

# Regresión lineal Simple

Cuando se estudian en forma conjunta dos características (variables estadísticas) de una población o muestra, se dice que estamos analizando una variable estadística bidimensional.

## ¿En qué consiste el problema de la regresión?

El problema de regresión consiste en hallar la mejor relación funcional entre dos variables  $X$  e  $Y$ .

Concretamente, dada una muestra de dos variables  $X, Y$  con  $(x_i, y_i), i = 1, 2, \dots, n$ , se desea estudiar como depende el valor de  $Y$  en función de  $X$

## Variables en un modelo de regresión

$X$ : variable de control, **exploratoria, predictora**, independiente o de regresión.

$Y$ : variable de **respuesta, resultado**, dependiente.

Si pensamos a manera de ejemplo, en un experimento donde la variable  $X$  es la que **controla el experimentador** y la variable  $Y$  es el **valor que se obtiene del experimento**

## Regresión y gráficos de dispersión

La relación funcional que se pueda dar entre dos variables, puede ser lineal, cuadrática, exponencial, potencial o cualquier otra.

Para identificar de manera exploratoria un modelo de ajuste para los datos, es posible apoyarse en los gráficos de dispersión, así un gráfico de dispersión debería ser la primera herramienta para explorar la relación entre dos variables.

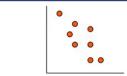
Generalmente se grafica la variable respuesta ( $Y$ ) en el eje vertical y la variable exploratoria en el eje horizontal ( $X$ )

## ¿Qué se busca en un gráfico de dispersión?

- Patrón general



Asociación positiva

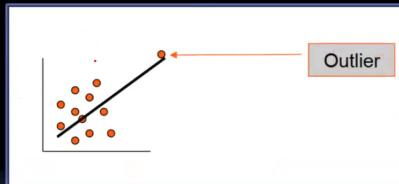


Asociación negativa

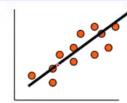


No asociación

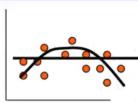
- Observaciones o datos atípicas (outliers)



- Forma de la relación

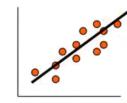


Relación lineal

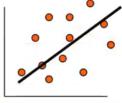


Relación curvilineal

- Intensidad de la relación



Asociación intensa



Asociación débil

## Regresión lineal simple

Sea:  $\rightarrow$  mayúscula

$X$ : Variable exploratoria, independiente

$Y$ : Variable aleatoria de respuesta, dependiente.

$Y|x$ : variable aleatoria  $Y$ , correspondiente a un valor fijo de  $x \in X$

$E(Y|x) = \mu_{Y|x}$ : es el promedio o valor esperado de los  $Y|x$

Si  $\mu_{Y|x}$  es una función lineal de  $x$  tal que:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

Entonces, se dice que existe una **regresión lineal simple** entre  $X$  e  $Y$ .

Como anteriormente se indicaba, para cada valor fijo de  $X$ , le corresponde un conjunto de valores  $Y$  asociados.

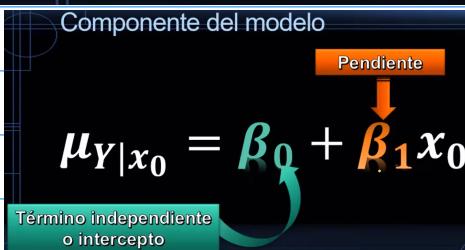
A manera de ejemplificar esta idea, se toma un valor  $x_0 \in X$ .

Para ese valor  $x_0$  suponga que se realiza un experimento

varias veces, obteniendo diferentes valores como resultado:

$Y|x_0 = \{y_{10}, y_{20}, \dots, y_{m0}\}$ , así un valor representativo de este conjunto sería un promedio  $\mu_{Y|x_0} = \frac{y_{10} + y_{20} + \dots + y_{m0}}{m}$ , por lo que el modelo de regresión lineal simple ayuda a predecir dicha media:

$$\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$$



El modelo de regresión lineal simple  $\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$  es un modelo ideal, un modelo teórico que describe a la población completa.

En la práctica, el trabajo que hacemos es estimar este modelo, a partir de la estimación de los parámetros:

$$b_0 \approx \beta_0$$

$$b_1 \approx \beta_1$$

Las estimaciones  $b_0, b_1$  se deben calcular a partir de la muestra  $(x_i, y_i), i = 1, 2, \dots, n$  y una vez halladas, se obtiene la ecuación de la recta de regresión para que estima al modelo teórico:

$$\hat{y} = b_0 + b_1 x$$

Así nuestro modelo quedaría de la siguiente manera:

Con:

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b \bar{x}$$

Se observan las edades y las estaturas de un grupo de niños como se muestra en la siguiente tabla:

X Edad	3,3	4,3	5,2	5,6	6,3	7,0	8,1	8,6	9,3	10,1
Y Estatura (cm)	94	112	115	123	117	117	125	139	134	140

$$\sum h = 10$$

Proponga un modelo de regresión lineal simple para los datos de la tabla anterior y conteste:

En la calculadora  
 menu → 6 → 2 → meter datos → AC →  
 optn → Summation

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b \bar{x}$$

Y ahí calcularemos las sumas

$$\sum xy = 8502,8 \quad \sum x = 67,8 \quad \sum y = 1216 \quad \sum x^2 = 509,54$$

$$n = 10$$

$$b = \frac{10 \cdot 8502,8 - 67,8 \cdot 1216}{10 \cdot 509,54 - 67,8^2} = 5,76$$

$$a = \frac{1216 - 5,76 \cdot 67,8}{10} = 82,55$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b \bar{x}$$

$$\hat{y} = 82,55 + 5,76x \rightarrow \text{Modelo de RLS}$$

$$\nearrow x=5$$

a. ¿Cuál es la estatura esperada para un niño de 5 años?

$$\hat{y} = 82,55 + 5,76 \cdot 5 \approx 111,35 \text{ cm}$$

$$\nearrow \hat{y}=111,35$$

b. ¿Cuál es la edad estimada de un niño que mide 125cm de estatura?

$$125 = 82,55 + 5,76x \rightarrow x = \frac{125 - 82,55}{5,76} \approx 7,37$$

$$\searrow x=7,37$$

c. ¿Cuál sería la estatura esperada para un niño a los 25 años?

$$\hat{y} = 82,55 + 5,76 \cdot 25 \approx 226,55 \text{ cm}$$

En un experimento donde se quería estudiar la asociación entre consumo de sal y la presión arterial, se asignó aleatoriamente a algunos individuos una cantidad diaria constante de sal en su dieta, y al cabo de un mes se les midió la tensión arterial media. Algunos resultados fueron los siguientes:

<i>X(sal en g)</i>	<i>Y (presión en mm de Hg)</i>
1.8	100
2.2	98
3.5	110
4.0	110
4.3	112
5.0	120

Determine la ecuación de la recta de regresión lineal para los datos de la tabla.

$$\hat{y} = a + bx \quad n = 6$$

$$\sum xy = 2302,2 \quad \sum x = 20,8 \quad \sum y = 650 \quad \sum x^2 = 79,82$$

$$b = \frac{6 \cdot 2302,2 - 20,8 \cdot 650}{6 \cdot 79,82 - 20,8^2} = 6,34$$

$$a = \frac{650 - 6,34 \cdot 20,8}{6} = 86,35$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

$$\hat{y} = 86,35 + 6,34x$$

**Ejemplo 88** Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días ( $X$ ) y el ingreso obtenido por su venta en dólares ( $Y$ ). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:

$X :$	2	3	5	9	10	16	19	20	24	27	32	41	55	60
$Y :$	100	200	250	400	500	850	930	900	1300	1360	1500	2050	2800	2900

1. Encuentre la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software

$$n = 14$$

$$\sum xy = 18780 \quad \sum x = 323 \quad \sum y = 16090 \quad \sum x^2 = 11871$$

$$b = \frac{14 \cdot 18780 - 323 \cdot 16090}{14 \cdot 11871 - 323^2} = 79,29$$

$$a = \frac{16090 - 79,29 \cdot 323}{14} = 8,52$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

$$\hat{y} = 8,52 + 79,29x$$

de a, lo da a

2. ¿Aproximadamente, al elaborar un software, cuánto es el valor promedio del ingreso fijo (ingreso que no depende del tiempo de elaboración)?

$$8,52$$

de b, lo da b

3. ¿Aproximadamente, al elaborar un software, cuánto es el aumento promedio del ingreso por día de elaboración?

$$79,29$$

#### Interpretación de los coeficientes

$a$  : valor promedio esperado cuando  $X$  es cero.

$b$  : razón de cambio del valor promedio de  $Y$  por cada unidad adicional en  $X$ .

Un vendedor de teléfonos celulares quiere poner a prueba cierta marca. Para esto, toma 7 teléfonos del mismo modelo y los carga hasta cierto porcentaje de la batería, y los deja proyectando el mismo video hasta que el celular se apague. Los resultados fueron los siguientes:

% de carga ( $X$ )	2	6	8	10	12	14	16
minutos antes de apagarse ( $Y$ )	1	5	8	11	15	20	25

1. Determine la ecuación de regresión lineal para la cantidad de minutos que dura en apagarse el celular proyectando el video en función del porcentaje de la carga.

$$n = 7$$

$$\sum xy = 1066 \quad \sum x = 68 \quad \sum y = 85 \quad \sum x^2 = 800$$

$$b = \frac{\sum xy - \bar{x} \cdot \bar{y}}{\sum x^2 - \bar{x}^2} = \frac{1066 - 68 \cdot 85}{800 - 68^2} \approx 1,73$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

$$a = \frac{85 - 1,73 \cdot 68}{7} \approx -9,66$$

$$\hat{y} = -9,66 + 1,73x$$

Con la llegada de las lluvias aumenta la cantidad de accidentes en carretera. Durante los primeros días de la estación lluviosa se han registrado las horas de lluvia en algunos sectores del GAM y la cantidad de accidentes reportados. Con la información se construyó la tabla

$y$	Accidentes	32	30	34	26	30	35	43	33
$x$	Horas	3	2	4	2	3	4	5	3

$$32 = 8$$

Utilice el método de mínimos cuadrados para estimar el parámetro  $\beta$  en la ecuación de regresión  $y = 20 + \beta x$ , en la que el número de accidentes reportados depende de las horas de lluvia.

Utilice el método de mínimos cuadrados para estimar el parámetro  $\beta$  en la ecuación de regresión  $y = 20 + \beta x$ , en la que el número de accidentes reportados depende de las horas de lluvia.

Aquí  $y = 20 + \beta x$ , entonces no se usa la calculadora de un solo se usa el método de mínimos cuadrados que consiste en escribir la función con el error, derivarlo, igualar a 0 y despejar

$$f(b) = \sum (y - a - bx)^2 \quad y = 20 + bx \quad n = 8$$

$$f(b) = \sum (y - 20 - bx)^2 \quad y - 20 - bx$$

Derivando en términos de  $b$

$$f'(b) = \varepsilon \cdot 2 \cdot (y - 20 - bx) \cdot \cancel{(y - 20 - bx)}$$

Igualando a 0

$$= -2 \varepsilon (y - 20 - bx) \cdot x = 0$$

Despejando el -2

$$= \varepsilon (y_i - 20 - bx_i) \cdot x_i = 0$$

Repartiendo el  $x_i$

$$= \varepsilon (x_i y - 20x_i - bx_i^2) = 0$$

Repartiendo la suma

$$= \varepsilon x y + \varepsilon -20x + \varepsilon -bx^2 = 0$$

Despejando  $b$

$$b = \frac{\varepsilon x y - 20 \varepsilon x}{\varepsilon x^2}$$

$$b = 9$$

$$\boxed{y = 20 + 9x}$$

Para calcular las  $\varepsilon$   
menu  $\rightarrow$  2  $\rightarrow$   
meter datos  $\rightarrow$  AC  
optn  $\rightarrow$  summation,  
se puede meter toda  
la formula de un  
solo con ( )

Interpretación de los coeficientes

$a$ : valor promedio esperado cuando  $X$  es cero.

$b$ : razón de cambio del valor promedio de  $Y$  por cada unidad adicional en  $X$ .

Ejemplo 89 Considera los datos de la siguiente tabla:

$$\begin{array}{ccccc} X & : & 2 & 3 & 5 \\ Y & : & 8 & 10 & 18 \end{array}$$

A partir de estos datos, estime el coeficiente  $\beta$  de la ecuación de regresión  $\mu_{Y|x} = 2 + \beta x$  utilizando el método de mínimos cuadrados.

$$y = 2 + \beta x \rightarrow y - 2 - \beta x \quad \sum xy = 926 \quad \sum x = 32$$
$$\sum x^2 = 288$$

$$\sum (y - 2 - \beta x)^2$$

$$\sum 2(y - 2 - \beta x) \cdot -x$$

$$-2 \sum (y - 2 - \beta x)x = 0$$

$$\sum (y - 2 - \beta x)x = 0$$

$$\sum (xy - 2x - \beta x^2) = 0$$

$$\sum xy - 2 \sum x - \sum \beta x^2 = 0$$

$$\sum xy - 2 \sum x = \beta \sum x^2$$

$$\beta = \frac{\sum xy - 2 \sum x}{\sum x^2} \quad \sum xy = 926 \quad \sum x = 32$$
$$\sum x^2 = 288$$

$$\beta = \frac{926 - 2 \cdot 32}{288} = 2,99$$

$$y = 2 + 2,99x$$

Ejercicio 44 Considera los datos de la siguiente tabla:

$x :$	1	5	10	20
$y :$	4	12	21	43

A partir de estos datos, estime el coeficiente  $\beta$  de la ecuación de regresión  $\mu_{Y|x} = \beta + \beta x$  utilizando el método de mínimos cuadrados.

$$R/ b = \frac{567}{281}$$

$$y = \beta + \beta x \rightarrow y - \beta - \beta x$$

$$\sum (y - \beta - \beta x)^2$$

$$\sum 2 \cdot (y - \beta - \beta x) \cdot (-1 - x)$$

$$\sum (\sum (-y + \beta + \beta x - xy + \beta x + \beta x^2)) = 0$$

$$-\sum y + \beta \sum 1 + \beta \sum x - \sum xy + \beta \sum x + \beta \sum x^2 = 0$$

$$\beta \sum 1 + \beta \sum x + \beta \sum x + \beta \sum x^2 = \sum xy + \sum y$$

$$\beta (\sum 1 + \sum x + \sum x + \sum x^2) = \sum xy + \sum y$$

$$\beta = \frac{\sum xy + \sum y}{\sum 1 + \sum x + \sum x + \sum x^2}$$

$$\sum xy = 1134$$

$$\sum x = 36$$

$$\sum y = 80$$

$$\beta = \frac{1134 + 80}{4 + 36 + 36 + 526}$$

$$\sum x^2 = 526$$

$$\sum 1 = 4$$

$$\beta = 20,166$$

$$\left\{ \begin{array}{l} k \\ \sum c = C \cdot k \end{array} \right.$$

$$i=1$$

$$y = 20,166 + 20,166x$$

$$\sum 1 - 4 = 4$$

$$i=1$$

**Ejercicio 43** Se espera que, por lo general, el número de horas de estudio ( $X$ ) en la preparación para hacer un examen tenga una correlación directa con la calificación obtenida ( $Y$ ) alcanzada en tal examen. Se obtuvieron las horas de estudio así como las calificaciones (en escala de 0 a 100) obtenidas por diez estudiantes seleccionados al azar de un grupo, los datos están resumidos en la siguiente tabla:

$$n = 10$$

$$\begin{array}{ccccc} \sum x & \sum y & \sum y^2 & \sum x^2 & \sum xy \\ 118 & 591 & 39013 & 1648 & 7956 \end{array}$$

1. Determine la ecuación de mínimos cuadrados para la calificación como función de las horas de estudio.

$$a + b$$

$$R/ \quad b = 3.842723005, a = 13.75586854$$

$$n = 10$$

$$b = \frac{10 \cdot 7956 - 118 \cdot 591}{10 \cdot 1648 - 118^2} = 3.87273005$$

$$10 \cdot 1648 - 118^2$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b \bar{x}$$

$$a = \frac{591 - 3.87273005 \cdot 118}{10} = 13.75586854$$

$$y = 13.75586854 + 3.87273005 X$$

2. ¿Aproximadamente, cuál es la calificación promedio de los estudiantes que no estudiaron para el examen?

$$R/ \quad a = 13.75586854$$

3. ¿Cuánto es el aumento promedio aproximado en la calificación por hora de estudio?  $R/ \quad b = 3.842723005$

### Interpretación de los coeficientes

$a$  : valor promedio esperado cuando  $X$  es cero.

$b$  : razón de cambio del valor promedio de  $Y$  por cada unidad adicional en  $X$ .

### Intervalos de confianza para $\alpha$ y $\beta$ cuando se cumplen las hipótesis de regresión

IC para  $\alpha$ :  $a \pm t_{\delta/2,\nu}s \sqrt{\frac{\sum x^2}{nS_{xx}}}$  con  $\nu = n - 2$

IC para  $\beta$ :  $b \pm t_{\delta/2,\nu}s \sqrt{\frac{1}{S_{xx}}}$  con  $\nu = n - 2$

### Algunas fórmulas útiles

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

### Independencia entre el predictor y el resultado

Una manera de determinar si la regresión lineal es significativa, es por medio de una prueba de hipótesis:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Note que la hipótesis nula ( $H_0$ ) plantea que no hay relación lineal entre las variables  $X$  e  $Y$ , mientras que la hipótesis alternativa ( $H_1$ ) defiende su existencia.

Ahora, el contraste no siempre requiere comparar  $\beta_1 = 0$ , podría ser de interés para el investigador hacer dicha comparación con otro valor para  $\beta_1$ . Seguidamente se presenta una prueba más general.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

$$s = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{n-2}}$$

### Prueba de hipótesis para una regresión lineal

Considere para la hipótesis nula:

$$H_0: \beta_1 = k \text{ (dependiendo de } H_1: \leq, \geq)$$

Opciones para la hipótesis alternativa:

$$H_1: \beta_1 > k$$

$$H_1: \beta_1 < k$$

$$H_1: \beta_1 \neq k$$

Estadístico de contraste o prueba:

$$T = \frac{b_1 - k}{\frac{s}{\sqrt{S_{xx}}}} = (b_1 - k) \frac{\sqrt{S_{xx}}}{s}$$

Los valores críticos y P-Valores se obtienen de la forma usual.

### Intervalo de confianza para $\mu_{Y|x}$ cuando se cumplen las hipótesis de regresión

Extremos del intervalo:  $\hat{y} \pm t_{\delta/2,\nu}s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$   
con  $\nu = n - 2$  y  $\hat{y} = a + bx$

### Intervalo de predicción para $Y|x$ cuando se cumplen las hipótesis de regresión

Extremos del intervalo:  $\hat{y} \pm t_{\delta/2,\nu}s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$   
con  $\nu = n - 2$  y  $\hat{y} = a + bx$

**Ejemplo 90** Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días ( $X$ ) y el ingreso obtenido por su venta en dólares ( $Y$ ). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:

$$\sum x = 323 \quad \sum y = 16040 \quad \sum x^2 = 11871$$

$$\sum y^2 = 29162000 \quad \sum xy = 587890$$

Asuma las hipótesis de regresión. En un ejemplo anterior se determinó que la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software es

$$\hat{y} = 49.2935x + 8.44282$$

1. Determine un intervalo de confianza del 90% para el valor promedio del ingreso fijo (ingreso que no depende del tiempo de elaboración)  $\rightarrow Y = \alpha$

$$a = 8.44282$$

$$a \pm t_{\delta/2, \nu} s \sqrt{\frac{\sum x^2}{n S_{xx}}} \text{ con } \nu = n - 2$$

$$b = 49.2935$$

$$b \pm t_{\delta/2, \nu} s \sqrt{\frac{1}{S_{xx}}} \text{ con } \nu = n - 2$$

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \sqrt{\frac{S_{yy} - b S_{xy}}{n-2}}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = (n-1)s_x^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = (n-1)s_y^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$\sum x = 323 \quad \sum y = 16040 \quad \sum xy = 587890 \quad \sum x^2 = 11871 \quad \sum y^2 = 29162000$$

$$h = 17 \quad v = 12 \quad \omega = 0.10 \rightarrow \frac{\omega}{2} = 0.05$$

$$t_{0.05, 12} = \pm 1.78229$$

$$S_{xx} = 11871 - \frac{323^2}{17} = 9918.928571$$

$$S_{yy} = 29162000 - \frac{(16040)^2}{17} = 10787742.86$$

$$S_{xy} = 587890 - \frac{323 \cdot 16040}{17} = 217827.2857$$

$$S = \sqrt{10787742.86 - 49.2935 \cdot 217827.2857}$$

$$S = 62,86323306$$

$$a \pm t_{\delta/2, \nu} s \sqrt{\frac{\sum x^2}{n S_{xx}}} \text{ con } \nu = n - 2$$

$$a = 8,99282 \quad b = 74$$

$$S = 62,86323306 \quad v = 72$$

$$\sum x^2 = 11871 \quad t_{0.05, 72} = \pm 1.78229$$

$$S_{xx} = 4918,928571$$

Límite inferior

$$8,99282 - 1,78229 \cdot 62,86323306, \boxed{\begin{array}{c} 11871 \\ 74,9918,928571 \end{array}}$$

$$\liminf = -70,636239$$

Límite superior

$$8,99282 + 1,78229 \cdot 62,86323306, \boxed{\begin{array}{c} 11871 \\ 74,9918,928571 \end{array}}$$

$$\limsup = 57,521879$$

El IC para 90% del valor promedio  
del ingreso fijo es

$$[-70,636239, 57,521879]$$

2. Determine un intervalo de confianza del 90% para el aumento promedio del ingreso por día de elaboración

6

$$\text{IC para } \beta: b \pm t_{\delta/2, \nu} s \sqrt{\frac{1}{S_{xx}}} \text{ con } \nu = n - 2$$

$$b = 49,2935$$

$$t_{0.05, 12} = \pm 1.78229$$

$$n = 14$$

$$S = 62,86323306$$

$$v = 12$$

$$S_{xx} = 4938,928571$$

Límite inferior

$$49,2935 - 1.78229 \cdot 62,86323306,$$

$$\boxed{47,6080717}$$

$$\text{Lím inf} = 47,6080717$$

Límite superior

$$49,2935 + 1.78229 \cdot 62,86323306,$$

$$\boxed{50,9789286}$$

$$\text{Lím sup} = 50,9789286$$

∴ El IC para 90% del valor promedio

del ingreso por día es

$$[47,6080717, 50,9789286]$$

Intervalo de confianza para  $\mu_{Y|x}$  cuando se cumplen las hipótesis de regresión

$$\text{Extremos del intervalo: } \hat{y} \pm t_{\delta/2, \nu} s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ \text{con } \nu = n - 2 \text{ y } \hat{y} = a + bx$$

*Promedio*

Determine el IC de 90% para el ingreso esperado por venta de software con un tiempo de elaboración de 30 días

$$\hat{y} = 49.2935x + 8.44282$$

*x = 30*

Intervalo de confianza para  $\mu_{Y|x}$  cuando se cumplen las hipótesis de regresión

$$\text{Extremos del intervalo: } \hat{y} \pm t_{\delta/2, \nu} s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ \text{con } \nu = n - 2 \text{ y } \hat{y} = a + bx$$

$$\hat{y} = 49.2935 \cdot 30 + 8.44282$$

$$S = 62,86323306$$

$$\hat{y} = 1987,29782$$

$$S_{xx} = 9918.928571$$

$$h = 14$$

$$x = 30$$

$$n = 12$$

$$\bar{x} = \frac{\sum x}{h} = \frac{323}{14} = 23.07$$

$$t_{0.05, 12} = \pm 1.78229$$

Límite inferior

$$1987,29782 - 1.78229 \cdot 62,86323306, \sqrt{\frac{1}{14} + \frac{(30 - 23,07)^2}{9918.928571}}$$

$$\text{Lim inf} = 1955,106337$$

Límite superior

$$1987,29782 + 1.78229 \cdot 62,86323306, \sqrt{\frac{1}{14} + \frac{(30 - 23,07)^2}{9918.928571}}$$

$$\text{Lim sup} = 2519,389306$$

{ El IC para 90% del valor promedio del ingreso esperado con elab de 30 días es

$$1955,106337, 2519,389306 [$$

Intervalo de predicción para  $Y|x$  cuando se cumplen las hipótesis de regresión

$$\text{Extremos del intervalo: } \hat{y} \pm t_{\delta/2, \nu} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ \text{con } \nu = n - 2 \text{ y } \hat{y} = a + bx$$

Determine un intervalo de predicción de 90% para el ingreso por venta de software con un tiempo de elaboración de 30 días

$$\hat{y} = 49.2935x + 8.44282$$

$$\hookrightarrow x = 30$$

Intervalo de predicción para  $Y|x$  cuando se cumplen las hipótesis de regresión

$$\text{Extremos del intervalo: } \hat{y} \pm t_{\delta/2, \nu} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ \text{con } \nu = n - 2 \text{ y } \hat{y} = a + bx$$

$$\hat{y} = 49.2935 \cdot 30 + 8.44282$$

$$S = 62,86323306$$

$$\hat{y} = 1787,29782$$

$$S_{xx} = 4918.928571$$

$$h = 14$$

$$x = 30$$

$$n = 12$$

$$\bar{x} = \underline{\Sigma x} = \frac{\sum x = 323}{h} = 23,07$$

$$t_{0.05, 12} = \pm 1.78229$$

$$14$$

Límite inferior

$$1787,29782 - 1.78229 \cdot 62,86323306, \sqrt{1 + \frac{1}{14} + \frac{(30 - 23,07)^2}{4918.928571}}$$

$$\text{Lim inf} = 1370,688773$$

Límite superior

$$1787,29782 + 1.78229 \cdot 62,86323306, \sqrt{1 + \frac{1}{14} + \frac{(30 - 23,07)^2}{4918.928571}}$$

$$\text{Lim sup} = 1603.807467$$

El IP de 90% para el ingreso por venta con elaboración de 30 días es

$$[1370.688773, 1603.807467]$$

Con un nivel de significancia del 10%, ¿Hay evidencia de que el aumento promedio del ingreso por día de elaboración sea de 50 dólares por día?

$$H_0: \beta = 50$$

$$H_1: \beta \neq 50$$

Para probar  $H_0: \beta = \beta_0$  cuando se cumplen las hipótesis de regresión

$$\text{Estadístico de prueba: } T = (B - \beta_0) \frac{\sqrt{S_{xx}}}{S} \text{ con } \nu = n - 2$$

$$\hat{y} = 49,2935x + 8,44282$$

$$\beta = 49,2935$$

$$\beta_0 = 50$$

$$n = 14$$

$$v = 12$$

$$S = 62,86323306$$

$$S_{xx} = 9418,928571$$

$$\alpha = 0,10 \rightarrow \begin{cases} \text{haz que dividir} \\ \frac{\alpha}{2} = 0,05 \end{cases} \text{ por ser 2 colas}$$

$$T_{0.05} = (49,2935 - 50) \cdot \sqrt{\frac{62,86323306}{9418,928571}} = -0,790708$$

$$T_C = t_{0.05} \cdot S_{xx} / v = \pm 1,78229$$

B) Como

$$t_{C1} = -1,78229 < t_{0.05} = -0,790708 \quad (t_{C2} = 1,78229)$$

No se rechaza  $H_0$

Con valor p

$$2P(|t| > -0,790708) = 0,97367 \quad \alpha = 0,10$$

B) Como  $p = 0,97367 > \alpha = 0,10$

No se rechaza  $H_0$

## Ejemplo

Una cooperativa de ahorro ha sufrido un cierre técnico debido a una deficiente administración. Esta entidad financiera estima que el tiempo (en días) para la devolución de dinero depende linealmente de la cantidad de dinero (millones de colones) que el cliente tiene en su cuenta. Los siguientes datos se refieren a información de 10 clientes sobre el dinero en su cuenta ( $x$ ) y el tiempo que han esperado para su devolución ( $y$ ).

$$\sum_{i=1}^{10} x_i^2 = 433$$

$$\sum_{i=1}^{10} x_i = 51$$

$$\sum_{i=1}^{10} y_i^2 = 1545$$

$$\sum_{i=1}^{10} y_i = 95$$

$$\sum_{i=1}^{10} x_i y_i = 808$$

- a) Según la recta de regresión lineal, cuántos días más, en promedio, debe esperar un cliente por la devolución de su dinero por cada millón de colones en su cuenta? (4 puntos)

$$10 \cdot 808 - 51 \cdot 95 \approx 1,871$$

$$b$$

$$10 \cdot 433 - 51^2$$

$$n = 10$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

Por cada millón debe esperar en promedio  
1,871 días

$$S_{xx} = 433 - \frac{51^2}{10} = 172,9$$

$$S_{yy} = 1545 - \frac{95^2}{10} = 692,5$$

$$S_{xy} = 808 - \frac{51 \cdot 95}{10} = 323,5$$

c) Construya un intervalo de predicción del 95% para el tiempo que debe esperar un cliente que tiene 5 millones de colones en su cuenta. (5 puntos)

$$b = 1,871 \text{ (incluso a)}$$

$$a = 95 - 1,871 \cdot 51 = -0,0921 \\ 10$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

$$y = -0,0921 + 1,871x \quad x = 5$$

$$y = -0,0921 + 1,871 \cdot 5 = 9,3729$$

Intervalo de predicción para  $Y|x$  cuando se cumplen las hipótesis de regresión

$$\text{Extremos del intervalo: } \hat{y} \pm t_{\delta/2, \nu} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \quad s = \sqrt{\frac{S_{yy} - b S_{xy}}{n-2}}$$

con  $\nu = n - 2$  y  $\hat{y} = a + bx$

$$s = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{n-2}} = \sqrt{\frac{692,5 - 1,871 \cdot 323,5}{10-2}} = 2,157$$

$$X = 5 \quad \bar{x} = \frac{\sum x}{n} = 5,1 \quad \alpha = 0,05 \quad n = 10 \\ \frac{\alpha}{2} = 0,025 \quad v = 8$$

$$t_{0,025, 8} = 2,306$$

$$S_{xx} = 933 - \frac{51^2}{10} = 172,9$$

$$a = 9,3729 - 2,306 \cdot 2,157 \cdot \sqrt{1 + \frac{1}{10} + \frac{(5 - 5,1)^2}{172,9}}$$

$$b = 9,3729 + 2,306 \cdot 2,157 \cdot \sqrt{1 + \frac{1}{10} + \frac{(5 - 5,1)^2}{172,9}}$$

[4.09575, 14.53005]

# Correlación y determinación

## Interpretación de los coeficientes

$\bar{x}$ : valor promedio esperado cuando  $X$  es cero.

$b$ : razón de cambio del valor promedio de  $Y$  por cada unidad adicional en  $X$ .

## Algunas fórmulas útiles

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$r^2$  = Coeficiente de determinación

$r^2$  = Proporción de Variación que depende de  $X$

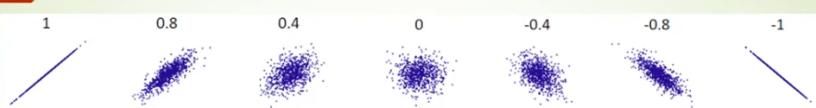
$1 - r^2$  = Proporción de Variación que NO depende de  $X$

$$r^2 = b \frac{s_x}{s_y} = \left( b \sqrt{\frac{S_{xx}}{S_{yy}}} \right)^2$$

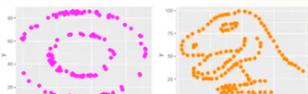
$r$  = Coeficiente de Correlación

$$r = b \frac{s_x}{s_y} = b \sqrt{\frac{S_{xx}}{S_{yy}}}$$

## Correlación y Determinación



Si  $r \approx 0$  no hay correlación lineal.  $-0.2 \leq r \leq 0.2$



Si  $r \approx \pm 1$  hay una buena correlación lineal.

Si  $r \approx 1$  hay correlación lineal positiva.  $r \geq 0.8$

Si  $r \approx -1$  hay correlación lineal negativa.  $r \leq -0.8$

Definición: coeficiente de correlación muestral

$$r = b \frac{s_x}{s_y} = b \sqrt{\frac{S_{xx}}{S_{yy}}}$$

## Interpretación de $r$

- El valor  $R = 1$ : indica correlación positiva perfecta.
- El valor  $R = -1$ : indica correlación negativa perfecta.
- Si  $R > 0$ , indica que la correlación es positiva, donde si  $X$  crece, también crece  $Y$
- Si  $R < 0$ , entonces, la correlación es negativa. En este caso, si  $X$  crece, decrece  $Y$ .

## Ejemplo

Una cooperativa de ahorro ha sufrido un cierre técnico debido a una deficiente administración. Esta entidad financiera estima que el tiempo (en días) para la devolución de dinero depende linealmente de la cantidad de dinero (millones de colones) que el cliente tiene en su cuenta. Los siguientes datos se refieren a información de 10 clientes sobre el dinero en su cuenta ( $x$ ) y el tiempo que han esperado para su devolución ( $y$ ).

$$\sum_{i=1}^{10} x_i^2 = 433$$

$$\sum_{i=1}^{10} x_i = 51$$

$$\sum_{i=1}^{10} y_i^2 = 1545$$

$$\sum_{i=1}^{10} y_i = 95$$

$$\sum_{i=1}^{10} x_i y_i = 808$$

b) ~~¿Qué porcentaje de la variación en los días de espera para la devolución del dinero depende de factores diferentes a la cantidad de dinero en la cuenta del cliente? (3 puntos)~~

No dependiendo de  $x$ , usar  $1 - r^2$

Algunas fórmulas útiles

$$S_{xx} = 433 - \frac{51^2}{10} = 172,9$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$S_{yy} = 1545 - \frac{95^2}{10} = 692,5$$

$$r^2 = b \frac{s_x}{s_y} = \left( b \sqrt{\frac{S_{xx}}{S_{yy}}} \right)^2$$

$$1 - r^2 = 1 - \frac{323^2}{172,9 \cdot 692,5} = 0,05793592032 \approx 0,6$$

Alrededor del 6% de variación

- Una fábrica recolectó la siguiente información de 8 de sus trabajadores sobre el número de minutos que llegaron tarde al trabajo (X) y el número de piezas defectuosas que fabricaron ese día (Y):

X	2	5	10	15	20	25	30	40	$n = 8$
Y	2	4	9	12	15	20	24	30	

Suponiendo que se cumplen las hipótesis de regresión.

- Encuentre la ecuación de regresión lineal para el número de piezas defectuosas que fabrica un empleado por día como función del número de minutos que llega tarde al trabajo.

$$n = 8 \quad \sum xy = 3019 \quad \sum y = 116 \\ \sum x = 147 \quad \sum x^2 = 3879$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \\ a = \frac{\sum y - b \sum x}{n} = \bar{y} - b\bar{x}$$

$$b = \frac{8 \cdot 3019 - 147 \cdot 116}{8 \cdot 3879 - 147^2} = 0.749230606$$

$$a = \frac{116 - 0.749230606 \cdot 147}{8} = 0.732887615$$

$$\boxed{y = 0.732887615 + 0.749230606}$$

- Aproximadamente, ¿cuál es el número promedio de piezas defectuosas que realiza un empleado que llega puntual a su trabajo?

R/ 0.732887615 piezas en promedio  $a$

- Aproximadamente, ¿cuánto es el aumento promedio en el número de piezas defectuosas que realiza un empleado por cada minuto que llega tarde al trabajo?

R/ 0.749230606 piezas por minuto  $b$

- Aproximadamente, ¿qué porcentaje de  $\underbrace{\text{variación de en el número de}}$  piezas defectuosas que elabora un empleado en un día se debe a otros factores distintos al número de minutos que llega tarde a trabajar?

$$\text{Si } n \text{ importar } x \rightarrow 1 - r^2$$

$$r^2 = b \frac{s_x}{s_y} = \left( b \sqrt{\frac{S_{xx}}{S_{yy}}} \right)^2$$

$$n = 8 \quad \sum x^2 = 3879 \quad \sum y^2 = 2396 \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$b = 0.749230606 \quad \sum x = 147 \quad \sum y = 116 \quad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xx} = 3879 - \frac{147^2}{8} = 1177.875 \\ S_{yy} = 2396 - \frac{116^2}{8} = 668$$

$$r^2 = b \frac{s_x}{s_y} = \left( b \sqrt{\frac{S_{xx}}{S_{yy}}} \right)^2$$

$$1 - r^2 = 1 - \left( 0.749230606 \cdot \sqrt{\frac{1177.875}{668}} \right)^2 = 0.009222877$$

- A un nivel de significancia del 5%, ¿existe evidencia de que el aumento promedio en el número de piezas defectuosas que realiza un empleado por cada minuto que llega tarde al trabajo es mayor a 0.7?

$$H_0: \beta = 0.7 (\leq)$$

$$n = 8$$

$$\alpha = 0.05$$

$$H_1: \beta > 0.7$$

$$b = 0.749230606$$

$$v = 6$$

$$\sum x^2 = 3879$$

$$\beta_0 = 0.7$$

$$T = (B - \beta_0) \frac{\sqrt{S_{xx}}}{S} \text{ con } v = n - 2$$

$$\sum x = 147$$

$$\sum xy = 3074$$

$$\sum y^2 = 2396$$

$$\sum y = 116$$

Como es 1

cola, se usa

directa

sin dividir

$$S_{xx} = 3879 \quad \frac{147^2}{8} = 1177.875$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = 2396 - \frac{116^2}{8} = 668$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = 3074 - \frac{147 \cdot 116}{8} = 882.5$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$s_B = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$$

$$S = \sqrt{\frac{668 - 0.749230606 \cdot 882.5}{8-2}} = 0.68361663$$

$$T = (B - \beta_0) \frac{\sqrt{S_{xx}}}{S} \text{ con } v = n - 2$$

$$t_{0.05} = (0.749230606 - 0.7) \cdot \frac{\sqrt{1177.875}}{0.68361663} = 2.47157$$

$$t_C = t_{0.05} \cdot 6 = 1.97318$$

**B1** Como  $t_{0.05} = 2.47157 > t_C = 1.97318$   
se rechaza  $H_0$

$$P(T > 2.47157) = 0.02478 < \alpha = 0.05$$

Se rechaza  $H_0$