4. ANÁLISIS DE REGRESIÓN SIMPLE INTRODUCCIÓN

En primer lugar estudiaremos ciertas medidas de variabilidad muy importantes, como lo es la variabilidad conjunta de dos variables, lo que se conoce como covarianza y coeficiente de correlación lineal. Luego, nos abocaremos a estudiar qué pasa cuando tenemos dos variables y deseamos efectuar inferencias acerca de los cambios que se producen en una de ellas cuando cambia la otra.

Para describir la forma de la relación que liga a estas dos variables utilizaremos los llamados modelos de regresión.

Por ejemplo, supongamos que la variable x define el precio anual del trigo y la variable y las hectáreas sembradas anualmente en la provincia de Córdoba. Si descubrimos la relación que liga al precio del trigo con la cantidad de hectáreas sembradas podremos predecir cuántas hectáreas se sembrarán cuando se producen variaciones en los precios del bien. Si descubrimos la relación que liga al precio del trigo con la cantidad de hectáreas sembradas podremos predecir cuántas hectáreas se sembrarán cuando se producen variaciones en los precios del bien.

Asimismo, cuando hablemos del grado de la relación que liga a dos variables también utilizaremos complementariamente el análisis de correlación.

Por ejemplo, si decimos que correlacionando las variables peso y altura de jóvenes obtuvimos un valor r = 0,80, estamos expresando que existe una fuerte correlación positiva entre ambas variables. En otras palabras, estamos diciendo que cuando aumenta el peso, aumenta la altura en la misma dirección y viceversa.

4.1. Covarianza y coeficiente de correlación lineal

Hasta ahora nos hemos limitado al estudio de datos univariables, es decir, aquellos que surgen de obtener una sola medición por unidad experimental.

De esta manera, se origina una sola variable aleatoria. Sin embargo, puede suceder que en lugar de una medición se efectúen dos o más mediciones de cada elemento de la muestra.

Por ejemplo, podemos estar interesados en observar la altura (x) y el peso (y) de todas las mujeres de una edad determinada.

En este caso, la muestra no respondería a una población univariada sino que, al realizarse dos mediciones, la muestra provendría de una población, bivariada.

Cuando se analizan datos de esta última población, el problema princIpal suele ser el de descubrir y medir la asociación o variación conjunta de ambas variables.

Siguiendo con el ejemplo, podemos observar si las mujeres más altas son en realidad las más pesadas.

En síntesis, cuando se extrae una muestra de n pares de valores x e y, interesa tener un indicador del grado de intensidad de la relación entre las dos variables que sea, a su vez, independiente de sus respectivas escalas de medición.

A este indicador lo llamaremos coeficiente de correlación lineal entre x e y (se denomina lineal cuando la forma de la relación que liga a ambas variables corresponde a la ecuación de una recta).

Para calcular el coeficiente de correlación lineal necesitamos establecer una medida de la variabilidad conjunta de ambas variables. Esta medida recibe el nombre de covarianza.

La covarianza tiene el mismo significado que la varianza pero, como su nombre lo indica, mide la covariabilidad de dos variables.

La fórmula de cálculo de la covarianza es la siguiente:

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

siendo x e y las dos variables de las cuales se va a estudiar la variabilidad conjunta.

Veamos como aplicamos la fórmula por medio de un ejemplo.

Supongamos que se extrae una muestra de farmacias de la ciudad de Córdoba. Los siguientes datos corresponden a los costos (x_i) y ventas (y_i) de las 12 farmacias seleccionadas en la muestra.

Costos (x_i)	Ventas(y_i)
11	19
10	15
14	20
13	14
12	16
20	33
21	32
15	18
22	29
18	22
19	23
16	20

Se tienen, entonces, dos variables (x_i e y_i) y se quiere estudiar la variabilidad conjunta de estas variables.

Si volvemos a mirar la fórmula, vemos que debemos calcular las medias de cada una de las variables.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 15{,}92 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 21{,}75$$

Para simplificar el cálculo de la covarianza utilizaremos una tabla conveniente:

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\ (y_i - \bar{y})$
11	19	-4,92	-2,75	13,53
10	15	-5,92	-6,75	39,96
14	20	-1,92	-1,75	3,36
13	14	-2,92	-7,75	22,63
12	16	-3,92	-5,75	22,54
20	33	4,08	11,25	45,90
21	32	5,08	10,25	52,07
15	18	-0,92	-3,75	3,45
22	29	6,08	7,25	44,08
18	22	2,08	0,25	0,52
19	23	3,08	1,25	3,85
16	20	0,08	-1,75	-0,14
				251,75

Luego,

$$\text{cov}(x,y) = \frac{251,75}{12} = 20,9792$$

Así como la covarianza mide la variabilidad conjunta de x e y, el coeficiente de correlación lineal mide el grado de relación que existe entre estas variables.

El coeficiente de correlación lineal varía entre - 1 y 1. Cuando se va acercando a 1 significa que las dos variables están correlacionadas positivamente, o sea, al variar una de ellas, la otra varía en la misma proporción y sentido.

Cuando el coeficiente se acerca a - 1, significa que las dos variables están correlacionadas negativamente: al variar una de ellas, la otra varía en la misma proporción pero en sentido inverso.

Cuando se acerca al valor 0, significa que las variables en estudio no están correlacionadas.

La fórmula para calcular el coeficiente de correlación lineal es la siguiente:

$$r_{x,y} = \frac{\text{cov}(x,y)}{D(x)\,D(y)}$$

En palabras, el coeficiente de correlación es igual al cociente entre la covarianza y el producto de las desviaciones estándar de las dos variables.

En consecuencia, para calcular el coeficiente de correlación lineal nos faltan calcular las desviaciones estándar de x e y.

$$D(x)=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}=3{,}8828$$

$$D(y)=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}=6{,}1390$$

Entonces:

$$r=\frac{20{,}9792}{3{,}8828 \times 6{,}1390}=\frac{20{,}9792}{23{,}8365}=0{,}88$$

Como el coeficiente de correlación está bastante cercano a uno, podemos concluir que las dos variables están correlacionadas positivamente.

Los costos y las ventas de las farmacias en estudio están relacionados.

Al aumentar los costos, aumentan las ventas o bien, al disminuir unos disminuyen también los otros.

La siguiente actividad tiene como objetivo que usted practique el cálculo del coeficiente de correlación lineal.