UNIDAD II. REGRESIÓN Y CORRELACIÓN LINEAL

¡Hola!

En la sesión pasada vimos como se pueden relacionar dos variables y calcular su grado y tipo de relación. En esta clase veremos que es posible generar un modelo matemático de relación de ambas variables suponiendo que su relación es lineal. A esto se le llama ajuste lineal.

Pon mucha atención Recuerda que es importante comprometerse con tu aprendizaje para tener éxito.

Espero lo disfrutes.

¡Comencemos!

Aun cuando el coeficiente de correlación mide la fuerza de una relación lineal, no nos dice nada acerca de la relación matemática entre las dos variables. En la sesión pasada se encontró que el coeficiente de relación para los datos de lagartijas/sentadillas era de 0.84. Esto, además del patrón del diagrama de dispersión, implica que hay una relación lineal entre el número de lagartijas y el de sentadillas que hace un estudiante. No obstante, el coeficiente de correlación no nos ayuda a pronosticar el número de sentadillas que una persona puede hacer con base en saber que puede hacer 28 lagartijas.

El análisis de regresión encuentra la ecuación de la recta que mejor describe la relación entre las dos variables. Un uso de esta ecuación es hacer predicciones. Hacemos uso de estas predicciones regularmente, por ejemplo, predecimos el éxito que un estudiante tendrá en la universidad con base en sus resultados de preparatoria y predecir la distancia necesaria para detener un auto con base en su velocidad. En general, el valor exacto de y no se puede pronosticar y nos satisface saber que las predicciones son razonablemente cercanas.

En el esquema siguiente se muestran los temas que abordaremos de la unidad II en esta sesión.

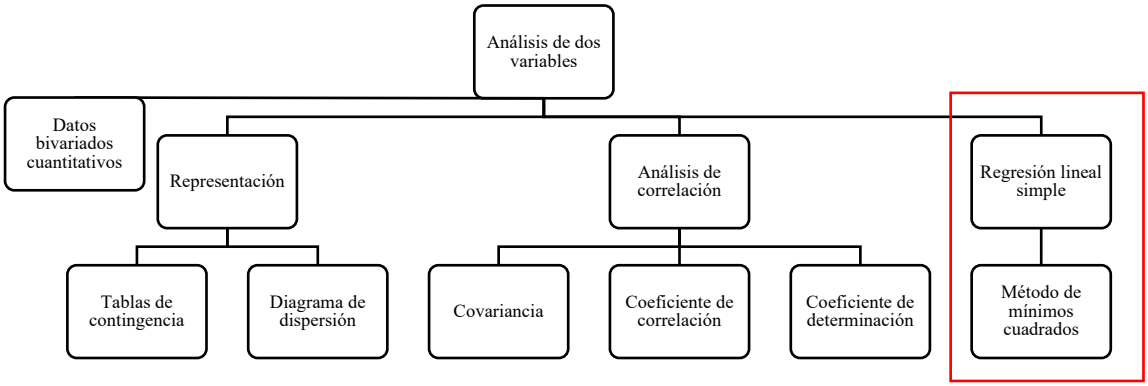

Figura. 1. Esquema de la unidad II.

1. REGRESIÓN LINEAL SIMPLE

La relación entre dos variables será una expresión algebraica que describe la relación matemática entre x y y. A continuación, veamos algunos ejemplos de diversas relaciones posibles, llamadas modelos o ecuaciones de predicción:

Lineal (línea recta):	$\hat{y} = b_0 + b_1 x$
Cuadrática:	$\hat{y} = a + bx + cx^2$
Exponencial:	$\hat{y} = a(b^x)$
Logarítmica:	$\hat{y} = a \log_b x$

En la siguiente gráfica se muestra la relación entre datos bivariados, que indican diferentes modelos.

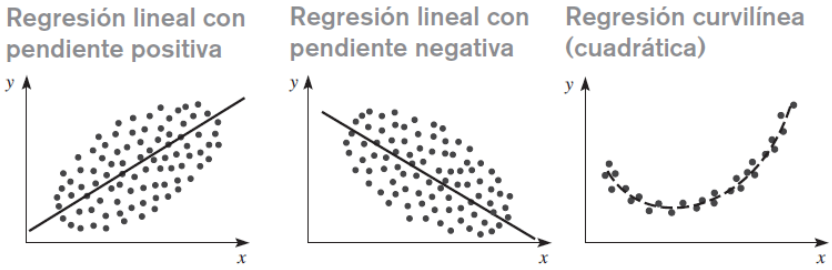

Figura 2. Gráficas de datos bivariados.

Si un modelo de recta parece apropiado, la recta de mejor ajuste se encuentra con el uso del método de mínimos cuadrados.

El término regresión fue introducido por Francis Galton (1822-1911). En un famoso artículo, Galton planteó que, a pesar de que había una tendencia en la que los padres de estatura alta tenían hijos altos y padres de estatura baja tenían hijos bajos, la estatura promedio de los niños nacidos de padres de una estatura dada tendía a moverse o regresar hacia la estatura promedio de la población total. En palabras de Galton, se trataba de una regresión a la mediocridad.

Fue tal el éxito de la contribución de las investigaciones de Galton que originaron la ley de regresión universal, que fue confirmada por Karl Pearson (1857-1936) y que constituye el pilar del análisis de regresión. En general, el análisis de regresión se centra en la exploración, explicación y estudio de dependencia de una variable mediante una o más variables explicativas, de ahí el nombre del método de predicción basado en este modelo.

En este tema haremos una introducción al método de regresión lineal simple. Recibe este nombre, porque:

- Regresión: utilizaremos información pasada
- Lineal: bajo el supuesto de que entre dos variables (X y Y) existe una relación lineal

- Simple: usaremos sólo una variable independiente para tratar de explicar la variable dependiente.

En otras palabras, ajustaremos una recta a los datos. "Ajustar" se refiere a construir la única recta que pase lo más cerca de todos los puntos ubicados en el diagrama de dispersión

1.1 Método de mínimos cuadrados

El análisis de regresión consiste en definir la variable independiente X que ayude a explicar (estimar) la variable dependiente Y, siempre que exista una relación lineal entre ellas, además de que ambas variables deben ser cuantitativas.

El método de mínimos cuadrados se usa para determinar la ecuación de la recta de regresión, es decir, por medio de él se encuentra la única recta que pasa lo más cerca que se puede de todos los puntos (observaciones) ubicados en un diagrama. La ecuación del método de mínimos cuadrados es:

$$Y' = b_1 X + b_0$$

donde Y' = valor estimado (aproximado) de Y

b_0 = ordenada al origen; es el valor de Y' cuando X es igual a cero.

b_1 = pendiente de la recta; es el cambio en Y' cuando X aumenta en una unidad.

El criterio de mínimos cuadrados requiere que encontremos las constantes b_o y b_1 tales que

$$\sum_{i=1}^{n} \left(Y_i - Y_i' \right)^2$$

sea tan pequeña como sea posible.

donde

Yi = el valor i de Y (valor real)

$Y'i$ = el valor i estimado de Y (valor sobre la recta de regresión), es decir, es la distancia que hay entre cada punto y la recta de regresión.

Minimizando esas distancias se obtienen b_0, la ordenada al origen, y b_1, la pendiente de la recta. Así, las ecuaciones para determinar b0 y b1 son:

$$b_1 = \frac{rs_y}{s_x} \text{ y } b_0 = \overline{Y} - b_1\overline{X}$$

Donde

r = coeficiente de correlación

s_y = desviación estándar muestral de Y

s_x = desviación estándar muestral de X

$\overline{Y}$ = media muestral de Y

$\overline{X}$ = media muestral de X

A b_0 y b_1 se les denomina coeficientes de regresión.

Ejemplo 2.1. Contracción del concreto

La contracción por desecación del concreto se define como la contracción de una mezcla endurecida de concreto debida a la pérdida de agua capilar. Todo concreto de cemento Portland experimenta contracción por desecación, o cambio en volumen hidráulico, a medida que el concreto envejece.

Dados los siguientes datos, realiza un ajuste de regresión lineal.

Contenido de agua $(kg/m)^3$	Contracción (10^{-6})
202	380
210	360
220	400
231	390
242	580
167	255

Tabla 1. Contracción del concreto.

Solución

Como observarás en la fórmula, requerimos calcular el coeficiente de correlación r, y las desviaciones estándar muestrales de X y Y.

Para eso conviene realizar una tabla de contingencia como en las sesiones pasadas que contengan todas las columnas necesarias para el cálculo.

Paso 1. Asignamos a las variables X y Y y reescribimos los valores en la columna 1 y 2.

X- Contenido de agua de una mezcla de concreto

Y- Contracción del concreto

Paso 2. Calcula el valor de la media de X y la media de Y

$$\bar{x} = \frac{202 + 210 + 220 + 231 + 242 + 167}{6} \qquad \bar{x} = 212$$

$$\bar{y} = \frac{380 + 360 + 400 + 390 + 580 + 255}{6} \qquad \bar{y} = 394.16$$

Paso 3. En la tercera columna calcula el valor de la resta de cada x menos la media de x.

Paso 4. En la cuarta columna calcula el valor de la resta de cada y menos la media de y.

Paso 5. En la quinta columna multiplica las columnas tres y cuatro.

Paso 6. En la sexta columna calcula el cuadrado de los valores de la columna 3.

Paso 7. En la sexta columna calcula el cuadrado de los valores de la columna 4.

Paso 8. Suma los valores de las columnas cinco, seis y siete.

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
202	380	-10	-14.17	141.67	100.00	200.69
210.00	360.00	-2.00	-34.17	68.33	4.00	1167.36
220	400	8	5.83	46.67	64.00	34.03
231.00	390.00	19.00	-4.17	-79.17	361.00	17.36
242	580	30	185.83	5575.00	900.00	34534.03
167.00	255.00	-45.00	-139.17	6262.50	2025.00	19367.36
			$\sum(X-\bar{X})(Y-\bar{Y})$	12015.00	3454.00	55320.83

Tabla 2. Cálculo de regresión lineal

Realizando los cálculos:

Covarianza

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \qquad S_{xy} = \frac{12015}{6-1} \qquad S_{xy} = 2403$$

Desviaciones estándar

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad S_x = \sqrt{\frac{3454}{6-1}} \qquad S_x = \sqrt{690.80}$$

$$S_x = 26.28$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad S_y = \sqrt{\frac{55320.83}{6-1}} \qquad S_y = \sqrt{11064.16}$$

$$S_y = 105.18$$

Coeficiente de correlación

$$r = \frac{S_{xy}}{S_x S_y} \qquad r = \frac{2043}{(26.28)(105.18)} \qquad r = 0.7243$$

Pendiente de la ecuación

$$b_1 = \frac{rS_y}{S_x} \qquad b_1 = \frac{(0.7243)(105.18)}{26.28} \qquad b_1 = 2.89$$

Ordenada en el origen

$$b_0 = \bar{Y} - b_1\bar{X} \qquad b_0 = 394.16 - 2.89(212) \qquad b_0 = -220.38$$

Ecuación de la regresión lineal

$$Y' = 2.89x - 220.38$$

Una estimación: se quiere saber la contracción que va a tener el concreto si tiene 250 kg/m^3

$$Y' = 2.89(250) - 220.38$$

$$Y' = 504.321x10^{-6}m$$

Ahora vamos a trazar la gráfica, para eso, rescribimos los datos en pares ordenados.

(202,380) (210,360) (220,400) (231,390) (242,580) (167,255)

Para grafica la recta, de la ecuación obtenemos dos puntos:

$$Y' = 2.89x - 220.38$$

$x = 180$	$Y' = 2.89(180) - 220.38$	Y=301.4
$x = 235$	$Y' = 2.89(235) - 220.38$	Y=460.8

Colocamos los puntos de la recta con rojo

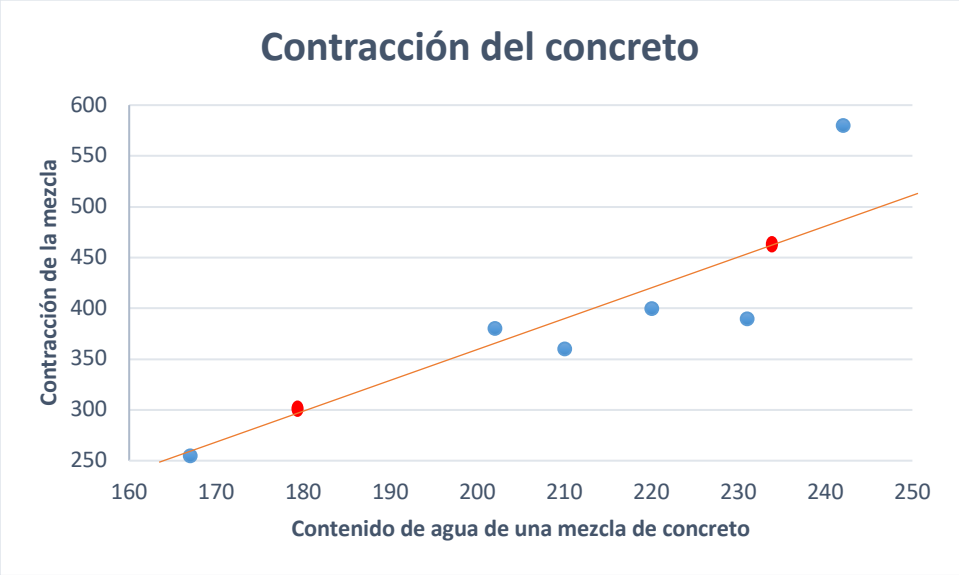

Figura 3. Gráfica de correlación con ajuste lineal.

Interpretación: La relación entre el contenido del agua y la contracción del concreto es lineal fuerte directa con un coeficiente de correlación r=0.72. La recta de ajuste lineal es $Y' = 2.89x - 220.38$, observa que la pendiente es positiva por lo que hablamos de una recta creciente.

Actividad 2.1

Tiempo estimado: 60 minutos

Los golfistas profesionales tienen un dilema clásico en golf: "haz un tiro largo para exhibirte, uno corto para ganar dinero". Es frecuente que el juego en corto (en el "green") lo que determina si ganan un torneo. El 7 de enero de 2005, en un artículo de USA Today titulado "En corto, la meta de Durant es mejorar", se publicó una tabla que indicaba los porcentajes de victorias para los jugadores del torneo PGA de golfistas profesionales en la temporada de 2004, para llegar a los "greens" desde varias distancias.

Yardas	Distancia media	Porcentaje de victorias
Más o menos 200	213	44
176–200	188	53
151–175	163	61
126–150	138	68
101–125	113	72
76–100	88	78
75 o menos	63	85

Fuente: PGA Tour Shotlink

Usando las distancias medias en yardas como variable independiente, x, y el porcentaje de victorias como la variable dependiente, y:

a. Construya un diagrama de dispersión.

b. ¿Parece haber una correlación lineal? Justifique su respuesta.

c. Calcule el coeficiente de correlación lineal, r.

d. Interprete el coeficiente de correlación hallado en la parte c. Comente sobre su dirección y fuerza.

e. ¿Parece haber una relación lineal? Justifique su respuesta.

f. Calcule la ecuación de la recta de mejor ajuste.

g. Grafique la recta de mejor ajuste sobre el diagrama de dispersión.

h. Pronostique el porcentaje promedio de victorias para un golfista profesional si llegó hasta el "green" desde una distancia de 90 yardas.

¡Muy bien!

Has llegado al final de la sesión y de la unidad II.

Como viste el tema de regresión y correlación es muy simple pero laborioso. Realiza un formulario y tenlo siempre a la mano. También ten cuidado con los cálculos, fíjate que teclees muy bien los números en la calculadora y si es necesario has una comprobación.

Realiza el Quizz para que veas lo que has aprendido.

¡Seguro podrás lograrlo!

Bibliografía

Banegas, A. L. (2012). *Probabilidad y estadística. Enfoque por competencias.* México: INTERAMERICANA EDITORES S.A. DE C.V.

Robert Johson, P. K. (2008). *Estadítica elemental: Lo esencial.* México: Cengage Learning Editores, S.A. de C.V.