CORRELACIÓN Y REGRESIÓN

Raúl David Katz

Correlación y regresión

Introducción

Hasta ahora hemos visto el modo de representar la distribución de frecuencias de los datos correspondientes a una variable (distribución unidimensional). Estas gráficas permiten reconocer la forma aproximada de la distribución de dichos datos, pero no permiten establecer una relación entre los datos correspondientes a diferentes variables.

En numerosas situaciones estadísticas resulta útil reconocer (si existe) una relación entre los datos correspondientes a dos variables. Por ejemplo: ¿existe alguna relación entre la vida útil de una herramienta y su velocidad de corte?, en caso afirmativo, ¿cómo es dicha relación?

Para captar la relación entre los datos correspondientes a dos variables cuantitativas resulta de utilidad la construcción de un diagrama de dispersión, que pasamos a tratar.

Diagrama de dispersión

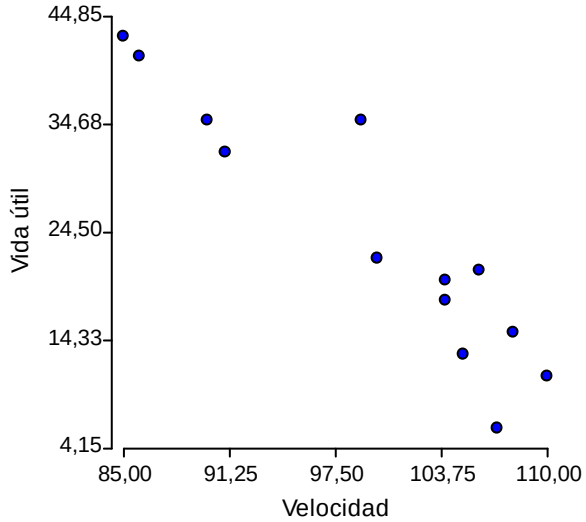
Los siguientes datos corresponden a la vida útil y a la velocidad de corte de una herramienta.

Velocidad de corte:	86	104	100	85	107	104	106	99	90	110	108	105	91
Vida útil:	41	18	22	43	6	20	21	35	35	11	15	13	32

Dados los pares de valores (x , y) entre los cuales se desea estudiar la relación, se representan a los mismos en un sistema de ejes cartesianos ortogonales, seleccionando una escala de modo que la lectura del diagrama resulte más fácil (generalmente se considera la diferencia entre el máximo y mínimo de cada variable y a esa diferencia se le asigna la misma longitud en cada eje). “La nube de puntos” que se obtiene cuando representamos los valores (x , y) se denomina **diagrama de dispersión**.

Si una de las variables se puede considerar como la variable que causa, o explica los cambios observados en la otra, a esa variable se la denomina **explicativa** y se la representa sobre el eje x. En este caso, a la otra variable se la denomina **variable respuesta** y se la representa sobre el eje y. Si no se quiere distinguir entre variable explicativa y variable respuesta, cualquiera de las dos puede representarse en el eje de las abscisas.

El siguiente diagrama de dispersión corresponde a los datos sobre la velocidad de corte y vida útil de una herramienta.



Para interpretar un diagrama de dispersión es necesario reconocer primero su aspecto general que debe revelar la dirección, la forma e intensidad de la relación entre las variables.

A partir del diagrama de dispersión se percibe una relación con tendencia lineal y pendiente negativa. Valores superiores al promedio de la velocidad de corte están en correspondencia con valores que son inferiores al promedio de la vida útil y valores inferiores al promedio de velocidad de corte están en correspondencia con valores que superan el promedio de la vida útil. En este caso decimos que las variables están **relacionadas negativamente**.

Asimismo decimos que dos variables están **relacionadas positivamente** cuando los valores que se representan sobre el eje x y que superan su promedio tienden a estar en correspondencia con los valores que se representan sobre el eje y superan su promedio, y los valores inferiores al promedio de cada variable también tienden a ocurrir conjuntamente.

Propuesta

Determina qué proporción de los datos se encuentra en cada cuadrante cuando el origen de coordenadas del sistema de referencia se toma en el punto de abscisa igual a la media aritmética de las velocidades de corte y ordenada igual a la media aritmética de las vidas útiles.

Propuesta

¿Considera que la relación entre la velocidad al corte de una herramienta y su vida útil tiende a una relación lineal?

Propuesta

La siguiente tabla muestra los datos correspondientes a 16 meses. La variable respuesta (y) es el promedio de los consumos diarios de gas durante el mes, medidos en m^3. La variable explicativa (x) es el promedio de los grados de calefacción demandada por día, durante el mes, medidos en grados Celsius. Por ejemplo, si la temperatura es de 1ºC entonces se demandan 17.5ºC de calefacción, para alcanzar un temperatura deseada de 18.5ºC

Temperaturas:	13.3	28.3	23.9	18.3	14.4	7.2	2.2	0	0	0.5	3.3
Consumo:	17.6	30.5	24.9	21.0	14.8	11.2	4.8	3.4	3.4	3.4	5.9
Temperaturas:		6.7	16.7	17.8	28.9	16.7					
Consumo:		-8.7	17.9	20.2	30.8	19.3					

a) Realiza el diagrama de dispersión.
b) ¿Considera que la relación entre las variables es lineal?
c) La relación entre las variables es positiva o negativa?

Si existe una "fuerte" relación entre dos variables, el conocimiento de una de ellas permite predecir el comportamiento de la otra, pero cuando la relación es débil, la información de una de las variables no ayuda demasiado a extraer conclusiones sobre la otra.

Propuesta

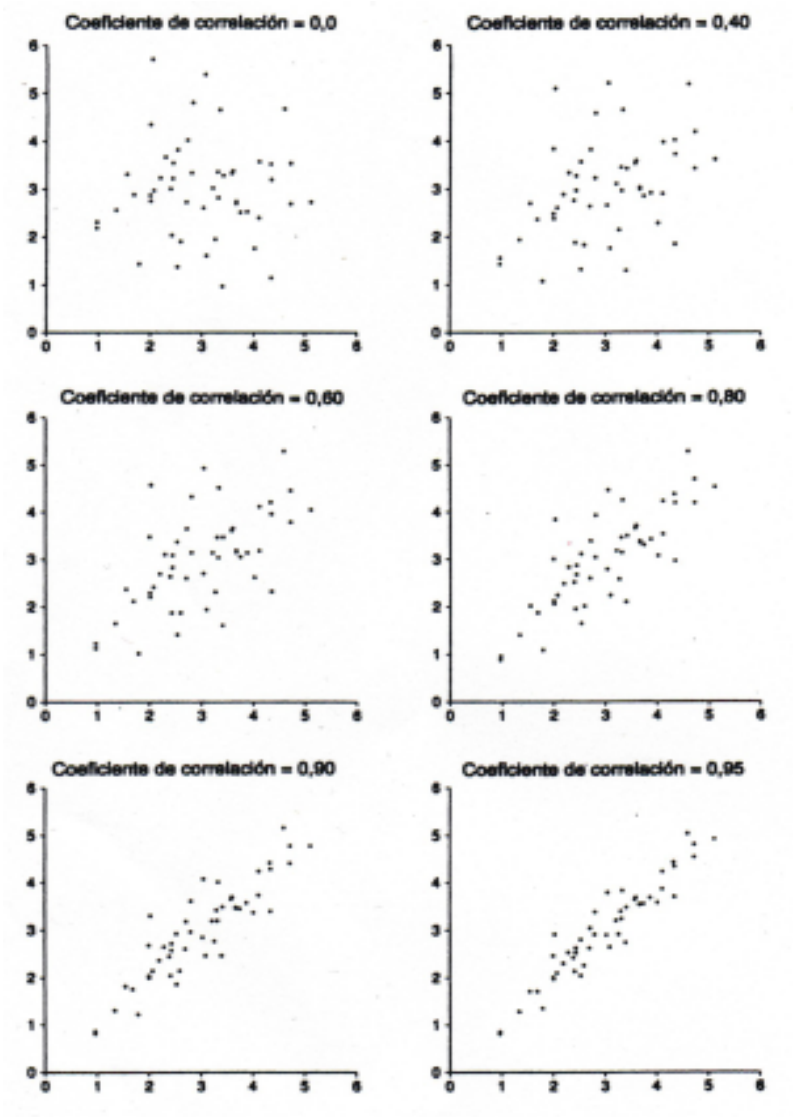
Los siguientes datos corresponden al consumo de combustible de un auto a medida que aumenta su velocidad.

Velocidad (km. /h):	10	20	30	40	50	60	70	80	90
Consumo (litros/100 km.)	21	13	10	8	7	5.90	6.30	6.95	7.57
Velocidad (km. /h):	100	120	130	140	150				
Consumo (litros/100 km.)	8.27	9.87	10.79	11.77	12.83				

a) Dibuja un diagrama de dispersión. ¿Cuál consideras que es la variable explicativa?
b) Describe la forma de la relación.

El coeficiente de correlación

Cuando el diagrama de dispersión muestra una nube de puntos muy agrupados en torno a una recta, se dice que existe una fuerte relación lineal entre las dos variables
El coeficiente de correlación es una medida de esa relación lineal o intensidad de la agrupación alrededor de una recta.



Definición y cálculo del coeficiente de correlación

Si (x_k , y_k) con k=1,2,.....N son las coordenadas de los puntos de una nube (considerados como una población de tamaño N), el coeficiente de correlación, que notamos con r, se define como:

$$r = \frac{1}{N} \sum_{i}^{N} \frac{x_i - \mu_x}{\sigma_x} \frac{y_i - \mu_y}{\sigma_y}$$

donde μ_x , σ_x , μ_y y σ_y representan la media y desviación estándar de los valores x_k e y_k respectivamente.

Si (x_k , y_k) con k=1,2,.....n son las coordenadas de los puntos de una nube (considerados como una muestra, de tamaño n), el coeficiente de correlación se define como:

$$r = \frac{1}{n-1} \sum_{1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

(El valor del coeficiente de correlación calculado sobre un conjunto finito de datos es independiente de considerar a estos como una muestra o una población)

Para calcular el coeficiente de correlación hay que estandarizar los observaciones de cada variable, es decir, a cada valor de la variable hay que restarle la media de esos valores y dividirlo por la desviación estándar de esos valores. El coeficiente de correlación es la media de los productos estandarizados de cada par de observaciones.

Un ejemplo

Calcular el coeficiente de correlación para los datos de la siguiente tabla

x	1	3	4	5	7
y	5	9	7	1	13

La media y desviación estándar de los valores de x son 4 y 2 respectivamente.
La media y desviación estándar de los valores de y son 7 y 4 respectivamente.

Luego:

x	y	x (estandarizada)	y (estandarizada)	producto
1	5	-1.5	-0.5	0.75
3	9	-0.5	0.5	-0.25
4	7	0.0	0.0	0.0
5	1	0.5	-1.5	-0.75
7	13	1.5	1.5	2.25

El valor del coeficiente de correlación es igual a 0.4, que es la media de los valores de la última columna.
Una manera práctica de determinar el coeficiente de correlación es la siguiente:

$\frac{1/N\sum x_i.y_i - \mu_x.\mu_y}{\sigma_x\sigma_y}$ donde N es la cantidad de pares ordenados de datos observados y μ_x, μ_y, σ_x y σ_y son las medias y desviaciones estándares de los valores de x e y respectivamente.

Propiedades del coeficiente de correlación

- El coeficiente de correlación asume valores en el intervalo [-1 , 1].

Cuando r = 1 todos los puntos de la nube se encuentran sobre una recta con pendiente positiva.
Cuando r = -1 todos los puntos de la nube se encuentran sobre una recta con pendiente negativa.
En cualquiera de estos dos casos, conociendo el valor de x, se puede predecir con certeza el valor que asume y.

- El coeficiente de correlación es un número adimensional (sin unidades de medida).
- El coeficiente de correlación entre dos variables no se modifica cuando:
  a) se suma un mismo número a todos los valores de una variable.
  b) se multiplica todos los valores de una variable por el mismo número positivo.

Observaciones:

Cuando r = 0 no hay relación lineal entre las variables, lo que no excluye la posibilidad de que exista otro tipo de relación.

El coeficiente de correlación mide el grado de relación lineal entre las variables, pero esa relación no es necesariamente de causa – efecto. La correlación ignora la distinción entre variables explicativas y variables respuesta.

Un ejemplo:
Se ha determinado que la cantidad de televisores que se venden por año y la cantidad de personas que padecen trastornos mentales (en el correspondiente año) guardan una fuerte relación lineal.
No son los programas de televisión que causan los trastornos mentales (podría ser). Existe un "factor de confusión", que es el aumento de la población, que genera un aumento tanto en la venta de televisores como en los que padecen de trastornos mentales.

Regresión

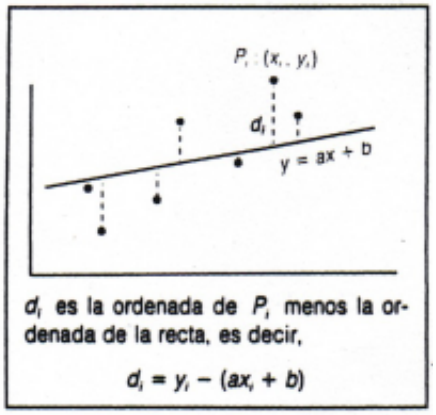
El coeficiente de correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas, sin realizar la distinción entre variable explicativa y variable respuesta.
Si un diagrama de dispersión muestra una relación lineal, nos gustaría dibujar una recta a través de la nube de puntos. A tal fin se suele utilizar el método de los mínimos cuadrados, que consiste en determinar una recta tal que la suma de los cuadrados de las distancias verticales d_i (marcadas en la figura) resulten mínimas.

¿Cuál es la recta $r: y = ax + b$ para la cual

$$\Sigma d_i^2 = \Sigma [y_i - (ax_i + b)]^2$$

es mínima?

Es decir, ¿cuánto deben valer a y b para que la suma anterior sea lo menor posible?



d_i es la ordenada de P_i menos la ordenada de la recta, es decir,

$$d_i = y_i - (ax_i + b)$$

Sea $y = a\,x + b$ la ecuación de la recta que se busca determinar con el criterio establecido.

Se demuestra que:

$a = r\,\frac{s_y}{s_x}$ y $b = \overline{y} - a\overline{x}$ de modo que la ecuación de la recta de regresión es:

$$y - \overline{y} = r\,\frac{s_y}{s_x}\,(x - \overline{x})$$

La variable que se representa sobre el eje de las abscisas se considera la variable explicativa y la variable que se representa sobre el eje de las ordenadas se considera la variable respuesta.
La recta de regresión describe los cambios en los valores medios de la variable respuesta (Y) a medida que cambian los valores de la variable explicativa (X). A esta recta se la denomina recta de regresión de Y sobre X.
De forma similar se obtiene la ecuación de la recta de regresión de X sobre Y, (ahora se considera a la variable Y como variable explicativa y a X como variable respuesta) a través de:

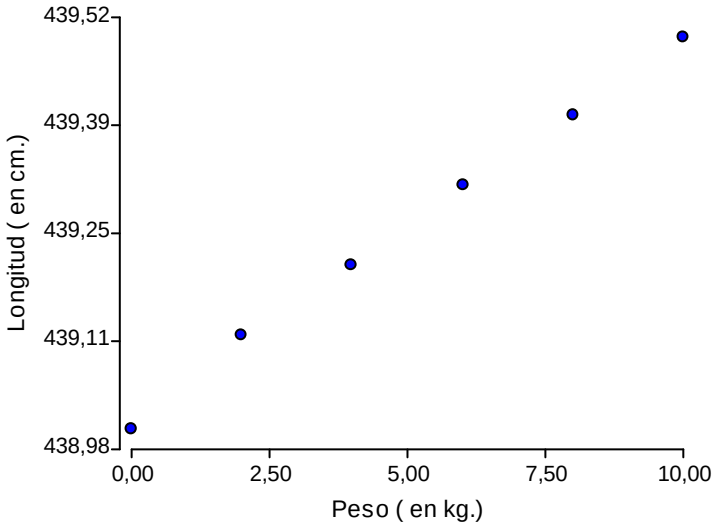
$$x - \overline{x} = r\,\frac{s_x}{s_y}(y - \overline{y})$$

Un ejemplo

Robert Hooke (Inglaterra, 1653-1703) estudió la relación entre la longitud de un resorte y el peso que de él pendía. Cuando aumentaba el peso, el resorte se alargaba más.
La siguiente tabla muestra los resultados de un experimento en el cual se colocaron diferentes pesos de una cuerda de piano.

Pesos:	0	2	4	6	8	10
Longitud de la cuerda:	439	439,12	439,21	439,31	439,40	439,50

El siguiente gráfico muestra que existe “una fuerte relacional lineal positiva” entre las variables peso y longitud que adquiere la cuerda con cada uno de ellos.



Si calculamos el coeficiente de correlación obtenemos un valor cercano a 1

El peso medio y la desviación estándar de los pesos son 5 y 3.74 respectivamente.
La longitud media y la desviación estándar de las longitudes son 439.26 y 0.18 respectivamente.
Por lo tanto la ecuación de la recta de regresión es :

$$y - 439.25 = 1.\frac{0.18}{3.74}(x - 5) \ ,$$

En consecuencia (redondeando los valores) se obtiene la ecuación $y = 0.05x + 439{,}01$

Observaciones:

- La estimación de la longitud del resorte cuando no tiene ningún peso colgado es de 439.01 cm.
- Cada kilogramo produce un alargamiento estimado en 0.05 cm.
- Para x = 9 Kg. la estimación de la longitud de alargamiento es de 439.46 Kg.

Propuesta

Halla la ecuación de la recta de regresión para los datos correspondientes a la velocidad de corte de una herramienta y la vida útil de la misma:

a) considerando a la velocidad de corte como variable explicativa,
b) Estima la vida útil media cuando la velocidad de corte es igual a 100.

Propuesta

a) Halla la ecuación de la recta de regresión para los datos correspondientes a las temperaturas medias diarias y consumo de gas
b) Estima el consumo medio de gas cuando la temperatura media diaria es igual a 20.
c) ¿Cuál es el aumento del consumo medio de gas ante un incremento unitario de la temperatura media diaria?

Bibliografía

1) de Guzmán, M. Colera, J.(1989) Matemáticas II. Anaya. Barcelona.
2) Freedman, D. y otros.(1993). Estadística. Antoni Bosch. Barcelona.
3) Moore, D. (1995). Estadística aplicada básica. Antoni Bosch. Barcelona.