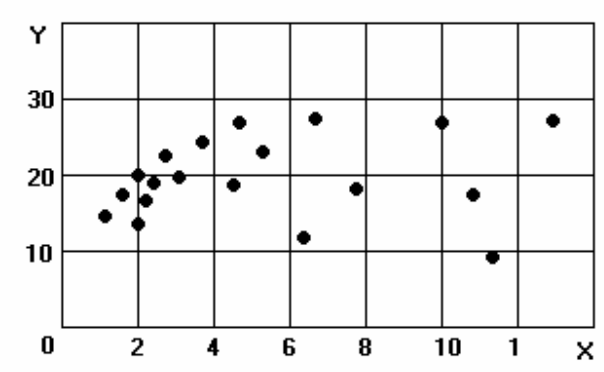
CORRELACION Y REGRESION

Jorge Galbiati Riesco

Se dispone de una muestra de observaciones formadas por pares de variables:

$(x_1, y_1), (x_2, y_2), .. , (x_n, y_n)$



A través de esta muestra, se desea estudiar la relación existente entre las dos variables X e Y. Es posible representar estas observaciones mediante un gráfico de dispersión, como el siguiente

También se puede expresar el grado de asociación mediante algunos indicadores, que se verán a continuación.

MEDIDAS DE ASOCIACION DE VARIABLES

Covarianza entre las variables *X* e *Y*. Es una medida de la variación conjunta. Se define como

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} S_{xy}$$

Puede tomar valores positivos o negativos.

Positivo, significa que ambas variables tienden a variar de la misma forma, hay una asociación positiva.

Negativo, significa que si una aumenta, la otra tiende a disminuir, y vice versa.

Covarianza cercana a cero indica que no hay asociación entre las variables.

Ejemplo 1

EL CLUB DE SALUD

Datos correspondientes a 20 empleados del club de salud de la empresa ABC.

X pulsasiones or minuto en reposo

Y tiempo en correr 1 milla (seg)

Fuente: S. Chatterjee - A. Hadi: " Sentivity Analysis in Linear Regression"

Empleado	x	y
1	67	481
2	52	292
3	56	357
4	66	396
5	65	345
6	80	469
7	77	425
8	65	393
9	68	346
10	66	401
11	70	267
12	59	368
13	58	295
14	52	391
15	64	487
16	72	481
17	57	374
18	59	367
19	70	469
20	63	252
sumas	1286	7656
promedios	64.30	382.80

Los promedios de PULSACIONES y de TIEMPO son 64.30 (puls/min) y 382.80 (seg), respectivamente.

Calcularemos de la covarianza entre estas dos variables:

Empleado	x-64.30	y-382.80	Producto
1	2.7	98.2	265.14
2	-12.3	-90.8	1116.84
3	-8.3	-25.8	214.14
4	1.7	13.2	22.44
5	0.7	-37.8	-26.46
6	15.7	86.2	1353.34
7	12.7	42.2	535.94
8	0.7	10.2	7.14
9	3.7	-36.8	-136.16
10	1.7	18.2	30.94
11	5.7	-115.8	-660.06
12	-5.3	-14.8	78.44
13	-6.3	-87.8	553.14
14	-12.3	8.2	-100.86
15	-0.3	104.2	-31.26
16	7.7	98.2	756.14
17	-7.3	-8.8	64.24
18	-5.3	-15.8	83.74
19	5.7	86.2	491.34
20	-1.3	-130.8	170.04
sumas	0.0000	0.0000	4788.20
promedios	0.0000	0.0000	239.410

La covarianza entre las variables PULSACIONES y TIEMPO es

$$\text{cov}(X,Y) = \mathbf{239.41}$$

Coeficiente de correlación lineal.

La covariaza tiene el inconveniente de que su valor no es acotado, por lo que, a partir de él es dificil juzgar si es grande o pequeña.

Se define el coeficiente de correlación, o simplemente correlación, que es una medida de asociación lineal independiente de las unidades de medida.

Es igual a la covarianza dividida por las desviaciones estándar:

$$corr(X,Y) = \frac{\text{cov}(X,Y)}{dsX * dsY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

El valor de la correlación entre cualquier par de variables es un número entre -1 y 1. n valor alto de correlación no indica que existe alguna relación de causa-efecto entre las variables.

Ejemplo 1 (continuación)

Se deben calcular las desviaciones estándar.

Para ello se deben elevar al cuadrado las observaciones centradas y promediar, obteniéndose las varianzas.

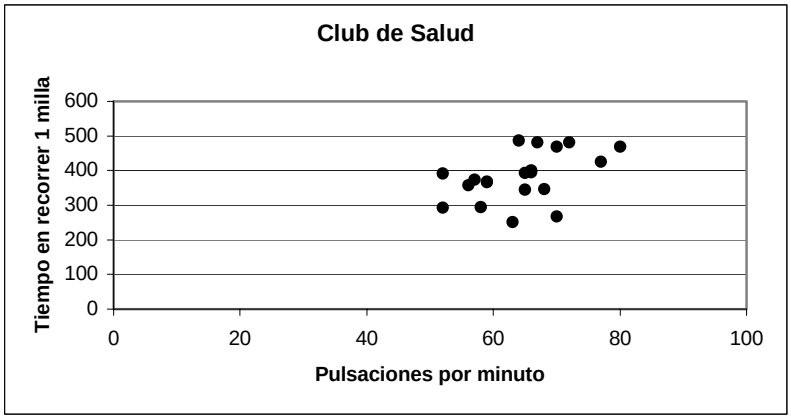
Las desviaciones standard son las raíces cuadradas de éstas.

Empleado	(x-64.30)2	(y-382.80)2
1	7.29	9643.24
2	151.29	8244.64
3	68.89	665.64
4	2.89	174.24
5	0.49	1428.84
6	246.49	7430.44
7	161.29	1780.84
8	0.49	104.04
9	13.69	1354.24
10	2.89	331.24
11	32.49	13409.64
12	28.09	219.04
13	39.69	7708.84
14	151.29	67.24
15	0.09	10857.64
16	59.29	9643.24
17	53.29	77.44
18	28.09	249.64
19	32.49	7430.44
20	1.69	17108.64
sumas	1082.20	97929.20
promedios (varianzas)	54.110	4896.460
Desv. estándar	7.356	69.975

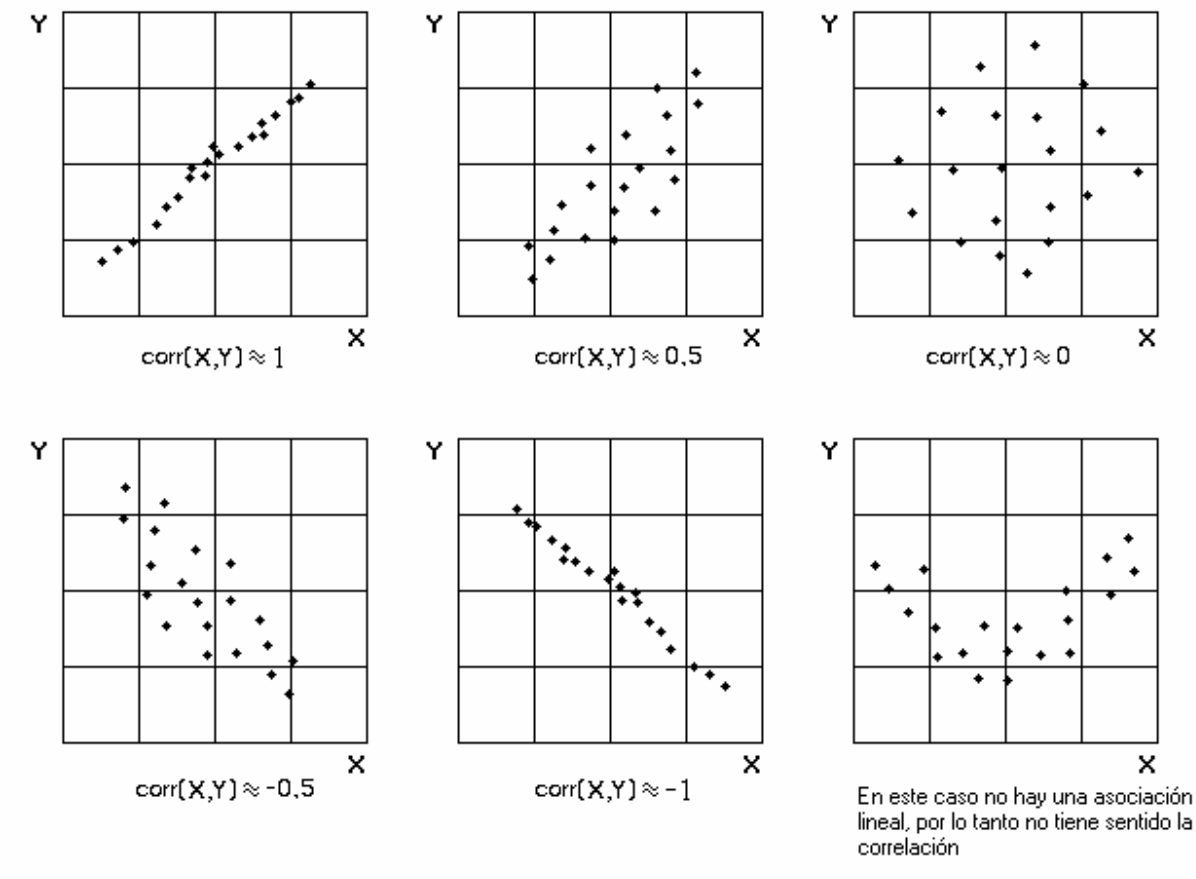
Para obtener el coeficiente de correlación se debe vividir la covarianza por el producto de ambas desviaciones estándar:

$$corr(X,Y) = 239.410/(7.356*69.975) = \mathbf{0.465}$$

El siguiente es un gráfico de dispersión que muestra estos datos.



La interpretación del coeficiente de correlación puede ilustrarse mediante los siguientes gráficos.



REGRESION LINEAL SIMPLE

Ahora asumiremos que si hay una relación de causalidad de la variable X (causa) hacia la variable Y (efecto). Además, se sabe que esa relación es de tipo lineal, dentro del rango de los datos.

Estableceremos un modelo para explicar la cai¡usa (Y) en términos del efecto (X), del tipo siguiente:

$$Y_i = a + bX_i + e_i \qquad \text{para } i = 1,2,\ldots, n$$

en que *a* y *b* son dos cantidades fijas (*parámetros* del modelo) y los e_i son cantidades aleatorias que representan las diferencias entre lo que postula el modelo $a + bx$ y lo que realmente se observa, *y*.

Por esa razón a los *e* los llamaremos "errores" o "errores aleatorios". Se asume que tienen valor esperado 0 y desviación standard común σ.

Ejemplo 2

Venta de automóviles.

Se piensa que si aumenta el porcentaje de comisión pagada al vendedor de automóviles, aumenta la venta.

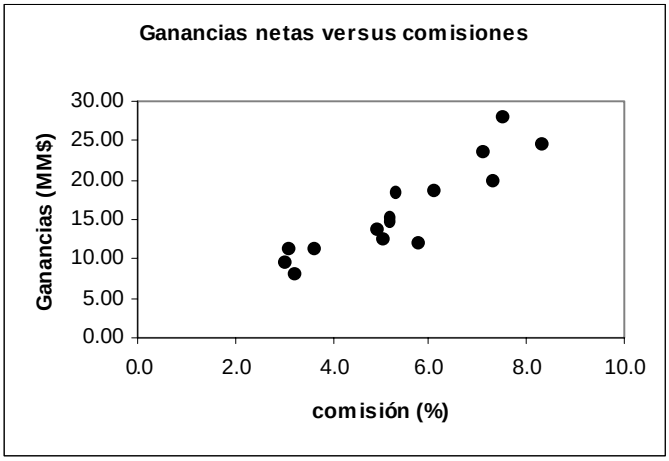
Estudio sobre 15 concesionarios similares:

X Comisión pagada por ventas (%)

Y Ganancias netas por ventas, en un mes determinado (Millones de $)

Concesionario	X	Y
1	3.6	11.28
2	5.2	14.74
3	5.3	18.46
4	7.3	20.01
5	5.0	12.43
6	5.2	15.37
7	3.0	9.59
8	3.1	11.26
9	3.2	8.05
10	7.5	27.91
11	8.3	24.62
12	6.1	18.80
13	4.9	13.87
14	5.8	12.11
15	7.1	23.68
sumas	80.600	242.180
promedios	5.373	16.145

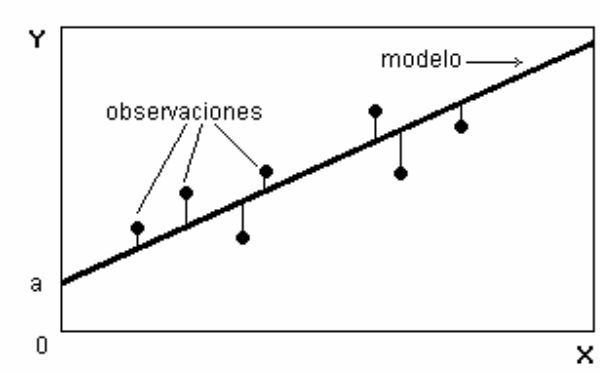
Representación de los datos en un gráfico de dispersión:



Se puede apreciar la relación lineal existente entre ambas variables observadas.

Nuestro problema es estimar los parámetros *a*, *b* y σ^2 para poder identificar el modelo.

Para estimar *a* y *b* se utiliza el método de *Mínimos cuadrados*, que consiste en encontrar aquellos valores de *a* y de *b* que hagan mínima la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta que representa el modelo, en el sentido vertical.



En la figura, son los cuadrados de los segmentos verticales cuya suma de cuadrados se debe minimizar, para determinar *a* y *b.* Estos segmentos representan los errores *e* del modelo. *b* se llama *pendiente* de la recta que representa los datos y *a* se llama

intercepto sobre el eje vertical.

La solución está dada por las siguientes fórmulas:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b\,\bar{x}$$

Ejemplo 2 (continuación)

Calculamos los promedios de ambas variables y se las restamos a los valores.

Promedio de las X: 5.373

Promedio de las Y: 16.145

Cálculo de de las desviaciones respecto de las medias, sus cuadrados y productos:

Concesionario	X	Y	(x-xmed)2	(y-ymed)2	Prod
1	3.6	11.3	3.145	23.671	8.628
2	5.2	14.7	0.030	1.975	0.244
3	5.3	18.5	0.005	5.358	-0.170
4	7.3	20.0	3.712	14.936	7.446
5	5.0	12.4	0.139	13.804	1.387
6	5.2	15.4	0.030	0.601	0.134
7	3.0	9.6	5.633	42.972	15.558
8	3.1	11.3	5.168	23.866	11.106
9	3.2	8.1	4.723	65.534	17.594
10	7.5	27.9	4.523	138.407	25.020
11	8.3	24.6	8.565	71.820	24.803
12	6.1	18.8	0.528	7.047	1.929
13	4.9	13.9	0.224	5.177	1.077
14	5.8	12.1	0.182	16.284	-1.722
15	7.1	23.7	2.981	56.771	13.010
sumas	80.600	242.180	39.589	488.225	126.043
promedios	5.373	16.145	1.979	24.411	6.302

Entonces, utilizando las fórmulas dadas más arriba, obtenemos los valores de los parámetros del modelo de regresión:

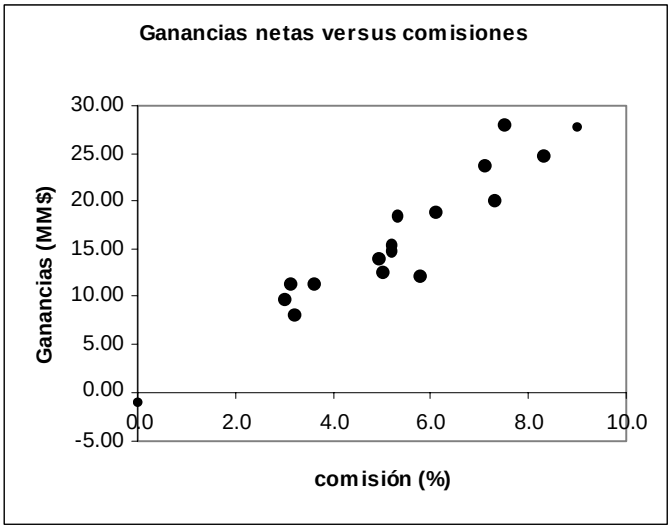
b=126.043/39.589=**3.184**

a=16.145-3.184*5.373=**-0.962**

El modelo, para estos datos, entonces, es

$$Y_i = -0.96 + 3{,}18X_i + e_i \qquad \text{para i=1,2,.. 15}$$

Representa una recta, cuyo intercepto con el eje vertical es -0.96, y su pendiente es 3.18, o sea, si el porcentaje de comisión *X* aumenta en 1%, la ganancia neta *Y* aumenta en 3.18 Millones de pesos.



VALORES AJUSTADOS AL MODELO.

El modelo de regresión lineal se puede utilizar para obtener valores de *Y* ajustados al modelo, Los valores puntuales se obtienen mediante la fórmula

$$Y_i = a + bX_i$$

en que *a* y *b* son los valores estimados por el procedimiento indicado anteriormente, y X_i toma los valores de la muestra. Los puntos que representan estos valores en el gráfico de dispersión, yacen sobre la recta.

Ejemplo 2 (continuación)

La tabla siguiente contiene los valores de *Y* ajustados , para cada valor de *X*, además de los valores de *Y* observados, a modo de comparación. Los ajustados se obtienen por la fórmula

$$Y_i = -0.96 + 3.18X_i$$

Concesionario	X	Y	Yajust	diferencia
1	3.6	11.28	10.499	0.781
2	5.2	14.74	15.593	-0.853
3	5.3	18.46	15.912	2.548
4	7.3	20.01	22.279	-2.269
5	5	12.43	14.957	-2.527
6	5.2	15.37	15.593	-0.223
7	3	9.59	8.589	1.001
8	3.1	11.26	8.908	2.352
9	3.2	8.05	9.226	-1.176
10	7.5	27.91	22.916	4.994
11	8.3	24.62	25.463	-0.843
12	6.1	18.8	18.459	0.341
13	4.9	13.87	14.638	-0.768
14	5.8	12.11	17.504	-5.394
15	7.1	23.68	21.643	2.037
sumas	80.600	242.180	242.180	0.000
promedios	5.373	16.145	16.145	0.000

Se puede observar que el promedio de los valores ajustados es igual al promedio de los valores observados, y que el promedio de las diferencias es cero.

La raíz cuadrada del promedio de los cuadrados de las diferencias entre los valores observados y ajustados, es una estimación de la varianza del error, σ^2. En el ejemplo, la suma de las diferencias al cuadrado es 86.933, luego la estimación de la desviación estándar del error es igual a

$$\sigma = \sqrt{86.933/15} = \sqrt{5.796} = \mathbf{2.41}$$ Millones de pesos

<u>Coeficiente de determinación</u>. Es una medida de bondad de ajuste del modelos de regresión lineal a los datos.
Es deseable que los valores de *Y* ajustados al modelo, sean lo más parecidos posible a los valores observados. Una medida de lo parecido que son, es el coeficiente de correlación.

Se define el *coeficiente de determinación, R^2,* como el cuadrado del coeficiente de correlación entre los valores de *Y* observados y los valores de *Y* ajustados. Sin embargo se puede demostrar que es igual a la siguiente expresión:

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}$$

El rango de R^2 es entre 0, cero ajuste, hasta 1, ajuste perfecto (cuando los puntos aparecen en un línea recta).

Ejemplo 2 (continuación)

Más arriba se calcularos las sumas de cuadrados y de productos, y dieron los siguientes valores:

$S_{xx} = 39.6$, $S_{yy} = 488.3$, $S_{xy} = 126.1$

Entonces el coeficiente de determinación es

$$R^2 = \frac{(126.043)^2}{39.589 * 488.225} = \mathbf{0.822}$$

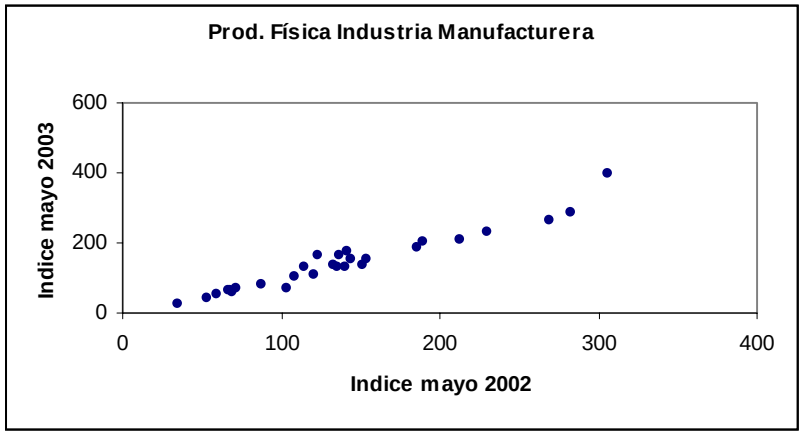
que señala que el ajuste del modelo a los datos es bueno.

Ejemplo 3

Los datos siguientes corresponde al Indice de Producción Física de la Industria Manufacturera, por agrupación, de los meses de mayo de 2002 y mayo de 2003, entregado por el Instituto Nacional de Estadísticas. Es un índice cuya base 100 es el promedio de producción de cada agrupación, en el año 1989.

Agrupaciones	Mayo 02	Mayo 03
Fabricac. de productos alimenticios	140.2	133.5
Industrias de bebidas	134.6	133.7
Industria del tabaco	151.1	140.5
Fabricac. de textiles	70.9	70.3
Fabricac. prendas de vestir, excepto calzado	34.7	30.5
Industria del cuero; produc. de cuero y sucedáneos	59.3	56.7
Fabricac. de calzado, exc. de caucho o plástico	52.6	45.3
Industria de madera y sus productos exc. muebles	132.3	141.6
Fabricac. de muebles y accesorios, exc. metálicos	114.0	132.4
Fabricac. de papel y productos de papel	189.5	205.3
Imprentas, editoriales e industrias conexas	107.5	108.0
Fabricac. de sustancias químicas industriales	229.4	231.4
Fabricac. de otros productos químicos	212.4	209.6
Refinerías de petróleo	136.0	165.2
Fabricac. prod. derivados de petróleo y carbón	143.2	156.2
Fabricac. de productos de caucho	141.4	177.4
Fabricac. de productos plásticos	305.8	399.7
Fabricac. de objetos de loza y porcelana	68.2	61.1
Fabricac. de vidrio y productos de vidrio	268.6	266.4
Fabricac. otros productos minerales no metálicos	185.6	186.5
Industrias básicas de hierro y acero	123.1	167.1
Industrias básicas de metales no ferrosos	119.8	108.7
Fabricac. prod. metálicos exc. maquinaria y equipo	153.6	153.5
Construcción de maquinaria, exc. la eléctrica	282.5	289.7
Construcción máq., aparatos y acces. eléctricos	87.0	83.0
Construcción de material de transporte	103.4	73.4
Fabricac. equipo profesional y artículos oftálmicos	67.7	64.1
Otras industrias manufactureras	66.0	67.5

El gráfico de dispersión es el siguiente:



Cálculos parciales, en que *X* es el índice mayo 2002, *Y* el índice mayo 2003:

$$n = 28 \qquad \bar{x} = 136.6 \qquad \bar{y} = 144.9$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = 134{,}913.6$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = 187{,}813.7$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 154{,}350.8$$

Estimación de los parámetros del modelo:

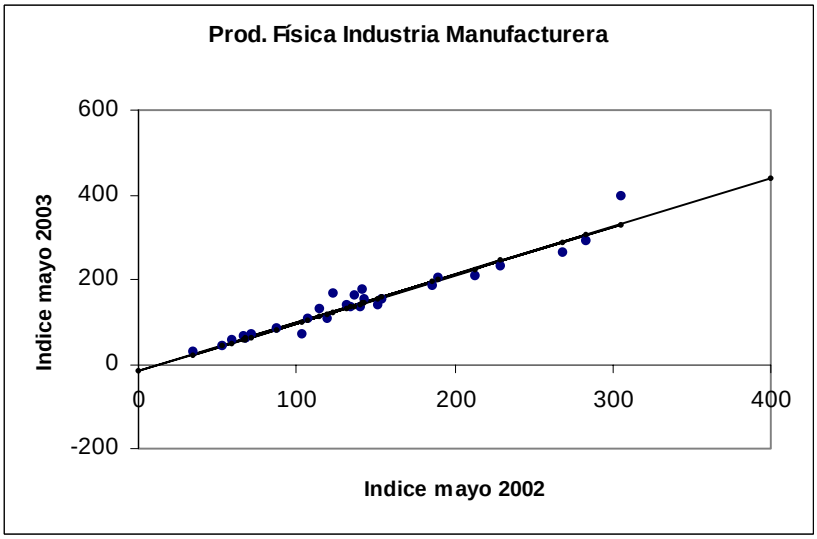
$$b = \frac{S_{xy}}{S_{xx}} = \frac{154{,}350.8}{134{,}913.6} = 1.14$$

$$a = \bar{y} - b\bar{x} = -13.61$$

Bondad de ajuste:

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{(154{,}350.8)^2}{(134{,}913.6) * (187{,}350.8)} = 0.940$$

que indica un muy buen ajuste. El siguiente gráfico muestra de recta de regresión estimada:



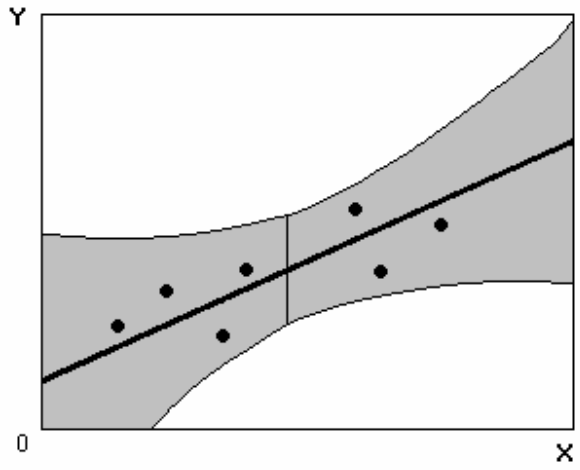
Predicción por bandas de confianza.

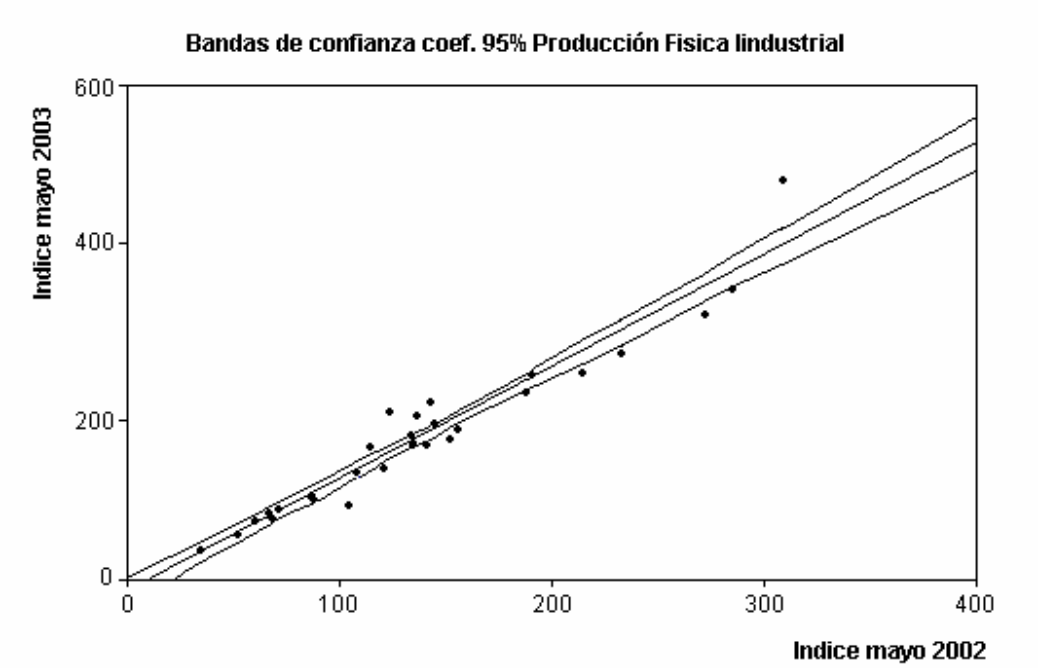
Se pueden hacer predicciones de valores Y para valores X que no están en el conjunto de observaciones, dentro o fuera de su rango, utilizando la fórmula de la regresión

lineal, con los parámetros *a* y *b* estimados.

También se pueden hacer predicciones por intervalos de confianza verticales, que tienen la ventaja de proporcionar una cuantificación del error de predicción. Los intervalos tienen la propiedad de ser de diferente ancho, según el valor de X, siendo más angostos cuando X es igual al promedio, ensanchándose a medida que nos alejamos del promedio. Cuando se sale del rango de los datos, se ensanchan más fuertemente. Esto significa que mientras más nos alejamos del centro de los valores de la variable *X*, más imprecisas serán nuestras estimaciones del valor de la variable *Y*, lo que parece razonable.

Si unimos los extremos superiores (o los inferiores) de todos los intervalos de confianza, se obtienen dos curvas con forma de hipérbola, como se muestra en la figura:





El gráfico siguiente muestra las bandas de confianza de coeficiente 95%, para el ejemplo de la producción física manufacturera.

Mientras mayor es el coeficiente de determinación R^2, más angostas son las bandas de confianza; lo mismo mientras mayor es la desviación estándar de las *X*, y lo mismo si el tamaño muestral aumenta. Y a medida que nos alejamos del promedio de las *X*, se ensanchan las bandas.

PREGUNTAS

1. Se tienen dos variables, relacionadas con las publicaciones en revistas de profesores universitarios:

X = Número de publicaciones.

Y = Número de veces que ha sido citado.

Utilizando regresión lineal, se estimó, en base a una muestra, que estas variables están relacionadas mediante el siguiente modelo lineal:

$$Y = 0.3 + 2.6\ X$$

¿Cómo se interpretan los dos parámetros de este modelo ?

2. Se tiene un conjunto de pares de datos (x,y), a los que se les estima una recta de regresión. La variable independiente es x, su rango es entre 150 y 230. Se hacen dos estimaciones de y por intervalos de confianza de coeficiente 95%, una para x=190 y otra para x=250. ¿Cuál es más precisa?

3. La relación entre el tiempo, en días, dedicado a elaborar un proyecto y el costo del proyecto se modeló mediante una regresión lineal, estimándose la siguiente expresión:

costo = 23 + 0.52 * tiempo

¿Cómo interpreta el número 23 ?
¿Cómo interpreta el número 0.52 ?

4. ¿Qué mide el coeficiente de determinación, en un modelo de regresión lineal?

5. Qué ventaja tiene el coeficiente de correlación, sobre la covarianza, como medidas de asociación entre variables?

6. Se aplicó regresión lineal para predecir la demanda de un producto, para el próximo año, utilizando los datos de seis años pasados. Interprete la siguente afirmación:

"La demanda proyectada para el próximo año será entre 855 y 955 en base a un intervalo de confianza de coeficiente 95%."

7. ¿Qué mide el coeficiente de correlación lineal de dos variables.

8. Se tienen dos variables, observadas en trabajadores de la salud:

X = años de servicio.

Y = asignaciones salariales actuales (miles de pesos).

Utilizando regresión lineal, se estimó, en base a una muestra, que estas variables están relacionadas mediante el siguiente modelo lineal:

Y = 200+ 15 X

¿Cómo se interpretan los dos parámetros de este modelo ?

9. ¿Qué mide el coeficiente de determinación, en una regresión lineal?

10. Se tienen dos variables, observadas en una muestra de estudiantes egresados de la enseñanza media:

X = promedio de notas de los cuatro años de enseñanza media.

Y = puntos PSU historia.

Utilizando regresión lineal, se estimó, en base a una muestra, que estas variables están relacionadas mediante el siguiente modelo lineal:

$$Y = 60 + 100\ X$$

¿Cómo se interpretan los dos parámetros de este modelo ?

11. Una institución ha encargado una serie de proyectos. Con los datos históricos, se quiso relacionar los montos de los proyectos con los tiempos de ejecución, obteniéndose los siguientes resultados:

Monto (M$) = 12620+ 476 x Tiempo (días) con un coeficiente de determinación R^2 = 0.86

12. Se tienen dos variables, observadas en una muestra de postulantes a un cargo profesional:

X = promedio de notas de sus años de estudio.

Y = puntaje obtenido en una prueba que se les aplicó.

Utilizando regresión lineal, se estimó, en base a una muestra, que estas variables están relacionadas mediante el siguiente modelo lineal:

$$Y = 3 + 1.5\ X$$

¿Cómo se interpretan los dos parámetros de este modelo ?

13. Se tiene un conjunto de pares de datos (x,y), a los que se les estima una recta de regresión. La variable independiente es x, su rango es entre 35 y 45. Se hacen dos estimaciones de y por intervalos de confianza de coeficiente 95%, una para x=40 y otra para x=50. ¿Cuál es más precisa?

14. El costo por días por trabajados en un proyecto está dado por la siguiente expresión:

El costo Monto (M$) = 2246+ 35 x Tiempo (días),

Que se obtuvo ajustando una regresión lineal a un conjunto de datos de proyectos ya realizados. Se obtuvo un coeficiente de determinación R^2 = 0.91.

Interprete los valores 2246, 35 y 0.91.