Tema 9: Correlación y Regresión lineal simple

Probabilidad y Estadística

Ingeniería Industrial

FCA-UNNE

2020

Métodos que relacionan dos o mas variables

- **Correlación lineal simple**
  - Definición
  - Coeficiente de correlación.
  - Características e interpretación.
- **Regresión lineal simple**
  - Recta de regresión. Cálculo de los estimadores a y b.
  - Prueba de hipótesis del coeficiente de regresión. Valor predictivo de la regresión.

Relación entre dos o mas variables

El tema que veremos a continuación considera la situación en la que que tenemos información simultánea de dos variables, es decir, las observaciones constan de una sucesión de parejas de datos (x1, y1), (x2, y2), . . . , (xn, yn). Esta situación corresponde al hecho de llevar a cabo dos mediciones o dos preguntas a cada unidad de análisis.

Por ejemplo

- Distancia recorrida y velocidad
- La nota del alumno del primer parcial y la del exámen final

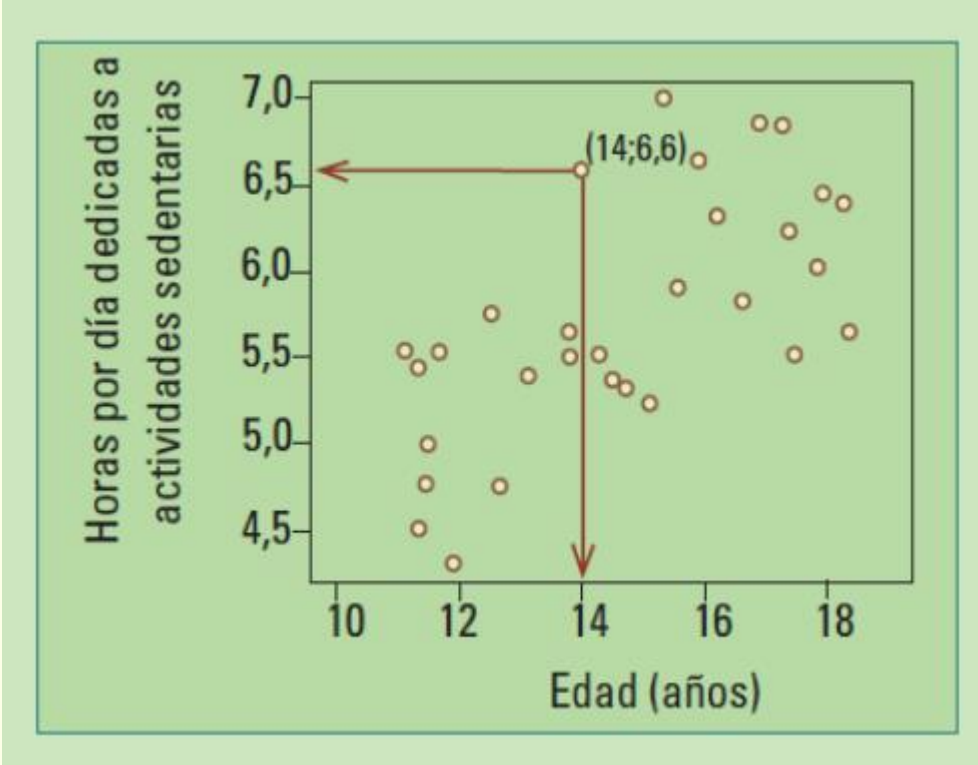
Ejemplos

- Cantidad de calorías consumidas y peso
- La presión arterial y la expectativa de vida
- Salario y Valor del metro cuadrado de su vivienda

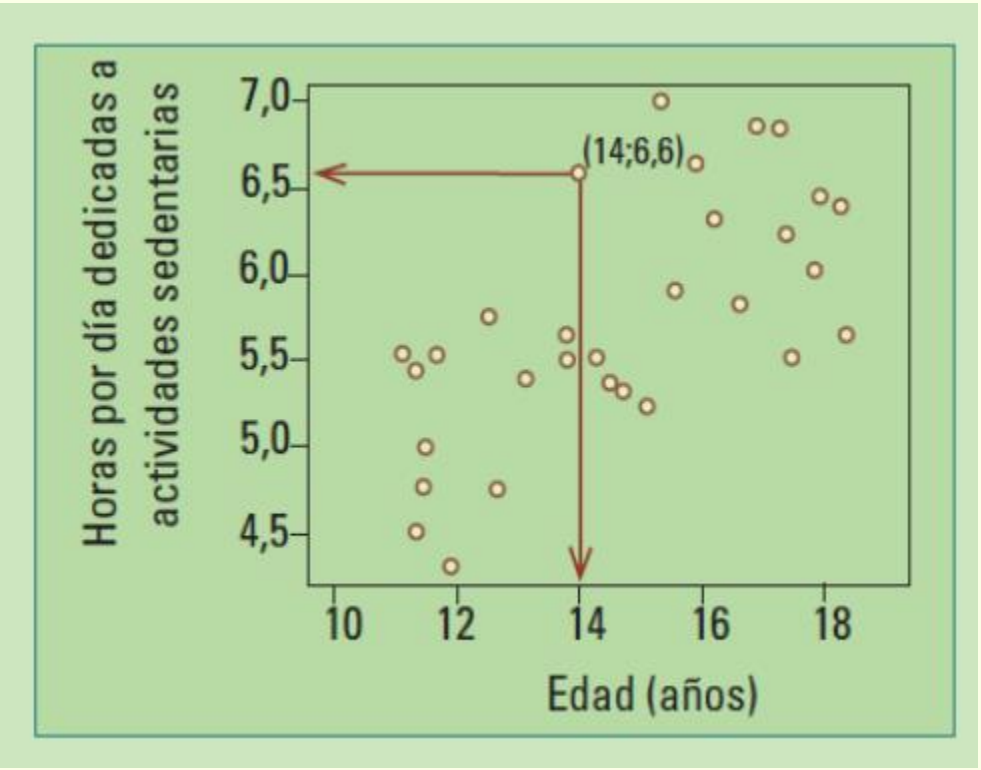
Dispersogramas

La forma gráfica más habitual de describir la relación entre dos variables cuantitativas es utilizando un **diagrama de dispersión.**
Cada **punto** corresponde a **un par de valores (x1, y1)**, medidos/observados sobre el mismo individuo.

Ejemplo : En un estudio de obesidad en la población, se solicitó a un grupo de 60 adolescentes que registrara durante un mes la cantidad de horas que dedicaban cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promediaran.



Dispersogramas



En un diagrama de dispersión observamos el patrón general de la relación entre las variables mirándolo de izquierda a derecha.

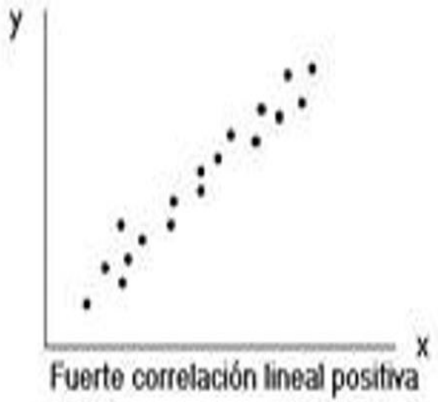
Si a medida que **x aumenta** (es decir, nos corremos hacia la derecha del gráfico), **y aumenta** también**,** esto indica una asociación lineal positiva entre las variables.

Correlación lineal simple

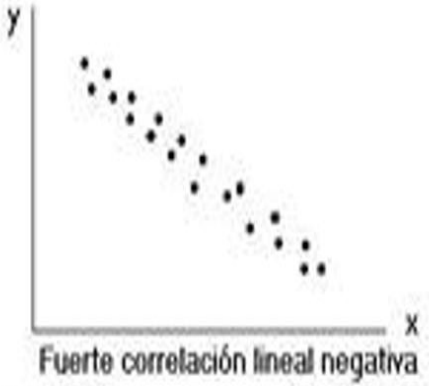
Dadas 2 variables cuantitativas aleatorias, la Correlación es una medida del grado en que las 2 variables varían conjuntamente o una medida de la intensidad de la asociación entre las variables.

La correlación mide el *grado de asociación* entre dos magnitudes cualesquiera.

Gráfico: Tipos genéricos de correlación entre dos variables X e Y.

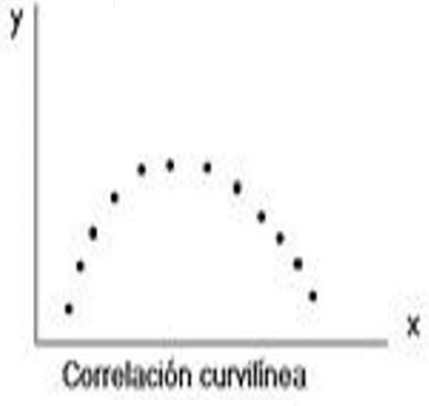


Fuerte correlación lineal positiva



Fuerte correlación lineal negativa

Ninguna correlación lineal aparente



Correlación curvilínea

Gráficos de arriba : x aumenta, y aumenta (relación positiva)
x aumenta, y disminuye (relación negativa)

Gráficos de abajo : x aumenta, y aumenta o disminuye (sin relación)
x aumenta, y aumenta para los primeros valores de x y luego disminuye (relación curvilínea)

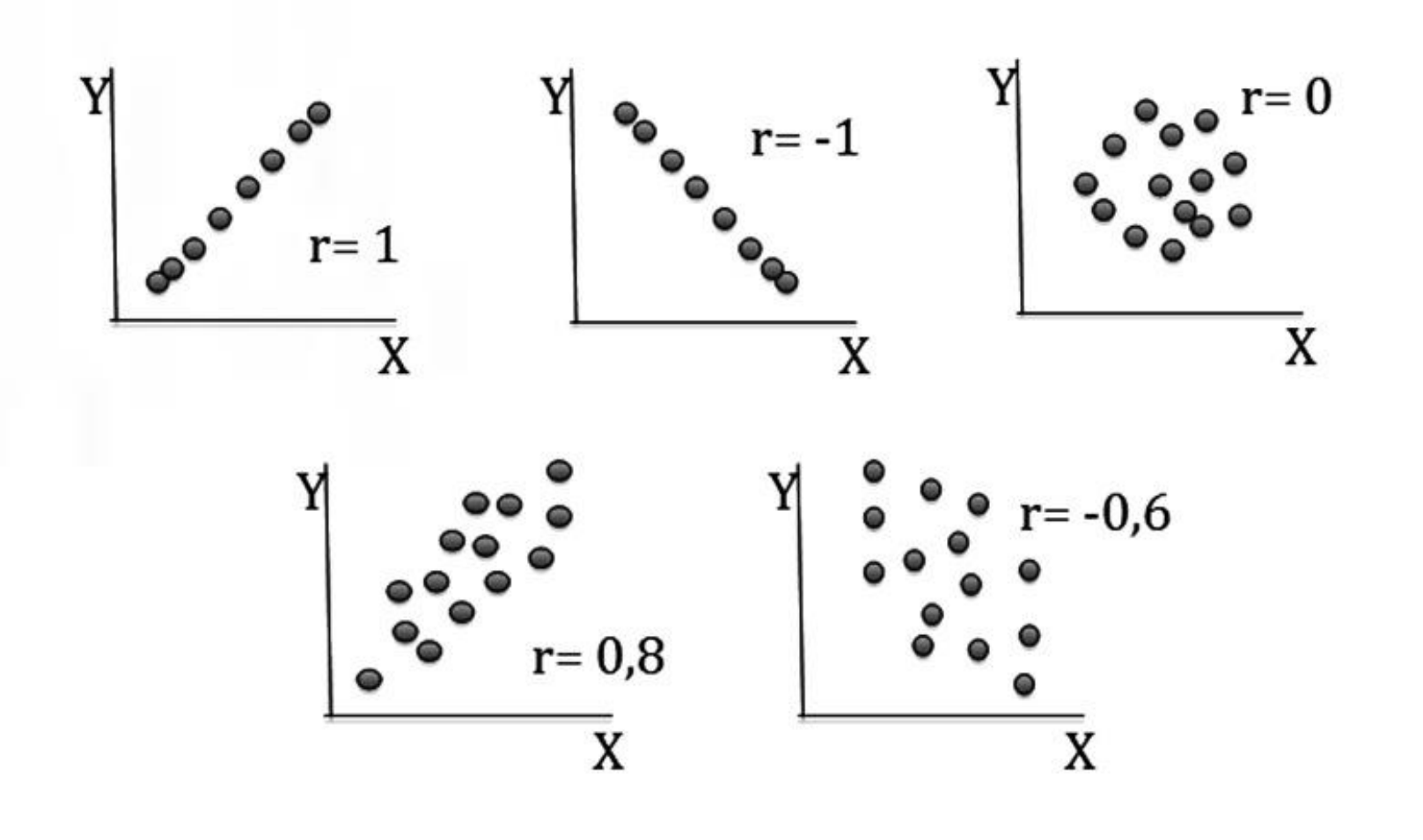
Coeficiente de Correlación

El Coeficiente de Correlación de Pearson (r) es un número que representa el grado de asociación lineal entre los pares de valores de dos variables continuas se simboliza con la letra ρ, no no depende de las escalas de las variables.

ρ tendrá valores comprendidos en el intervalo (-1,1)

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

Estimador r del coeficiente de correlación



Cálculo del Coeficiente de Correlación de Pearson

$$r = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sqrt{\left[\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}\right]\left[\sum_{i=1}^{n} y^2 - \frac{\left(\sum_{i=1}^{n} y\right)^2}{n}\right]}} = \frac{SP}{\sqrt{SCx * SCy}}$$

La correlación es el cociente entre la Covarianza entre x e y y la raíz cuadrada de la varianza de x por la varianza de y.

$$\rho = \frac{Co\,\mathrm{var}\,x, y}{\sqrt{Varx * Vary}}$$

Prueba de hipótesis de r

Cómo saber si hay correlación significativa o no?
Que valor tiene que tener r para decir que hay correlación?

Para responder a estas preguntas se plantean las siguientes hipótesis?

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

$$t_c = \frac{r - \rho}{S_r} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} =$$

Si P-Valor de *tc* < *α* entonces se rechaza H_0

Prueba de hipótesis respecto a si ρ=0

Aunque el coeficiente sea r = 0,975, se debe probar la significancia de ese coeficiente!
Sobre todo cuando la muestra es chica.
Si queremos descartar que los datos muestrales proviene de una población con ρ = 0 (α = 5%, n=12).

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

$$t_c = \frac{r-\rho}{S_r} = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0{,}975}{\sqrt{\frac{1-0{,}951}{10}}} = \frac{0{,}975}{0{,}07} = 13{,}9$$

El P- valor de (tc=13.9, t 10 gl) es P-valor= 0.000000036
entonces se rechaza $H_{0.}$ por lo tanto se concluye que existe correlación entre las variables.

REGRESION LINEAL SIMPLE

Qué es la Regresión Lineal Simple

Técnica estadística para modelar la relación entre variables.

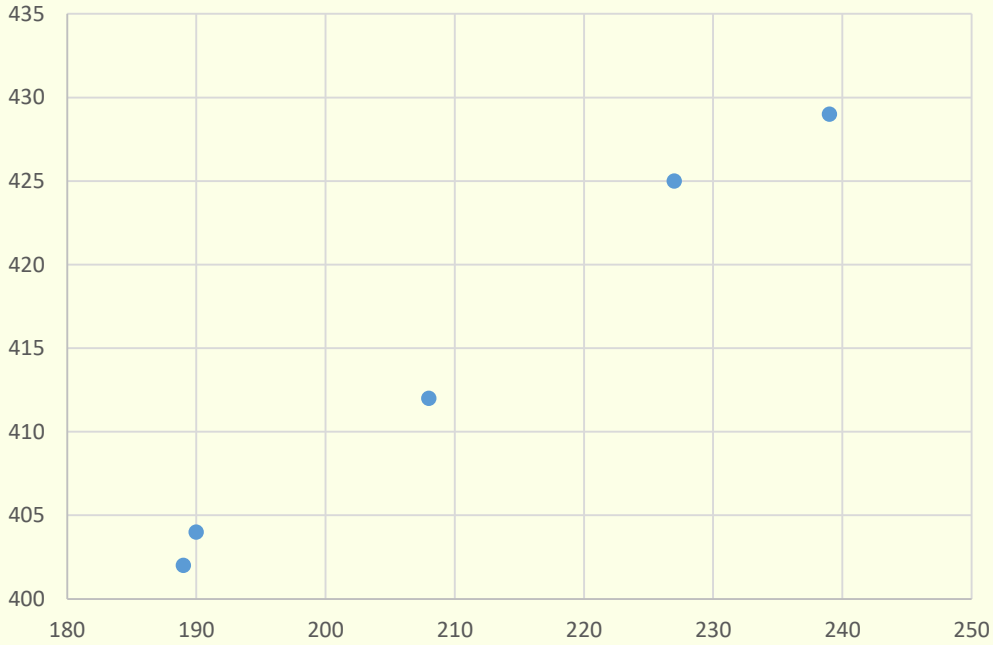
Tiene aplicaciones en mucha ciencias como ser:

- Ingeniería
- Ciencias físicas y químicas
- Economía
- Ciencias biológicas
- Ciencias de la salud
- Ciencias sociales

EJEMPLO

Una compañía desea hacer predicciones del valor anual de sus ventas totales en cierto país a partir de la relación de estas con la renta nacional.
Se dispone de los siguientes datos:

X (renta)	189	190	208	227	239
Y(Valor de las Ventas)	402	404	412	425	429



Mas ejemplos

¿Las ventas dependen de la inversión en publicidad?
La inversión realizada en el proceso productivo y el rendimiento.
Número de horas de estudio y calificación obtenida.
Número de cursos de capacitación y número de accidentes laborales.

En análisis de regresión se puede realizar a partir de datos obtenidos de:

-Estudios retrospectivos basados en datos históricos

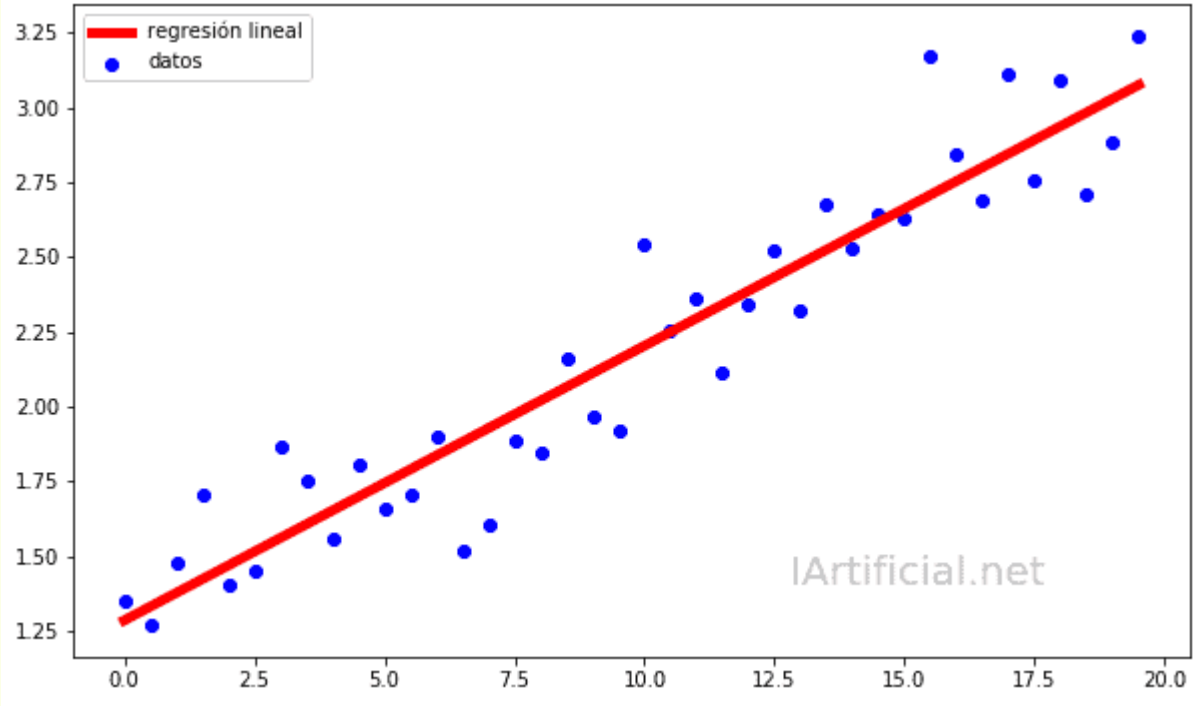
-Estudios observacionales

-Experimentos diseñados

Regresión Lineal Simple

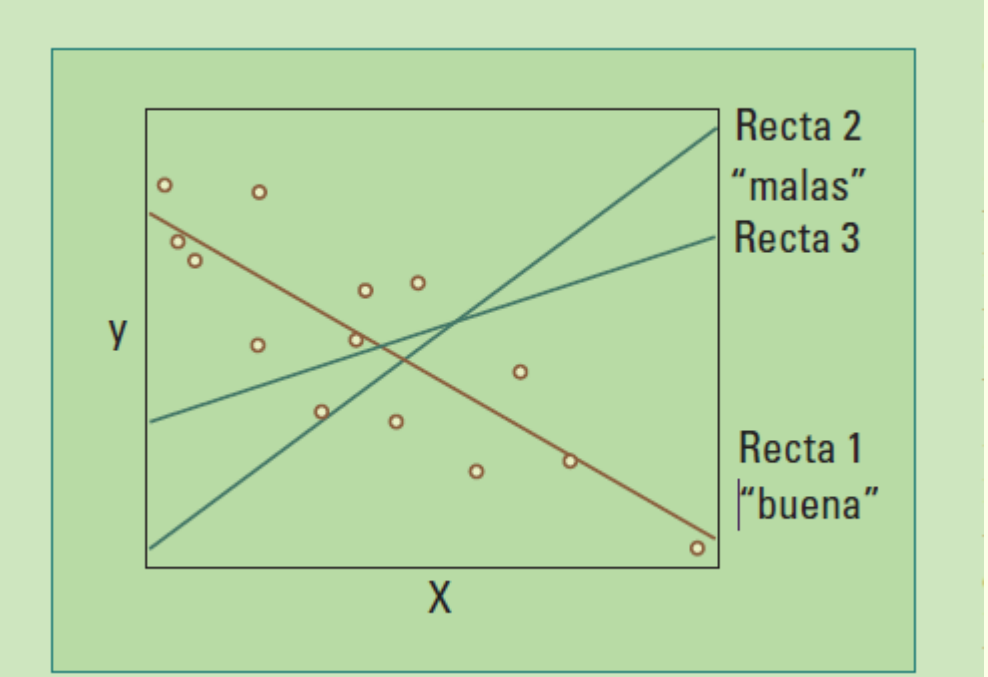
Cuando un diagrama de dispersión muestra un patrón lineal es deseable resumir ese patrón mediante la ecuación de una recta.

Esa recta debe representar a la mayoría de los puntos del diagrama, aunque ningún punto esté sobre ella.



La recta de la figura representa bien la dirección y el sentido de la asociación entre los valores de **X** e **Y**, pasa “cerca” de la mayoría de los puntos del diagrama de dispersión; decimos que es una recta “buena”.

Muchas rectas se pueden trazar sobre una nube de puntos pero algunas no son rectas que describen la relación entre x e y, en la figura las rectas 2 y 3 no representa bien la dirección y el sentido de la asociación entre los valores de **X** e **Y.**



Pasamos ahora a una definición mas formal de la recta de regresión

Modelo de Regresión Lineal Simple

Pasamos ahora a una definición mas formal de la recta de regresión, a diferencia de la expresión matemática de la recta, en estadística incorporamos un término denominado error, el cual representa la distancia entre la recta y el punto observado.

$$y = \alpha + \beta x + \varepsilon$$

y = variable respuesta o dependiente

x = variable regresora o independiente

α = ordenada al origen

β = pendiente

α y β coeficientes de regresión (parámetros desconocidos)

ε = componente aleatorio del error $(0, \sigma^2)$

Se expresan con letras griegas la ordenada al origen y la pendiente porque se refieren a parámetros poblacionales desconocidoa que serán estimados con datos muestrales.

Estimación de los parámetros por el método de Mínimos Cuadrados

Existen varios métodos para determinar a y b, por ejemplo encontrando la recta que pasa por 2 puntos, o bien el método del gradiente. En este curso veremos el método de mínimos cuadrados, que encuentra la recta que minimiza la suma de los errores.

$$Y = \alpha + \beta X + \varepsilon$$

$$\hat{y}_i = a_i + b_i x_i$$

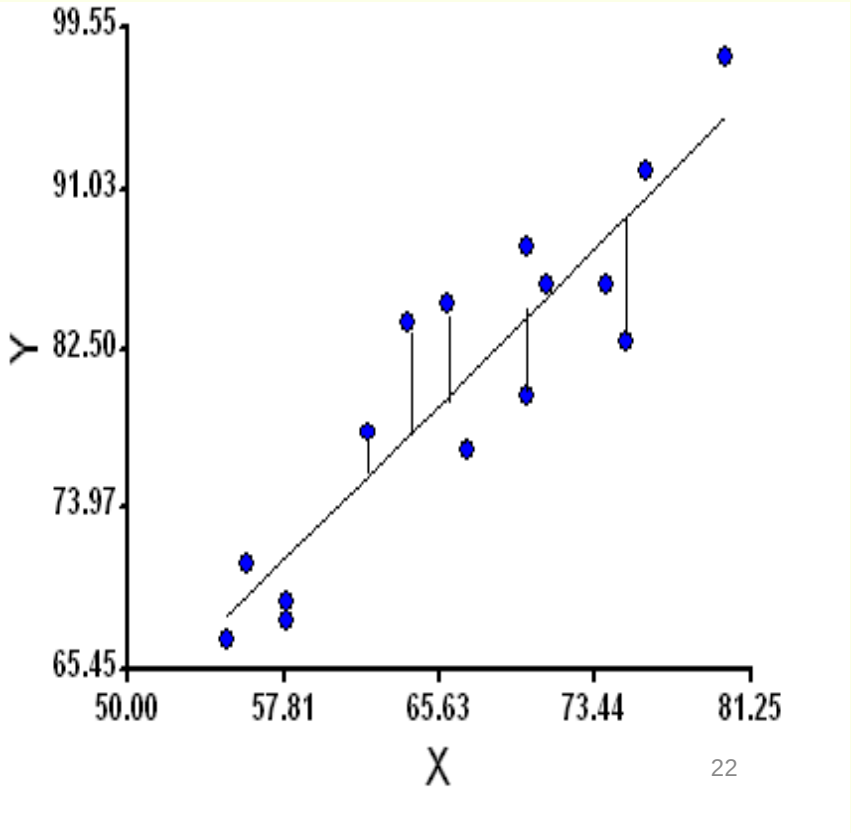
i= 1, 2, …, n (modelo muestral)

$$e_i = y_i - \hat{y}_i = y_i - (a_i + b_i x_i)$$

Criterio de mínimos cuadrados:

El método de MC da como valor de estimación de parámetros aquellos que minimizan la siguiente expresión:

$$S(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [(y_i - (a + b x_i)]^2$$



Estimación de los parámetros

La función S tendrá un mínimo cuando sus derivadas parciales en relación a “a” y “b” sean nulas.

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

Simplificando estas dos ecuaciones se llega a un sistema de ecuaciones normales:

$$n.a + b.\sum x_i = \sum y_i$$

$$a.\sum x_i + b.\sum x_i^2 = \sum x_i y_i$$

Estimación de los parámetros

Resolviendo el sistema de 2 ecuaciones con 2 incógnitas anterior, se obtiene la siguiente solución.

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

La propuesta es calcula b y luego a con el promedio de y y de x.

Propiedades de los estimadores por mínimos cuadrados

1- La suma de los residuales en cualquier modelo de regresión que contenga ordenada al origen siempre es igual a cero:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} e_i = 0$$

2- La suma de los valores observados es igual a la suma de los valores ajustados (predichos):

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3- La suma de los residuales, ponderados por los valores de la variable regresora, siempre es igual a cero:

$$\sum_{i=1}^{n} x_i e_i = 0$$

4- La línea de regresión de mínimos cuadrados siempre pasa por el **centroide** de los datos, que es el punto $\left(\overline{y}, \overline{x}\right)$

Prueba de hipótesis de la regresión

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

El procedimiento de prueba para esta hipótesis se puede establecer por dos métodos. El primero:

$$t_c = \frac{b-0}{S_b} = \frac{b-0}{\sqrt{\frac{\sum yx^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}}$$

Se rechazaría H_0 si P-valor de t_c < α

Rechazar H_0 implica que hay relación lineal entre x e y.

INTERVALO DE CONFIANZA PARA β

Si los errores se distribuyen en forma normal e independiente, entonces la distribución de muestreo tanto de a y b es t con n-2 grados de libertad.

$$P\left[b - t_{5\%} Sb \leq \beta \leq b + t_{5\%} Sb\right] = 1 - \alpha = 0.95$$

COEFICIENTE DE DETERMINACIÓN

$$0 \leq R^2 \leq 1$$

Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.

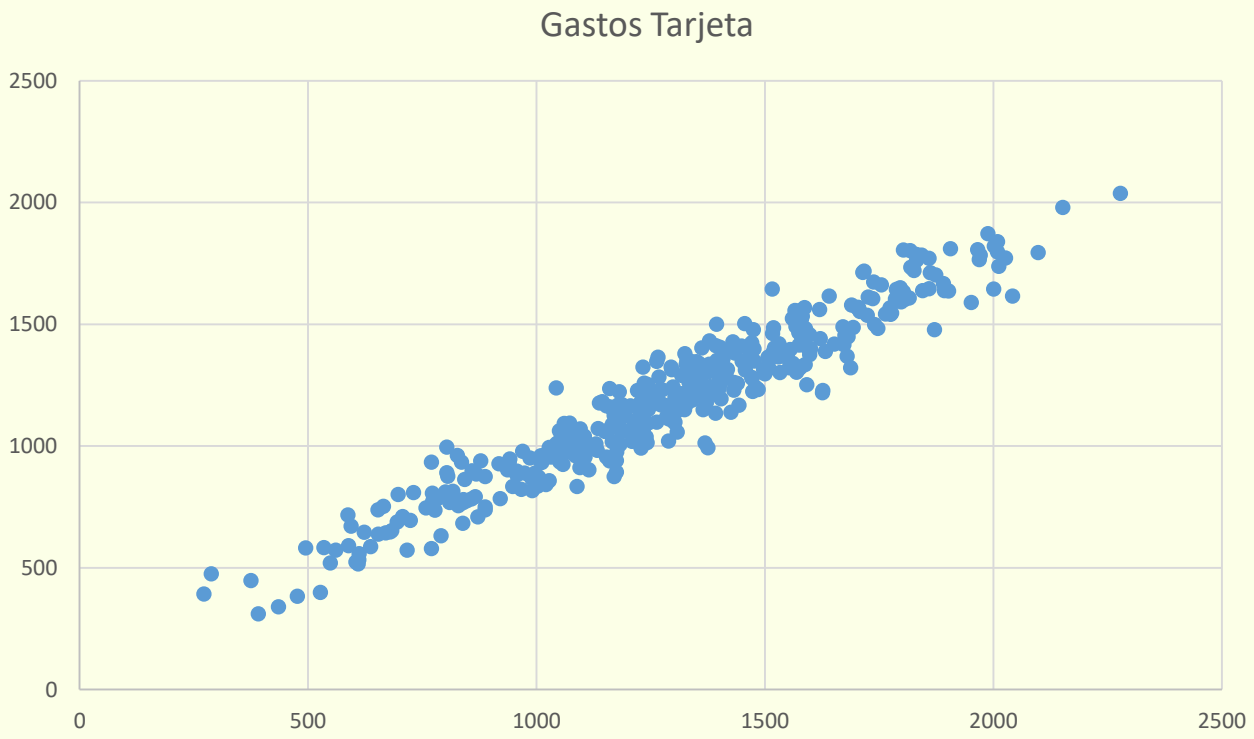
El estadístico R^2 se debe usar con precaución, porque siempre es posible conseguir que R^2 sea grande agregando términos suficientes al modelo.

¡VALORES ALTOS DE R^2 NO GARANTIZAN UN BUEN AJUSTE!

EJEMPLO 1

Variable dependiente (y): Gastos con tarjeta de crédito

Variable regresora (x): Salario



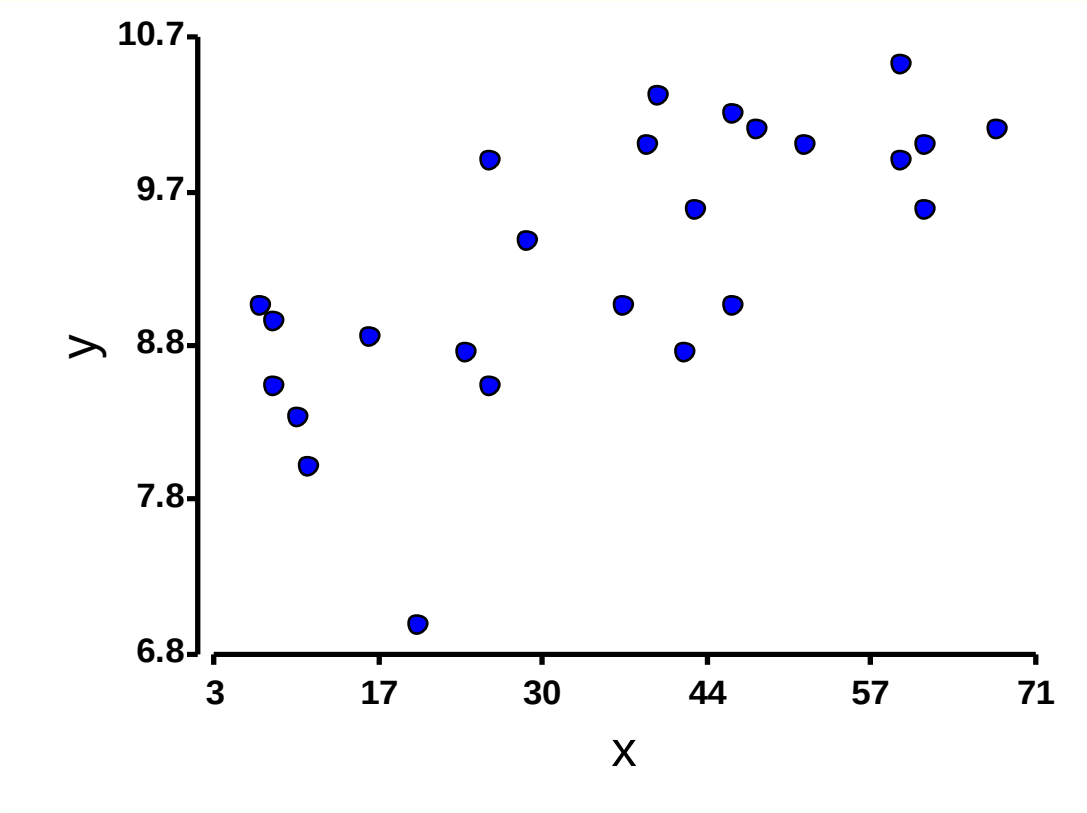
Variable	N	R^2				
GastosTarjeta	400	0.92				
Coeficientes de regresión y estadísticos asociados						
Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	93.61	16.5	61.17	126	5.67	<0.0001
Ingreso	0.85	0.01	0.82	0.87	68.94	<0.0001

$$y = 93.61 + 0.85\,x$$

EJEMPLO 2

Variable dependiente (y): pH

Variable regresora (x): profundidad (cm) en muestras de suelo de una región



Variable	N	R^2
Y	25	0.50

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor
const	8.16	0.26	7.62	8.71	31.16	<0.0001
X	0.03	0.01	0.02	0.04	4.82	0.0001

$$y = 8.16 + 0.03\ x$$

En el ejemplo del contenido de nitrógeno en suelo (X) y los valores promedio de nitrógeno por planta (Y) hallar la ecuación de regresión:

	X	Y	X Y	X^2	Y^2
	0,42	0,13	0,055	0,176	0,017
	0,45	0,15	0,068	0,203	0,023
	0,50	0,16	0,080	0,250	0,026
	0,55	0,17	0,094	0,303	0,029
	0,68	0,18	0,122	0,462	0,032
	0,69	0,18	0,124	0,476	0,032
	0,70	0,19	0,133	0,490	0,036
	0,73	0,20	0,146	0,533	0,040
	0,80	0,20	0,160	0,640	0,040
	0,90	0,21	0,189	0,810	0,044
	0,92	0,22	0,202	0,846	0,048
	0,94	0,23	0,216	0,884	0,053
Total	8,28	2,22	1,589	6,073	0,42

$$b = \frac{1.5289 - (8.28) * (2.22)/12}{6.073 - (8.28)2/12}$$

a =(2.22/12)-b(8.28/12)

La ecuación de regresión es entonces:

$$\hat{y} = 0,075 + 0,16x$$

Prueba de hipótesis de b:

Dado el coeficiente b del ejercicio anterior, probar si la muestra proviene de una población con β = 0. Nivel 5%

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

$$t_c = \frac{b-\beta}{S_b} = \frac{b-0}{\sqrt{\frac{S_{yx}^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}} = 13,93$$

$$t_c = \frac{b-\beta}{S_b} = \frac{b-0}{\sqrt{\frac{S_{yx}^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}} = 13,93$$

Por ser el t_c superior al valor de tabla correspondiente (2,228), se rechaza H_0. Por lo que se concluye que el coeficiente de regresión poblacional es diferente de cero. La diferencia entre la pendiente estimada y la pendiente poblacional hipotética no se puede explicar (en este caso) por las variaciones aleatorias solamente.

Intervalo de confianza para β:

$$P\left(b - t_\alpha S_b \leq \beta \leq b + t_\alpha S_b\right) = 1 - \alpha$$

$$P\left(b - t_\alpha S_b \leq \beta \leq b + t_\alpha S_b\right) = 1 - \alpha$$

$$P\left(0{,}133 \leq \beta \leq 0{,}186\right) = 0{,}95$$