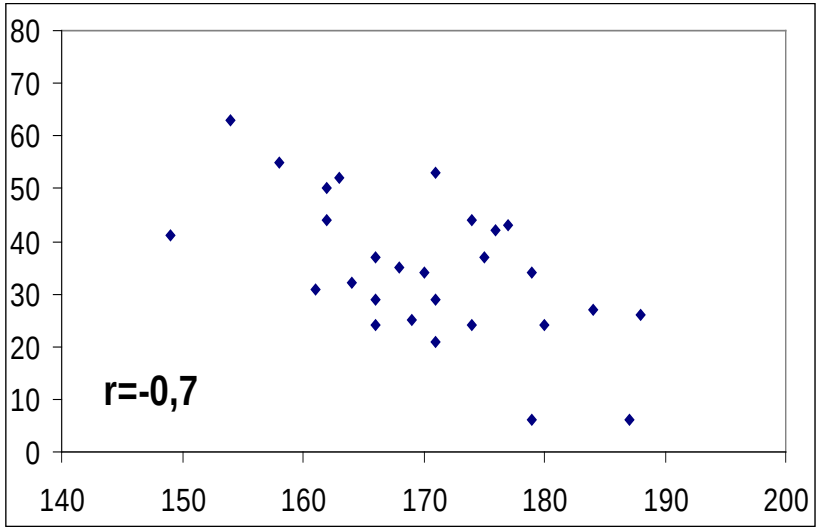


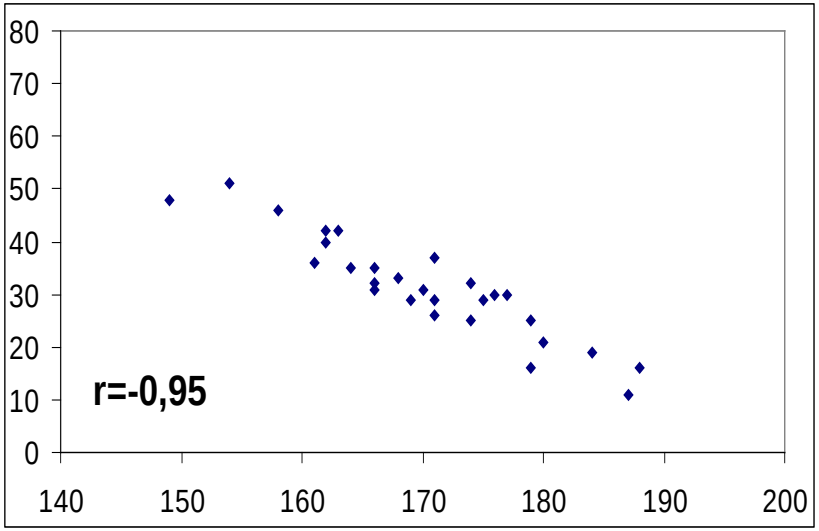


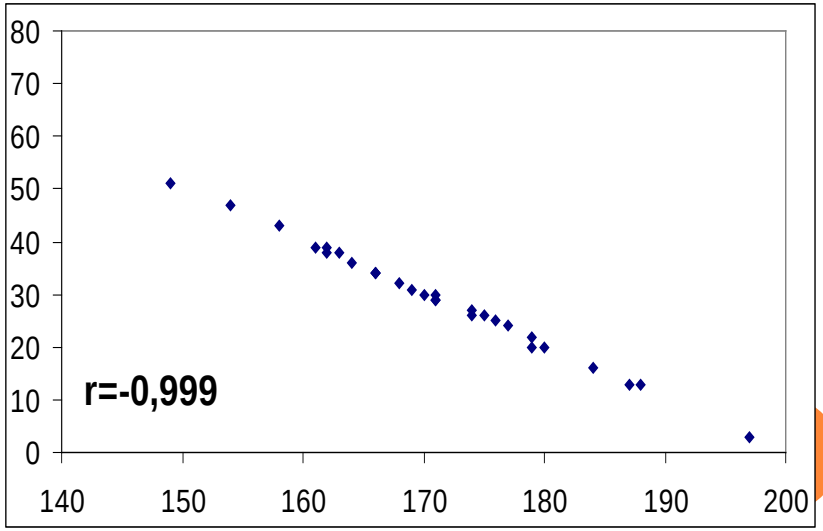


Ejemplos de coeficientes de correlación









Estudio de la varianza

- La varianza total queda definida como $s_y{}^2$

$$s_y{}^2 = s_y{}^2(1 - r^2) + r^2 s_y{}^2$$

Donde:

$s_y{}^2(1 - r^2)$ Varianza residual o no explicada

$r^2 s_y{}^2$, representa la varianza explicada por el Modelo.

Si el ajuste es perfecto la varianza residual es 0

OBJETIVO DE LAS RECTAS DE REGRESIÓN

El objetivo de las rectas de regresión, es predecir el valor de la variable dependiente, en función de la variable independiente. Si hablamos de la recta de regresión $^{y}/_{x}$, la predicción de Y para $X = x_0$ será simplemente el valor obtenido en la recta de regresión de $^{y}/_{x}$ al sustituir el valor de x por x_0. La fiabilidad de esta predicción será tanto mayor cuando mayor sea la correlación entre las variables (es decir, cuanto mayor sea r)

Calcular la recta y/x

x	y
6	6,5
4	4,5
8	7
5	5
3,2	4

x^2	y^2	$x_i y_i$
36	42,25	39
16	20,25	18
64	49	56
25	25	25
12,25	16	14
153,25	**152,5**	**152**

$$y - \bar{y} = \frac{S_{xy}}{S_x{}^2}(x - \bar{x})$$

$$S_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x}\bar{y}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x} = \frac{26,2}{5} = 5,3$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\bar{y} = \frac{27}{5} = 5,4$$

$$S_x{}^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2$$

$$S_{xy} = \frac{152}{5} - 5,3 * 5,4 = 1,78$$

$$S_x{}^2 = \frac{153,25}{5} - 5,3^2 = 2,56$$

$$\frac{S_{xy}}{S_x{}^2} \rightarrow \frac{1,78}{2,56} = 0,7$$

$$y - 5,4 = 0,7(x - 5,3) \rightarrow y = 0,7x + 1,69$$

Otros ajustes de regresión

Regresión exponencial

$$y = ab^x$$

Linealizamos, para ello tomamos logaritmos:

$$\log y = \log a + x \log b$$

- $\log y = V$
- $\log a = A$
- $\log b = B$

$$V = A + Bx$$

Los coeficientes A y B se calculan:

- $B = \frac{S_{xv}}{{s_x}^2}$
- $A = \bar{V} - \mathrm{B}\bar{x}$

Para calcular la regresión exponencial deshacemos el cambio:

- $a = e^A$
- $b = e^B$

Ejemplo regresión exponencial

x	y
1	1,25
2	5
3	11,25
4	20
5	30,5
15	68

$$y = ab^x$$

Linealizamos:

$$logy = loga + xlogb$$

$$logy = V$$
$$loga = A$$
$$logb = B$$

Luego tenemos transformada nuestra exponencial en la ecuación de una recta:

$$V = A + Bx$$

$$V - \bar{V} = \frac{S_{xv}}{{s_x}^2}(x - \bar{x})$$

x	y	V= lny	x^2	vx
1	1,25	0,2231	1	0,2231
2	5	1,6094	4	3,2188
3	11,25	2,4203	9	7,2609
4	20	2,9957	16	11,9828
5	30,5	3,4177	25	17,088
Σ15	Σ68	Σ10,6662	Σ55	Σ39,7741

$$S_{xv} = \frac{\sum xv}{N} - \bar{x}\,\bar{v}$$

$$\bar{x} = \frac{15}{5} = 3$$

$$\bar{V} = \frac{10{,}666}{5} = 2{,}1332$$

$$S_{xv} = \frac{39{,}774}{5} - 3.2{,}1332 = 1{,}5552$$

$$s_x{}^2 = \frac{55}{5} - 3^2 = 2$$

$$V - 2{,}1332 = \frac{1{,}5552}{2}(x - 3)$$

$$V = -0{,}1996 + 0{,}7776x$$

- Deshacemos los cambios y tenemos:

$$e^{A} = a$$
$$e^{-0,1996} = 0,819$$
$$e^{B} = b$$
$$e^{0,7776} = 2,176$$

- Luego la regresión exponencial buscada es:

$$y = 0,819 \cdot 2,176^{x}$$

Bondad del ajuste:

x	y	ŷ	e=ŷ-y	e2
1	1,25	1,7794	0,529	0,2798
2	5	3,86	-1,138	1,295
3	11,25	8,37	-2,88	8,2944
4	20	18,18	-1,82	3,3124
5	30,5	39,45	8,95	80,102
Σ15	Σ68	Σ71,6394	Σ3,641	Σ93,2836

$$ECM = \frac{\sum e^2}{N}$$

$$\hat{y} = 0{,}819 \cdot 2{,}176^{x}$$

$$\hat{y}_1 = 0{,}819 \cdot 2{,}176^{1} = 1{,}7794$$

$$ECM = \frac{93{,}283}{5} = 18{,}656$$

Regresión parabólica

$$Y = a + bx + cx^2$$

$$ECM = \frac{1}{N}\sum\left(y_i - a - bx_i - cx_i^2\right)^2$$

- Derivamos con respecto a los parámetros a, b y c igualamos a 0 y nos queda:

$$\frac{dECM}{da} = \frac{2}{N}\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2(-1) = 0$$

$$\frac{dECM}{db} = \frac{2}{N}\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2(-x_i) = 0$$

$$\frac{dECM}{dc} = \frac{2}{N}\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2\left(-x_i^2\right) = 0$$

Regresión parabólica

- Así llegamos a un sistema de tres ecuaciones con tres incógnitas en el que tendremos que calcular el valor de los parámetros a, b y c.

$$\sum_{i=1}^{N} y_i = Na + b\sum_{i=1}^{N} x_i + c\sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i\, y_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i^3$$

$$\sum_{i=1}^{N} x_i^2 y_i = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i^3 + c\sum_{i=1}^{N} x_i^4$$

Ejemplo de regresión parabólica

x	y	x^2	x^3	x^4	xy	x^2y
1	1,25	1	1	1	1,25	1,25
2	5	4	8	16	10	20
3	11,25	9	27	81	33,75	101,5
4	20	16	64	256	80	320
5	30,5	25	125	625	152,5	762,5
Σ15	Σ68	Σ55	Σ225	Σ979	Σ277,5	Σ1205,25

$$\sum_{i=1}^{N} y_i = Na + b\sum_{i=1}^{N} x_i + c\sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i\, y_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i^3$$

$$\sum_{i=1}^{N} x_i^2 y_i = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i^3 + c\sum_{i=1}^{N} x_i^4$$

$$68 = 5a + 15b + 55c$$
$$277{,}5 = 15a + 55b + 225c$$
$$1205 = 55a + 225b + 979c$$

Resolvemos el sistema y queda:

$$a = -0{,}47$$
$$b = 0{,}51$$
$$c = 1{,}14$$

$$y = -0{,}47 + 0{,}51x + 1{,}14x^2$$

Bondad del ajuste

x	y	ŷ	e=ŷ-y	e2
1	1,25	1,18	0,07	0,0049
2	5	5,11	-0,11	0,0121
3	11,25	11,32	-0,07	0,0049
4	20	19,81	0,19	0,0361
4	30,5	30,58	-0,08	0,0064
	Σ68	Σ68	Σ0	Σ0,0644

$$ECM = \frac{\sum e^2}{N}$$

$$\hat{y} = -0{,}47 + 0{,}51x + 1{,}14x^2$$

$$\hat{y}_1 = -0{,}47 + 0{,}51 \cdot 1 + 1{,}14 \cdot 1^2 = 1{,}18$$

$$ECM = \frac{0{,}0644}{5} = 0{,}01288$$