Correlación y regresión lineal simple

Por Tevni Grajales G.

Esta vez deseamos retomar el tema ya presentado y titulado "medidas de asociación". Sabemos que se trata de determinar el grado de relación o correspondencia entre dos conjuntos de valores denominados variables. Cuando la relación tiene un valor positivo significa que a valores altos en una variable corresponden valores altos en la otra variable. Y la relación con signo negativo significa que las variables están relacionadas de manera inversa de modo que cuando el valor aumenta en una, disminuye en la otra.

Para quienes desean disponer de algún criterio guiador para interpretar las correlaciones, se presenta la siguiente escala, sin antes recordar que es el sentido común y la lógica lo que en muchos casos puede guiar en determinar la importancia de una relación observada.

Coeficiente de correlación				Interpretación
.80	a	1.00		Una alta relación de dependencia
.60	a	.79		Una relación entre moderada a acentuada
.40	a	.59		Una mediana relación
.20	a	.39		Una ligera relación
.00	a	.19		Una relación fortuita o insignificante

Se puede utilizar la correlación para encontrar la relación entre dos diferentes medidas u observaciones en un mismo grupo de individuos y objetos. Para determinar la relación en características de dos grupos relacionados (EJ. Padres e hijos, esposos y esposas, etc.).

Como ya fue mencionado, algunas variables se correlacionan de manera lineal, es decir que su relación sigue el mismo comportamiento (dirección) a lo largo de todos los posibles valores de las variables. Pero en algunos casos las variables cambian la dirección de su relación a medida que varían los valores observados. En este caso se trata de una relación no lineal. En esta ocasión vamos a referirnos a correlaciones lineales para lo cual se utiliza el coeficiente producto-momento de Pearson. (Cuando se trata de relaciones curvilineales se debe usar la razón de correlación eta).

El coeficiente de correlación de Pearson

Una forma para calcular el coeficiente de correlación, según nuestro tema anterior, consiste en utilizar los valores z de cada observación. En esta ocasión vamos a presentar otra forma esta vez utilizando el cuadrado de la diferencia entre cada valor observado y su media.

$$r = \Sigma xy \,/\sqrt{(\Sigma x^2)(\Sigma y^2)}$$

Para facilitar los cáculos de este coeficiente, es posible asumir una media cualquiera y luego proceder en los cálculos. El coeficiente de correlación no se ve alterado por esta decisión.

La significatividad de la correlación

Una vez calculado el coeficiente de correlación se hace necesario determinar si tal correlación podría obtenerse como resultado de una selección aleatoria de una muestra procedente de una población no correlacionada. En otras palabras, no existiendo correlación, esta muestra resultó ser una de las pocas que se ubican en los extremos de una distribución de muestras. La pregunta es ¿la correlación que se muestra es real o producto del azar? Así que asumiendo la hipótesis nula que los valores de estas dos variables no están relacionados, o que el coeficiente de correlación es igual a cero, se procede con una prueba t de significatividad. Donde t es igual a:

$$t = r\sqrt{N - 2} / \sqrt{1 - r^2}$$

En esta ecuación, N representa el número de casos o valores pareados.

Supongamos que usted a correlacionado las calificaciones de matemática e inglés de 27 alumnos y encontró un coeficiente de correlación 0 .75. Si desea determinar la significatividad de esta correlación, procede así:

- N = 27
- r = .75
- N - 2 = 25
- √ 25 = 5
- $r^2 = .5625$

t = .75 √ 25/√ 1 - .56

t = .75 (5) / √ .44

t = 3.75/.66

t = 5.6818

Este valor t obtenido o calculado se debe comparar con el valor t crítico que ofrece la una tabla de valor t de student, usada con anterioridad. También se dispone de una tabla para pares de correlaciones elaborada por Fisher y Yates, la cual aparece en los anexos de libros de estadística, esa tabla determina cuánto tiene que ser el coeficiente de correlación según los grados de libertad para que sea significativo a un nivel alpha determinado. Se sigue el mismo procedimiento aprendido en el tema anterior, si el valor t observado es mayor que el valor t crítico se descalifica (rechaza) la hipótesis nula y se considera aceptable al hipótesis de investigación según el nivel alpha determinado. Se debe recordar que los grados de libertad están determinados por el total de pares menos dos.

Tanto el cálculo del coeficiente de correlación como de su significatividad se explican de la manera más sencilla posible dado que se espera que el estudiante utilice procesadores estadísticos para el análisis de sus datos, pero al mismo tiempo se espera que tenga una idea de la forma como se calculan los resultados que se obtienen por medio de la computadora. Así le será más fácil, al estudiante, entender e interpretar el significado de los resultados obtenidos.

La correlación es utilizada para determinar la confiabilidad y validez de pruebas o instrumentos de medición. Tema que se considera el material preparado por este servidor sobre estadística multivariante.

La correlación, la regresión y la predicción

Una vez que se conoce la correlación entre dos variables es posible predecir el valor que le correspondería a un caso en una de las variables, si conoce su calificación o valor en la otra variable. (Recuérdese que se supone que las variables tienen una correlación lineal).

Su pude decir que cuando dos variables están relacionadas entre sí, dichas variables se encuentran una en función de la otra. El término función, se utilizó por primera vez en 1692 en un artículo sobre matemática publicado en el Acta Eruditorem y se atribuyó a Gottfried von Leibniz, en 1749 Leonhard Euler lo definió como *la cantidad de una variable que es dependiente de otra cantidad* y más tarde Lejeune Dirichlet (1837) propuso que desde el punto de vista matemático la función es la correspondencia que asigna un valor único a la variable dependiente para cada valor permitido en una variable independiente.

Cuando compramos harina la pagamos por su peso (masa), podemos decir que el precio de la harina está en función de su peso (masa). Por ejemplo, si su precio es de .50 centavos de dólar por kilo, dos kilos costarán un dólar, medio kilo veinticinco centavos, diez kilos cinco dólares, etc. Esto se representa así:

$$f(x) = .50\ x$$

En este caso f(x) está definido sólo para que $x \geq 0$.

De manera que una función es una asociación entre dos o más variables, en la cual a cada valor de cada una de las variables independientes o argumentos, corresponde exactamente un valor de la variable de dependiente en un conjunto denominado específicamente, dominio de la función.

La función de una variable puede ser escrita f (x) lo cual se lee " f de x" o de manera más completa "el valor de la función f en x". De manera que si *y* es una variable que está en función de *x,* se acostumbra escribir la variable dependiente en el lado izquierdo del signo de igualdad en la ecuación, así:

$$y = x + 2 \text{ que es lo mismo que decir } f(x) = x + 2$$

Según la expresión anterior, tanto y como f(x) es la variable dependiente, siendo x la variable independiente.

También podría ser la expresión que aparece a continuación:

$$x = y + 2 \text{ que es lo mismo que decir } f(y) = y + 2$$

Nótese que en este caso, tanto x como y pueden intercambiar su papel como variable dependiente e independiente. Esto sucede así porque cuando se trata de la relación entre dos variables, los términos dependiente o independiente no trascienden el simple significado de determinar cuál variable es la que se pretende predecir en contraste con la variable que se utiliza para hacer la predicción. En otras palabras, la existencia de la relación, no permite por sí misma determinar que una variable se dependiente (efecto) de la independiente (causa). No hay lugar para la determinación de causalidad. Esa es la razón por la cual autores como Kachigan (1991) prefieren utilizar los términos, variable criterio y variable (s) predictora (s).

Cuando se trata de una terminología apropiada al método de investigación, es preferible utilizar los términos criterio y predictora en la correlación dejando los términos dependiente e independiente para aquellos estudios en los cuales, en efecto, se aspira a determinar causalidad. Como por ejemplo, estudios experimentales.

A continuación se presenta la gráfica No.1 en la que se describe la correlación significativa encontrada al encuestar doce estudiantes que comieron nieve (helado) de su preferencia por ocho días seguidos. Al final del octavo día se les solicitó de determinaran el grado de deseo y de agrado que sentían al comer nieve en esa ocasión.

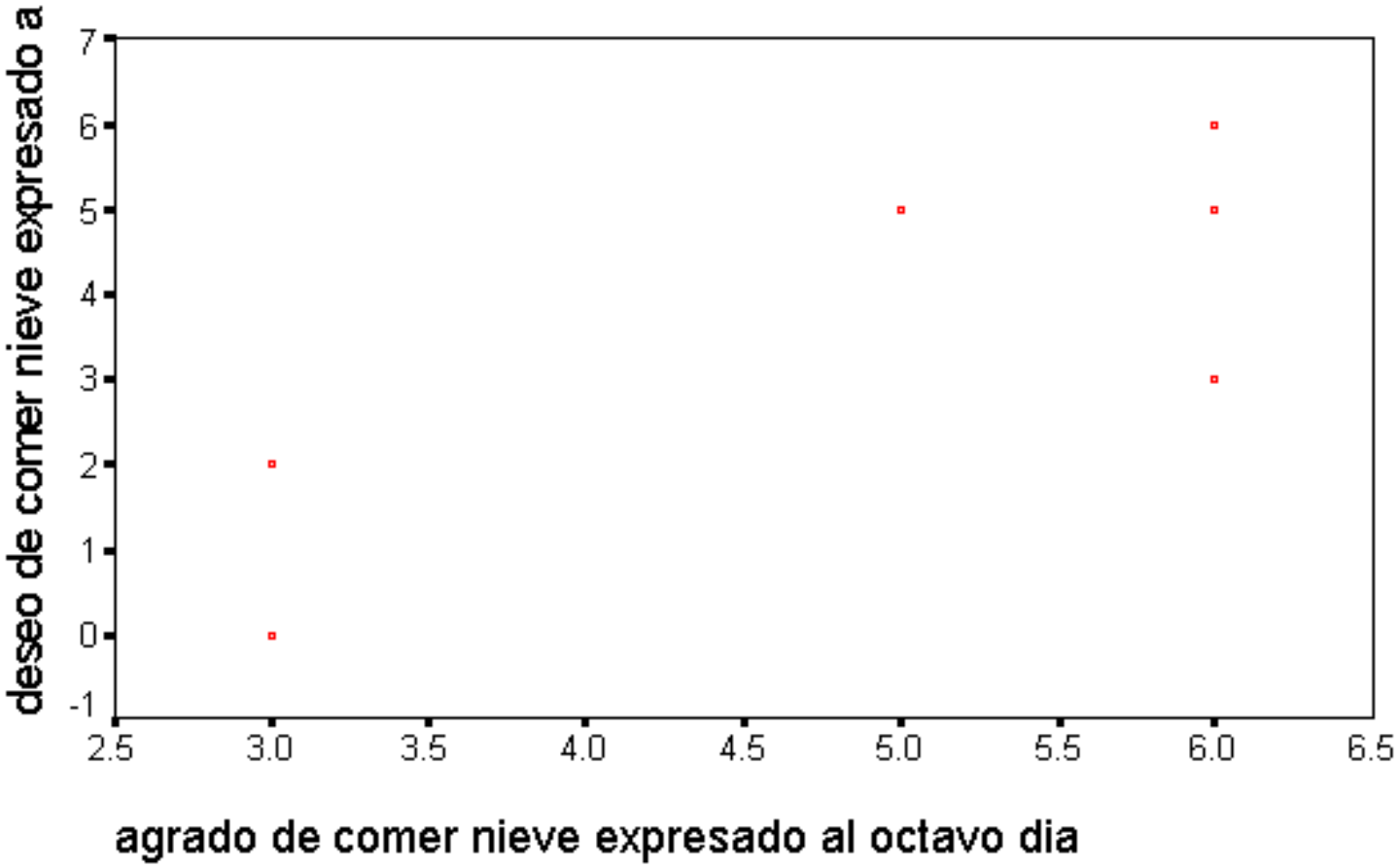

Figura 1
Relación entre el agrado por el helado y el deseo de comer helado al final de una semana de consumo diario del helado.

Pero antes de avanzar con el concepto de regresión lineal y la predicción, veamos los resultados obtenidos al calcular el coeficiente de correlación de Pearson y el grado de significatividad entre ambas variables.

Tabla 2.
Coeficiente de correlación de Pearson y su significativad entre el deseo y el agrado de comer helado al octavo día.

- - Correlation Coefficients - -

	DCNOCHOR	ACNOCHOR
DCNOCHOR	1.0000 (10) P= .	.8233 (10) P= .003
ACNOCHOR	.8233 (10) P= .003	1.0000 (10) P= .

(Coefficient / (Cases) / 2-tailed Significance)

Los resultados se pueden ver en la Tabla 1 que representa la forma como el programa estadístico SPSS Windows 6.1 ofrece los resultados. Allí podemos observar que las variables, agrado por comer nieve y el deseo de comer nieve en el

octavo día, tienen una correlación de .8233 con una significatividad menor a .05 (P=.003).

Esto descalifica con un margen de error menor al 5% a la hipótesis nula que niega la existencia de relación y conduce a considerar que en el 95% de las muestras obtenidas de la misma población, al observar esta relación, es posible obtener un coeficiente de correlación semejante a .82

Antes de cerrar este tema de las correlaciones, vamos a señalar tres asuntos muy importantes y útiles

1. Un coeficiente de correlación no muestra el porcentaje de relación entre dos variables como algunas veces se piensa. Sin embargo, el valor que se obtiene al elevar al cuadrado el coeficiente de correlación entre dos variables, muestra el porcentaje de la variabilidad (varianza) de una de las variables, que puede ser atribuido a la variabilidad (varianza) de la otra. De manera que ese cuadrado ofrece una aproximación del porcentaje de relación que existe entre las dos variables. (Para el ejemplo que hemos tenido en esta sección, correspondería decir que la correlación de .82 puede estar mostrando aproximadamente un 67% de relación entre las dos variables observadas).

2. Los coeficientes de correlación no son directamente proporcionales. Como por ejemplo, en el caso de nuestra correlación .82 muestra más que el doble de relación que una correlación de .41. Se puede obtener una proximación del valor relativo de estas dos correlaciones si elevamos al cuadrado ambos coeficientes y luego dividimos el más grande entre el más pequeño. Por ejemplo $(.82)^2 = .67$, $(.41)^2 = .1681$. De manera que al dividir .67 entre .1681 tenemos 3.98 lo que significa que la correlación .82 manifiesta cuatro veces más relación que la correlación .41.

3. Cuando dos variables muestran una correlación positiva no necesariamente significa que tienen una relación tan alta como lo muestra el coeficiente de correlación. Siendo que ellas pueden estar relacionadas con otra (s) variable (s) o factor (es) comunes. Puede que se usted observe una relación significativa entre la satisfacción en el trabajo y el grado de motivación del maestro, pero también hay que recordar que estas dos variables están relacionadas también con ciertas características del individuo como puede ser su nivel de inteligencia emocional. De manera que si podemos aislar el aporte que hace la inteligencia emocional en la variabilidad de la satisfacción y de la motivación, notaremos que la correlación entre estas últimas será menor. Esta técnica es conocida como correlación parcial y permite estimar la correlación residual entre dos variables habiendo retirado el efecto de una o más variables que intervienen.

Volviendo a la figura 1 y al observar la ubicación de las diversas puntuaciones por caso (tratándose apenas de diez casos), se nota una aparente correlación positiva, los valores altos en una variable tienden a ser altos en la otra. Ahora podríamos intentar dibujar una línea que pasara por entre los puntos a una distancia equidistante a todos los puntos y tendríamos una representación gráfica de una correlación que suponemos lineal entre las dos variables.

Esa recta imaginaria que hemos trazado pasaría entre los diez puntos cortando el eje y (de las coordenadas) en algún punto por encima o por debajo de cero (en el caso particular de nuestro ejemplo pasa por el punto +3) y la recta imaginaria tendría una pendiente o caída cuyo valor es = .5238.

Si recordamos lo mencionado al inicio de este tema, dijimos que la asociación entre dos o más variables se puede representar una función. En este caso particular se trata de la función entre dos variables. Determinemos que deseamos observar la función de x que es la variable sobre el eje de las abscisas, (agrado que siente respecto a la nieve expresado al octavo día). La variable que estaría en función de x (fx) sería la otra variable que aparece representada por los valores en el eje de las coordenadas y (deseo que siente por comer nieve al octavo día).

Por cada valor de x hay un correspondiente valor en y. Es decir por cada unidad de medida que cambia x, el valor de y cambia en cierta proporción. Esta proporción del cambio en y puede determinarse si se logra calcular la pendiente de la recta imaginaria que hemos trazado entre los puntos que representan la correlación. Y como ya mencionamos anteriormente, esta pendiente es = .5238. No vamos a detenernos a estudiar los detalles respecto a cómo se llega a este valor. Lo vamos a dejar para la clase introductoria del curso sobre estadística multivariante. Pero si vamos a señalar que consiste en un valor denominado B que podemos obtener al correr una regresión lineal simple en un paquete estadístico como el SPSS. También se obtiene el valor de una constante el cual representa el valor que tiene el punto en el que la recta imaginaria corta el eje y (coordenadas). Con estos dos valores podemos determinar la función o la ecuación de la recta que estamos tratando de encontrar.

En la Tabla 2 a continuación aparece los resultados obtenidos en el SPSS para Windows 6.1 al correr una regresión lineal simple entre las variables de este ejemplo. Allí se ofrecen un gran número de datos, los que nos interesan en este momento

aparecen bajo en la sección titulada variables en la ecuación (variables in the ecuation). Tenemos en la primera columna el nombre de la variable: DCNOCHOR (deseo de comer nieve al octavo día) y la variable Constante. En la segunda columna encabezada por una B aparecen los valores correspondientes a el valor correspondiente a la proporción del cambio en la variable dependiente por cada unidad de la variable independiente DCNOCHOR = .52 y el valor de la constante es el punto que se la línea corta el eje y.

Tabla 2

Resultados de la regresión lineal simple entre las variables

* * * * M U L T I P L E R E G R E S S I O N * * * *

Equation Number 1 Dependent Variable. ACNOCHOR agrado de comer nieve expresado

Block Number 1. Method: Enter DCNOCHOR

Variable(s) Entered on Step Number 1.DCNOCHOR deseo de comer nieve expresado al octavo

Multiple R = .82333

R Square = .67787

Adj R Square= .63761

Standard Error= .74001

Analysis of Variance	DF	Sum of Squares	Mean Square
Regression	1	9.21905	9.21905
Residual	8	4.38095	.54762

F = 16.83478 Signif F = .0034

---------- Variables in the Equation ----------

Variable	B	SE B	95% Confdnce Intrvl B		Beta
DCNOCHOR	.523810	.127664	.229415	.818204	.823329
(Constant)	3.000000	.585032	1.650916	4.349084	

---------- Variables in the Equation ----------

Variable	Tolerance	VIF	T	Sig T
DCNOCHOR	1.000000	1.000	4.103	.0034
(Constant)			5.128	.0009

Es probable que alguno de los lectores se considere perdido con esta explicación, siendo que hemos dado un salto sobre los detalles de cómo se calculan estos valores para sólo limitarnos a decir dónde se encuentran los datos requeridos cuando el programa estadístico ofrece los resultados. Bien, por ahora y dados los objetivos de hacer de usted una persona que pueda leer e interpretar resultados, no necesariamente un experto en estadística, vamos a limitarnos a esto.

De manera que usted tiene una ecuación que representa matemáticamente la recta que pasa por entre los valores correlacionados. Esa recta en este caso es una función de x.

La ecuación de una recta, de acuerdo con lo explicado anteriormente se representa por:

$$y = a + b_{(x)}$$

a = valor del punto en que la recta cruza, corta el eje de las coordenadas (y)

b = proporción de variabilidad de la variable y por cada unidad de la variable x (pendiente)

x = es cualquier valor de x que desee utilizarse para predecir su correspondiente valor en y

Para el caso particular de nuestro ejemplo

a = es el valor que aparece en la columna identificada como B para la constante = 3.0000

b = es el valor que aparece en la columna identificada como B para la DCNOCHOR = .5238

x = es cualquier valor de x que desee utilizarse para predecir su correspondiente valor en y

De manera que con la siguiente ecuación, podemos predecir cualquier valor de y a partir de algún valor de x que escojamos:

$$y = 3 + .52_{(x)}$$

Supongamos que uno de los sujetos de la investigación no contestó la pregunta relacionada con cuánto le agrada la nieve en la octava noche pero si contestó la pregunta cuánto deseo tengo de comer nieve con un valor = 4. En ese caso podemos utilizar esta ecuación para predecir el valor que corresponde a este sujeto en la variable que no contestó.

Así:

y	=	3 + .52 (x)
y	=	3 + .52 (4)
y	=	3 + 2.8
y	=	5.8

Podemos concluir que dada la correlación significativa que existe entre las variables deseo y agrado de comer nieve al octavo día en este grupo de personas, es probable que este sujeto de la investigación se ubique alrededor de un valor 5.8 en su agrado por la nieve al octavo día del experimento.

Montemorelos, 10 de octubre de 1999