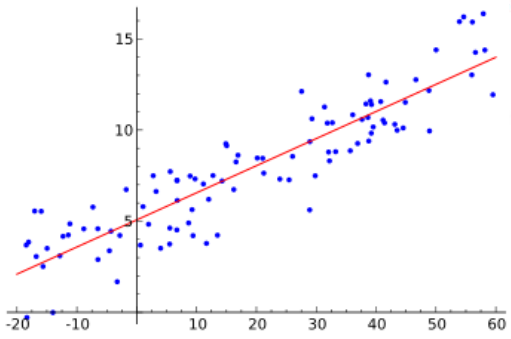
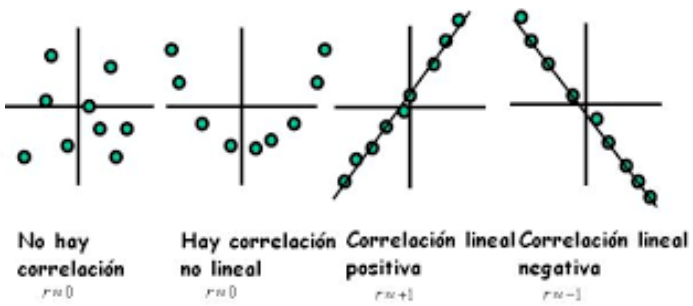
CONCEPTOS BÁSICOS DE

ESTADISTICA INFERENCIAL





Inferencia Estadística

Se ocupa de estudiar los métodos necesarios para extraer, o inferir, conclusiones válidas e información sobre una población a partir del estudio experimental **de una muestra** de dicha población.

Regresión lineal

Relaciones o dependencias entre las dos variables x e y

➢**Funcional**: Exista una relación matemática exacta que ligue ambas variables (ej. el radio y el área de un círculo).

➢**Aleatoria**: Cuando, aunque no exista entre las variables una relación exacta, se puede observar (aunque no siempre es el caso) una cierta tendencia entre los comportamientos de ambas (ej. El peso y la altura de un individuo).

El primer paso para el estudio de la relación entre las variables consiste en la construcción y observación de un diagrama de dispersión $y = f(x)$

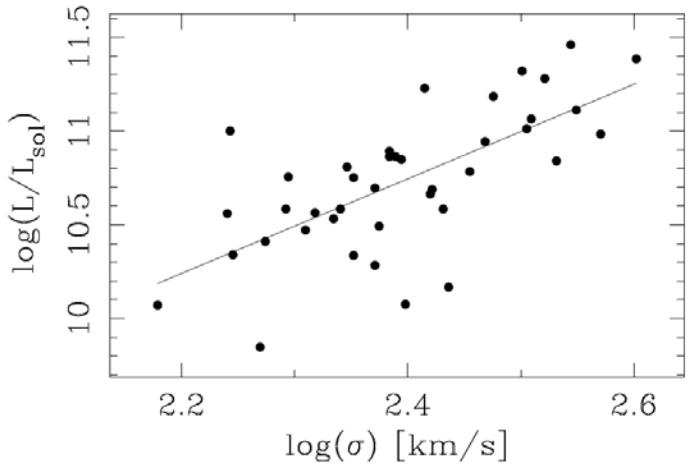


Fig. Ejemplo de diagrama de dispersión. Los datos corresponden a las medidas de dispersión de velocidades y luminosidad en una muestra de 40 galaxias elípticas realizadas por Schechter (1980).

El problema de la regresión se concreta entonces en ajustar una función a la nube de puntos representada en dicho diagrama

$y = f(x)$ $\quad x = f(y)$

Se conoce como **línea de regresión** a la representación gráfica de la función que se ajusta a la nube de puntos del diagrama de dispersión.

Cuando dicha nube se distribuya aproximadamente a lo largo de una línea recta ajustaremos **una recta de regresión**.

Modelo de regresión lineal simple

Modelo de Regresión lineal:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_0, β_1 Parámetros del modelo ε Variable aleatoria (Error)

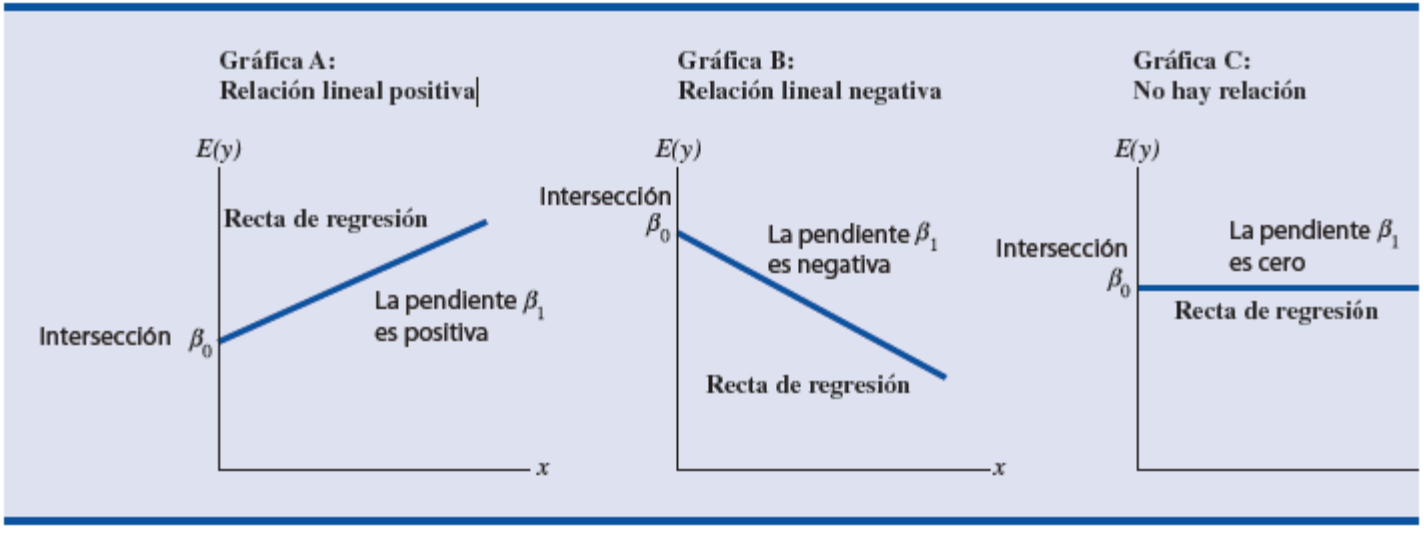
Ecuación de regresión lineal simple

$$E(y) = \beta_0 + \beta_1 x$$

$E(y)$:Es la media o valor esperado de y *para* un valor dado de x.

β_0 es la intersección de la recta de regresión con el eje y, β_1 *es la pendiente*

EJEMPLOS DE LÍNEAS DE REGRESIÓN EN LA REGRESIÓN LINEAL SIMPLE



Regresiones otras formas funcionales de la regresiones

$$y = a + bx + cx^2 \qquad y = ab^x$$

Ecuación de regresión estimada

Se calculan estadísticos muestrales (que se denotan b_0 y b_1) *como estimaciones de* los parámetros poblacionales β_0 y β_1 . Sustituyendo en la ecuación de regresión b_0 y b_1 *por los valores de los estadísticos muestrales* β_0 y β_1 se obtiene la **ecuación de regresión estimada.**

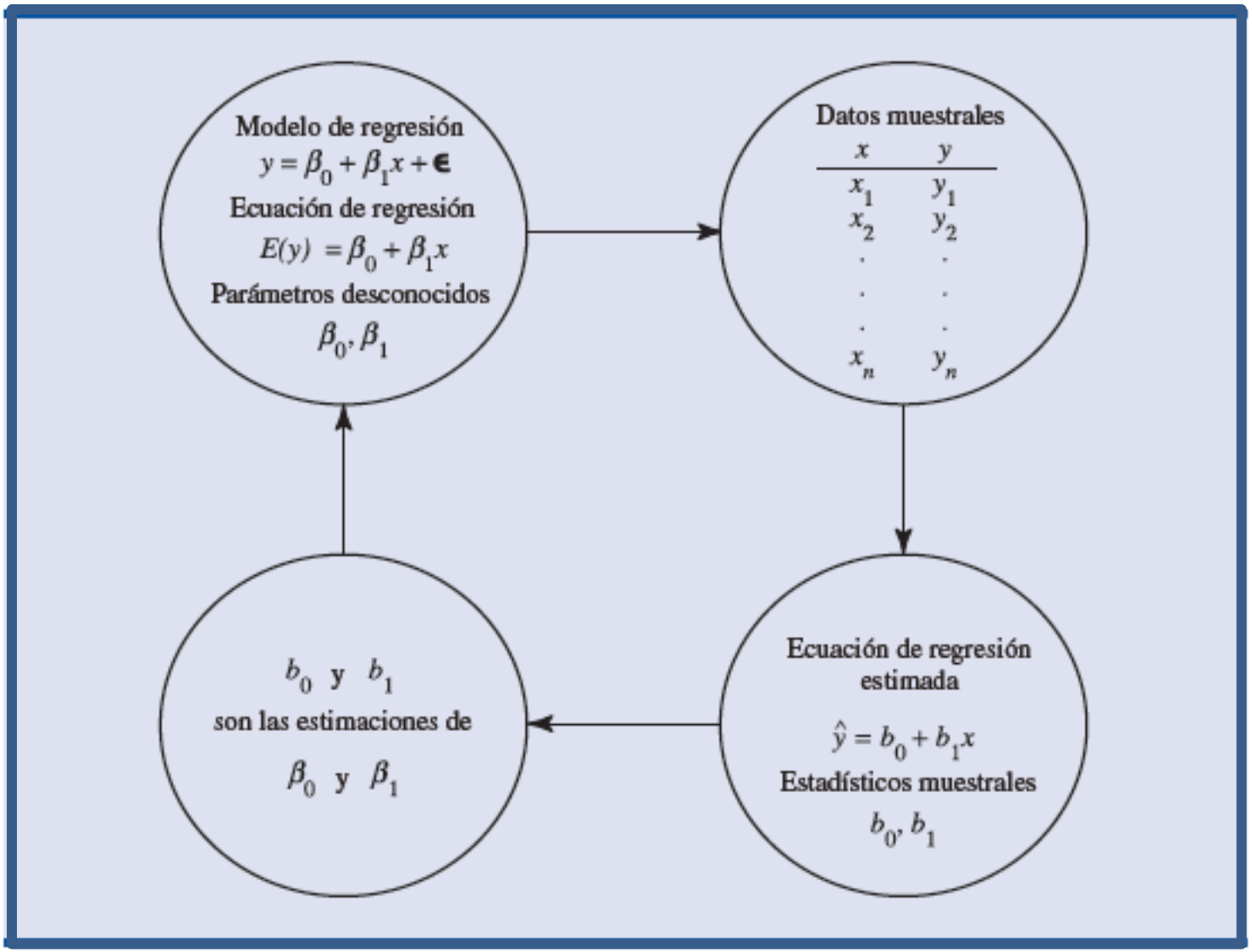
$$\hat{y} = b_0 + b_1 x$$

A la gráfica de la ecuación de regresión simple estimada se le llama *recta de regresión estimada*

b_0 es la intersección de la recta de regresión con el eje *y,* b_1 *es la pendiente*

En general, $\hat{y}$ es el estimador puntual de *E(y), el valor medio de las y para un valor dado de x*

PROCESO DE ESTIMACIÓN EN LA REGRESIÓN LINEAL SIMPLE



Ajuste de una recta de regresión

El método de mínimos cuadrados es un método en el que se usan los datos muestrales para hallar la **ecuación de regresión estimada.**

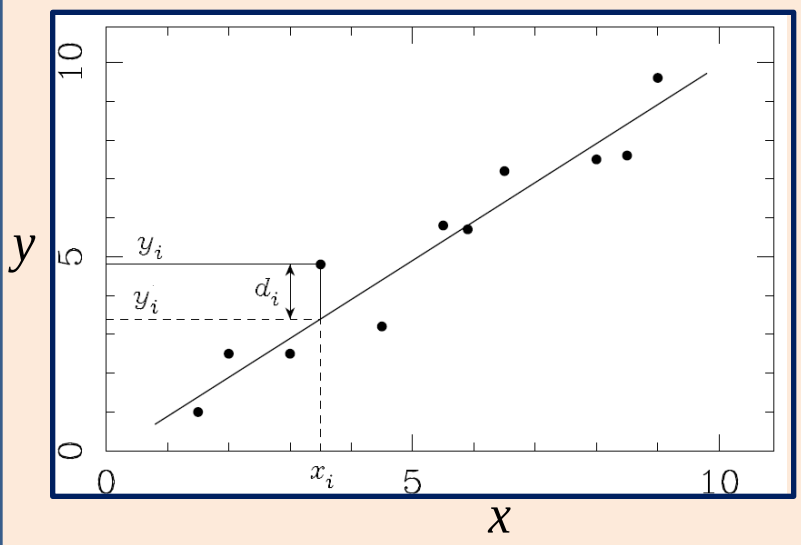
Sea una muestra de tamaño n en que la variable estadística bidimensional toma los valores.

$$(x_1, y_1), (x_2, y_{21}), \ldots, (x_n, y_1).$$

A cada valor x_i de la variable x le corresponde entonces un valor y_i de la variable y, pudiendo además asociársele un valor y_i^*, que sería el dado por la recta que queremos calcular. Es decir

$$\hat{y}_i = b_0 + b_1 x_i$$

Sea d_i a la diferencia entre los dos valores, observado y dado por la recta, de la variable y en cada punto.



$$d_i = \hat{y}_i - y_i.$$

Para que la recta a determinar sea la que mejor se ajuste a la nube de puntos de entre todas las rectas posibles, dichas distancias d_i deberían ser lo más pequeñas posible. Es decir, hay que minimizar los d_i

Tomar los cuadrados de las distancias, para que así no se anulen desviaciones positivas y negativas.

El problema se reduce a minimizar la expresión

$$\min \sum (y_i - \hat{y})^2$$

y_i valor observado de la variable dependiente en la observación *i*

$\hat{y}$ valor estimado de la variable independiente en la observación *i*

PENDIENTE E INTERSECCIÓN CON EL EJE *y* DE LA ECUACIÓN DE REGRESION ESTIMADA

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

x_i valor la variable independiente en la observación *i*

y_i valor de la variable dependiente en la observación *i*

$\bar{x}$ media de la variable independiente

$\bar{y}$ media de la variable dependiente

n numero total de observaciones

$$\bar{y} = b_0 + b_1 \bar{x} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

La recta de regresión debe pasar por $(\bar{x}, \bar{y})$, es decir, por el centro de la nube de puntos.

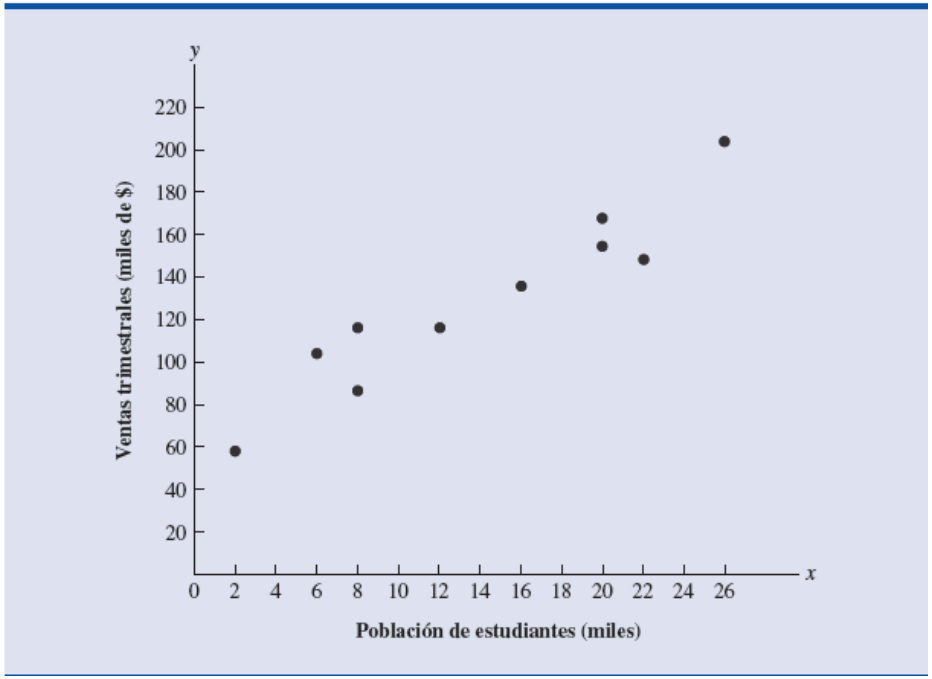
la ecuación de regresión estimada es $\hat{y} = b_0 + b_1 x$

El desarrollo anterior puede generalizarse para calcular expresiones similares para la regresión parabólica y, en general, polinómica.

POBLACIÓN DE ESTUDIANTES Y VENTAS TRIMESTRALES EN 10 RESTAURANTES DE LA CIUDAD

Restaurante i	Población de estudiantes (miles) x_i	Ventas trimestrales (miles de $) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

DIAGRAMA DE DISPERSIÓN EN EL QUE SE MUESTRA LA POBLACIÓN DE ESTUDIANTES Y LAS VENTAS TRIMESTRALES



Restaurante i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totales	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n} \quad \bar{x} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{n} \quad \bar{y} = \frac{1300}{10} = 130$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{2840}{568} \Rightarrow b_1 = 5$$

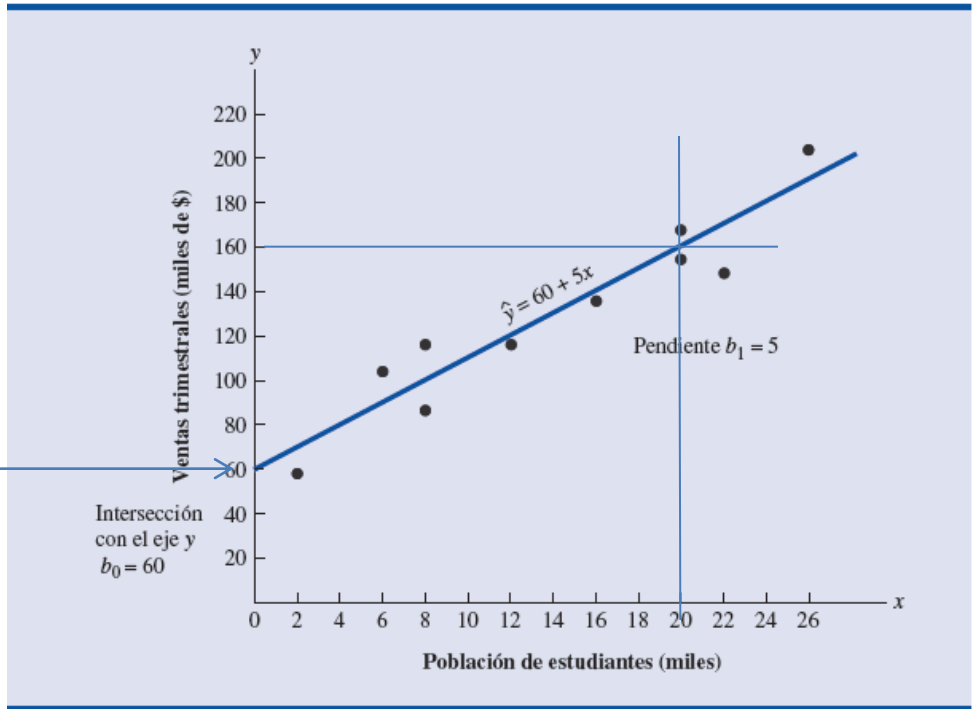
$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_0 = 130 - 5 \times 14 \Rightarrow b_0 = 60$$

la ecuación de regresión estimada es

$$\hat{y} = 60 + 5x$$

$$\hat{y} = 60$$

$$\hat{y} = 60 + 5(20) = 160$$



residual *i*

Diferencia que existe, en la observación *i*, entre el valor observado de la variable dependiente y_i, y el valor estimado de la variable dependiente $\hat{y}_i$. Representa el error que existe al usar $\hat{y}_i$ para estimar y_i.

Para la observación *i*, *el residual* es $y_i - \hat{y}_i$

❑*La suma de los cuadrados de estos residuales o errores es la cantidad que se minimiza empleando el método de los mínimos cuadrados.*

SUMA DE CUADRADOS DEBIDAAL ERROR

$$SCE = \sum (y_i - \hat{y}_i)^2$$

❑La diferencia $y_i - \bar{y}_i$ *proporciona una medida* del error que hay al usar $\bar{y}$ para estimar

SUMA TOTAL DE CUADRADOS

$$STC = \sum (y_i - \bar{y})^2$$

❑*suma de cuadrados debida a la regresión*

SUMA TOTAL DE CUADRADOS

$$SCR = \sum (\hat{y}_i - \bar{y})^2$$

RELACIÓN ENTRE STC, SCR Y SCE

$$STC = SCR + SCE$$

COEFICIENTE DE DETERMINACIÓN

Se usa para evaluar la bondad de ajuste de la ecuación de regresión estimada

Coeficiente de Determinación $r^2 = \frac{SCR}{STC}$ *(0,1)*

COEFICIENTE DE CORRELACIÓN MUESTRAL

$r_{xy} = (\text{signo de } b_1)\sqrt{r^2}$ *(–1 , +1)*

b_1 pendiente de la ecuación de regresión estimada $\hat{y} = b_0 + b_1 x$

El signo del coeficiente de regresión muestral es positivo si la ecuación de regresión tiene pendiente positiva ($b_1 > 0$) *y es negativo si la ecuación de regresión estimada tiene pendiente negativa* ($b_1 < 0$).

Conclusiones

✓El método de mínimos cuadrados proporciona una ecuación de regresión estimada que minimiza la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente y_i *y* los valores estimados de la variable dependiente .

✓El criterio de mínimos cuadrados permite obtener la y_i . ecuación de mejor ajuste. Si se empleara otro criterio, como minimizar la suma de las desviaciones absolutas entre y_i y $\hat{y}_i$ *se obtendría una ecuación* diferente. En la práctica el método de mínimos cuadrados es el método más usado.

CÁLCULO DE SCE EN EL EJEMPLO

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Ventas pronosticadas $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Error al cuadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
				SCE =	1530

$$\bar{y} = \frac{1300}{10} = 130$$

$$STC = \sum (y_i - \bar{y})^2$$

$$STC = 15730$$

$$SCR = STC - SCE \qquad SCR = 15730 - 1530 \Rightarrow SCR = 14200$$

COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{SCR}{STC} \qquad r^2 = \frac{14200}{15730} \Rightarrow r^2 = 0.9027$$

90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre el tamaño de la población de estudiantes y las ventas. Sería bueno que la ecuación de regresión tuviera un ajuste tan bueno.

COEFICIENTE DE CORRELACIÓN MUESTRAL

$$r_{xy} = (\text{signo de } b)\sqrt{r^2} \qquad r_{xy} = (+)\sqrt{0.9027} \qquad r_{xy} = 0.9501$$

se concluye que existe una relación lineal fuerte entre *x* y *y*.

Medidas de la asociación entre dos variables

Covarianza: Es una medida descriptiva de la asociación entre dos variables.

En una muestra de tamaño *n con observaciones* (x_1,y_1), (x_2,y_2), *etc., se* define como sigue:

covarianza muestral

$$Cov \equiv s_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

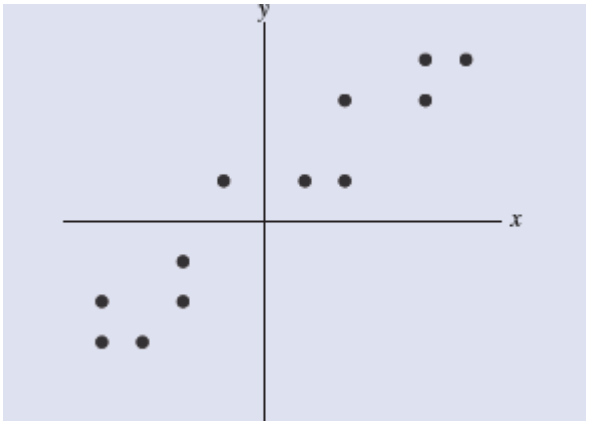
covarianza poblacional

$$Cov \equiv \sigma_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

Esta fórmula aparea cada x_i *con una* y_i

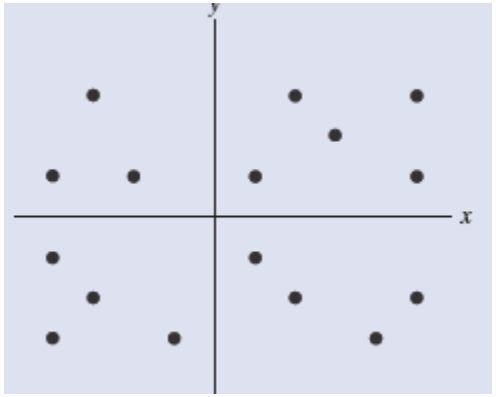
INTERPRETACIÓN DE LA COVARIANZA MUESTRAL

$S_{xy}>0$



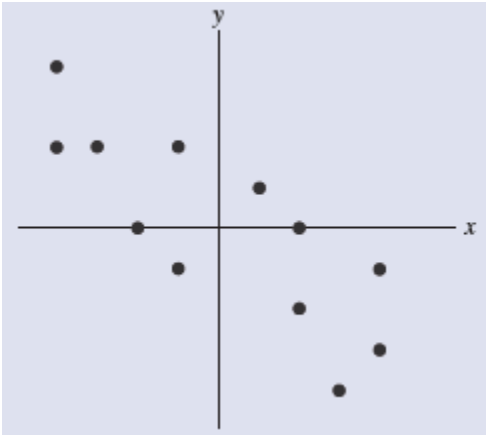
(*x* y *y* es lineal positiva)

$S_{xy}=0$



(no hay relación lineal entre *x y y)*

$S_{xy}<0$



la relación entre *x y* es lineal negativa)

Un valor positivo grande de la varianza indica una relación lineal positiva fuerte y que un valor negativo grande indica una relación lineal negativa fuerte.

Una medida de la relación entre dos variables, a la cual no le afectan las unidades de medición empleadas para *x y y, es el* ***coeficiente*** **de correlación.**

Coeficiente de correlación del producto MOMENTO DE PEARSON: Datos muestrales

$$r_{xy} = \frac{s_{xy}^2}{s_x s_y} \qquad s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad s_x = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

r_{xy} coeficiente de correlación muestral
s_{xy} covarianza muestral
s_x desviación estándar muestral de *x*
s_y desviación estándar muestral de *y*

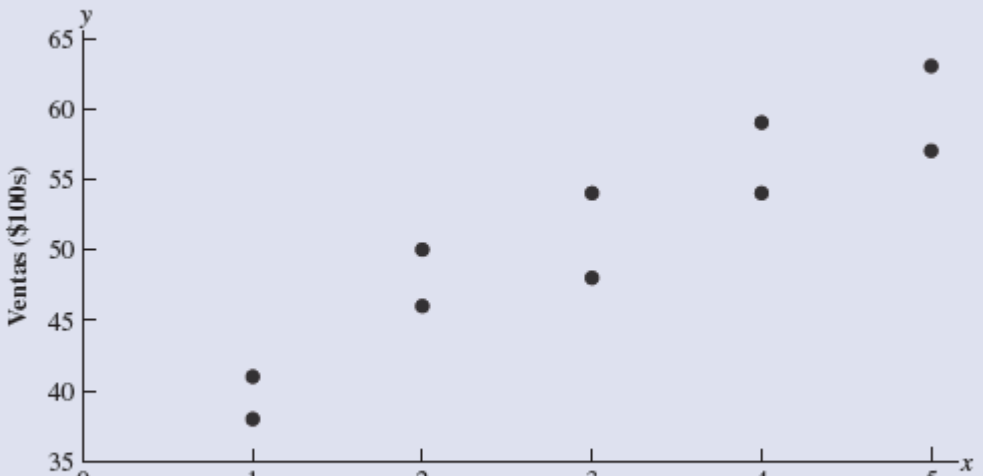
En general, si todos los valores del conjunto de datos caen en una línea recta con pendiente positiva, el coeficiente de correlación será 1; es decir, un coeficiente de correlación de 1 corresponde a una relación lineal positiva perfecta entre *x* y *y*. *Por otra parte, si* los puntos del conjunto de datos caen sobre una línea recta con pendiente negativa, el coeficiente de correlación muestral será *-1;* un coeficiente de correlación de *-1* corresponde a una relación lineal negativa perfecta entre *x y y*

El administrador de la tienda desea determinar la relación entre el número de comerciales televisados en un fin de semana y las ventas de la tienda durante la semana siguiente.

DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

Semana	Número de comerciales x	Volumen de ventas ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO



Covarianza muestral

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totales	30	510	0	0	99

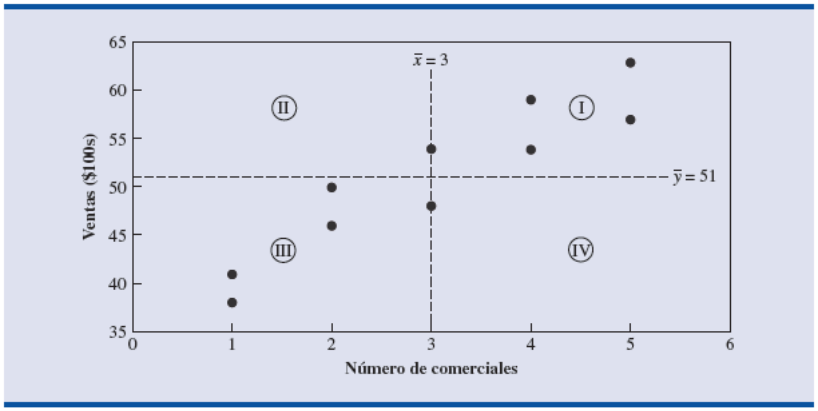
$n=10$

$$\bar{x} = \frac{30}{10} = 3$$

$$\bar{y} = \frac{510}{10} = 51$$

$$s_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}. \qquad s_{xy}^2 = \frac{99}{10-1}. = 11 \quad \Rightarrow s_{xy}^2 = 11$$

Tracemos una línea vertical punteada en $\bar{x} = 3$ y una línea horizontal punteada en $\bar{y} = 51$. Estas líneas dividen a la gráfica en cuatro cuadrantes.



Correlación positiva

desviación estándar muestral de las dos variables.

$$r_{xy} = \frac{s_{xy}^2}{s_x s_y}$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad s_x = \sqrt{\frac{20}{10-1}} \Rightarrow s_x = 1.49$$

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} \qquad s_x = \sqrt{\frac{566}{9}} \Rightarrow s_y = 7.93$$

$$r_{xy} = \frac{11}{1.49 \times 7.93} \Rightarrow r_{xy} = +0.93$$

se concluye que existe una relación lineal fuerte entre el número de comerciales y las ventas

debilidades

➢**Tanto la recta de regresión como el coeficiente de correlación no son robustos, en el sentido de que resultan muy afectados por medidas particulares que se alejen mucho de la tendencia general**.

➢**No hay que olvidar que el coeficiente de correlación no es más que una medida resumen. En ningún caso puede substituir al diagrama de dispersión, que siempre habrá que construir para extraer más información. Formas muy diferentes de la nube de puntos pueden conducir al mismo coeficiente de correlación.**

➢El que en un caso se obtenga un coeficiente de correlación bajo no significa que no pueda existir.

➢**correlación entre las variables. De lo único que nos informa es de que la correlación no es lineal (no se ajusta a una recta), pero es posible que pueda existir una buena correlación de otro tip**o.

➢Un coeficiente de correlación alto no significa que exista una dependencia directa entre las variables. Es decir, no se puede extraer una conclusión de causa y efecto basándose únicamente en el coeficiente de correlación. En general hay que tener en cuenta que puede existir una tercera variable escondida que puede producir una correlación que, en muchos casos, puede no tener sentido.