UNIVERSIDAD DE SAN PEDRO SULA

5. Regresión Lineal Múltiple

Introducción

La regresión lineal simple es en base a una variable independiente y una dependiente; en el caso de la regresión línea múltiple, solamente es una variable dependiente y más de una independiente.

Análisis de regresión múltiple

Técnica para desarrollar la ecuación de regresión y proporcionar los valores estimados para más de una variable independiente.

Desde el punto de vista de la matemática, una ecuación lineal tiene la forma de Y=a + bX; en donde "b" se conoce como la "pendiente de la ecuación" y "a" es el intercepto en el eje Y. La fórmula general de una ecuación de regresión lineal se denota como:

ECUACIÓN DE REGRESIÓN MÚLTIPLE	$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$

Donde:

$\hat{Y}$	:	Valor de pronóstico
X_i	:	Variables independientes
a	:	Intersección en Y
b_i	:	Pendientes de la ecuación de regresión múltiple

Las tablas que forman la muestra tendrían la siguiente forma:

Descripción	(Y)	(X_1)	(X_2)	(X_3)

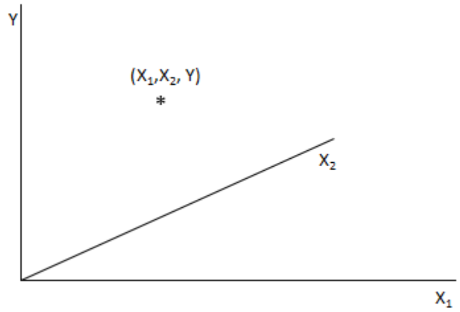
Si la ecuación solamente tiene dos variables independientes, su notación será la siguiente:

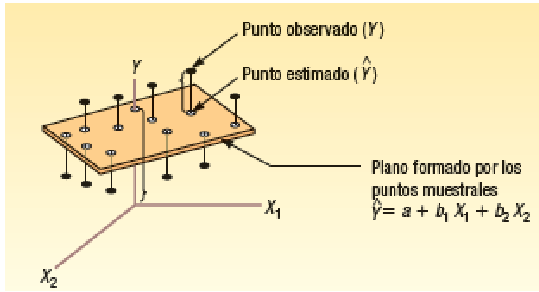
ECUACIÓN DE REGRESIÓN MÚLTIPLE DE 3 VARIABLES.	$\hat{Y} = a + b_1X_1 + b_2X_2$

Los coeficientes b_i se obtienen mediante el método de mínimos cuadrados; pero, usualmente se utilizan paquetes estadísticos para obtener los resultados. Entre los paquetes más populares se puede mencionar:

- Excel
- Minitab
- SPSS
- PhStat2
- SAS

USAP
UNIVERSIDAD DE SAN PEDRO SULA





Ejemplo 5.1

1. Suponer que el rendimiento por galón de combustible de un vehículo tiene una relación directa con el octanaje de la gasolina (X_1) y una relación inversa con el peso del automóvil (X_2). El valor de pronóstico indica la cantidad de millas por galón que recorrería el vehículo y la ecuación de regresión lineal múltiple es la siguiente:

$$\hat{Y} = 6.3 + 0.2X_1 - 0.001X_2$$

Calcular el pronóstico para un vehículo con peso de 2,000 libras y con 92 octanos.

Desarrollo

Clasificación de las variables:

$X_1 = 92$ Octanaje
$X_2 = 2000$ Peso del vehículo

$$\hat{Y} = 6.3 + 0.2X_1 - 0.001X_2$$
$$\hat{Y} = 6.3 + 0.2(92) - 0.001(2000)$$
$$\hat{Y} = 6.3 + 18.4 - 2$$
$$\hat{Y} = 22.7$$

El automóvil recorre 22.7 millas por galón

2. El director de marketing en Productos Ivory estudia las ventas mensuales de la Distribuidora Internacional; seleccionó tres variables independientes como estimadores de las ventas: población regional, ingreso *per cápita* y la tasa de desempleo regional. La ecuación de regresión se calculó (en dólares):

$$\hat{Y} = 64100 + 0.394X_1 + 9.6X_2 - 11600X_3$$

¿Cuáles son las ventas anuales estimadas para una región particular con una población de 796,000 habitantes, un ingreso per cápita de L.6,940 y una tasa de desempleo de 6%

Desarrollo

Clasificación de las variables

$X_1 = 64100$ Población regional
$X_2 = 6940$ Ingreso per cápita
$X_3 = 6.0$ Tasa de desempleo

$$\hat{Y} = 64100 + 0.394X_1 + 9.6X_2 - 11600X_3$$
$$\hat{Y} = 64100 + 0.394(64100) + 9.6(6940) - 11600(6)$$
$$\hat{Y} = 374\ 748$$

Las ventas anuales que se podrían obtener son de 374 748 dólares

UNIVERSIDAD DE SAN PEDRO SULA

Error estándar de estimación múltiple

Por lo general, no todos los puntos del diagrama de dispersión están ubicados sobre la recta que resulta de la ecuación de regresión lineal; en el trazo de toda ecuación de regresión siempre se presenta un error por las múltiples variaciones que se presentan; a este error se le llama “Error estándar de estimación”.

ERROR ESTÁNDAR DE ESTIMACIÓN: Medida de dispersión de los valores observados respecto de la recta de regresión.” (Lind |Marchal |Wathen, 2008, p.478).

Si el error estándar de la estimación es pequeño, los datos están relativamente cercanos a la recta de la ecuación de regresión lineal y significa que se predice Y con poco error; si es grande, los datos están muy dispersos de la recta de la ecuación de regresión lineal y no proporciona una estimación precisa de Y.

ERROR ESTÁNDAR DE LA ESTIMACIÓN MÚLTIPLE	$s_{Y \cdot X} = \sqrt{\dfrac{\sum\left(Y - \hat{Y}\right)^2}{n-(k+1)}}$

Donde:

Y : Observación actual

$\hat{Y}$: Valor de pronóstico

n : Tamaño de la muestra

k : Número de variables independientes

Ejemplo 5.2

1. En una empresa inmobiliaria de desea determinar el gasto de electricidad durante el invierno en la costa este de Estados Unidos. Al departamento de investigación se le pidió desarrollar algunas directrices respecto de los costos de calefacción de casas unifamiliares. Se considera que tres variables se relacionan con los costos de calefacción: 1) la temperatura externa diaria media, 2) el número de pulgadas de aislamiento en el ático y 3) la antigüedad en anos del calentador. Para el estudio, se seleccionó una muestra aleatoria de casas de venta reciente. Se determinó el costo de calefacción de cada casa en enero pasado, así como la temperatura externa en enero en la región, el número de pulgadas de aislamiento en el ático y la edad del calentador. Obteniendo los siguientes resultados:

Casa	Costo de Calefacción $ (Y)	Temperatura externa media °F (X_1)	Aislamiento del ático (Pulg.) (X_2)	Antigüedad del calentador (años) (X_3)
1	250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	200	10	6	7
8	355	7	10	10

UNIVERSIDAD DE SAN PEDRO SULA

La ecuación para calcular el pronóstico es la siguiente:

$$\hat{Y} = 427.194 - 4.583X_1 - 14.83X_2 + 6.101X_3$$

Calcular el error estándar de la estimación.

Desarrollo

- **Calcular el pronóstico de la tabla:**

Casa	Costo de Calefacción $ (Y)	Temperatura externa media °F (X_1)	Aislamiento del ático (Pulg.) (X_2)	antigüedad del calentador (años) (X_3)	Pronóstico $\hat{Y}$
1	250	35	3	6	258.9
2	360	29	4	10	296.0
3	165	36	7	3	176.7
4	43	60	6	9	118.1
5	92	65	5	6	91.8
6	200	30	5	5	246.1
7	200	10	6	7	335.1
8	355	7	10	10	307.8
				Σ	1,830.4

- Calcular el error estándar de la estimación

$$s_{Y \cdot X} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - (k + 1)}}$$

$$s_{Y \cdot X} = \sqrt{\frac{1830.4}{8 - (3 + 1)}}$$

$$s_{Y \cdot X} = \sqrt{\frac{1830.4}{4}}$$

$$s_{Y \cdot X} = 21.39$$

2. Celulon, fabricante de aislamiento para casas, desea desarrollar guías para informar a constructores y consumidores sobre la forma como el espesor del aislamiento en el ático de una casa y la temperatura externa afectan el consumo de gas natural. En el laboratorio se varió el espesor del aislamiento y la temperatura, algunos resultados fueron:

Consumo de gas mensual (pies cúbicos) Y	Espesor del aislamiento (pulgadas) X_1	Temperatura externa (°F) X_2
30.3	6	40
26.9	12	40
22.1	8	49
25.3	10	41

Con base en los resultados muestrales, la ecuación de regresión es:

$$\hat{Y} = 62.65 - 1.86X_1 - 0.52X_2$$

Calcular el error estándar de la estimación

USAP
UNIVERSIDAD DE SAN PEDRO SULA

Desarrollo

– Calcular el pronóstico de la tabla

Consumo de gas mensual (pies cúbicos) Y	Espesor del aislamiento (pulgadas) X_1	Temperatura externa (°F) X_2	$\hat{Y}$
30.3	6	40	30.58
26.9	12	40	19.42
22.1	8	49	22.18
25.3	10	41	22.62
		Σ	94.80

- Calcular el error estándar de la estimación

$$s_{Y\cdot X} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-(k+1)}}$$

$$s_{Y\cdot X} = \sqrt{\frac{94.80}{4-(2+1)}}$$

$$s_{Y\cdot X} = \sqrt{\frac{94.8}{1}}$$

$$s_{Y\cdot X} = 9.74$$

Coeficiente de determinación múltiple

Porcentaje de variación en la variable dependiente, *Y*, explicada por el conjunto de variables independientes, X_1, X_2, X_3,…, X_k.

Las características del coeficiente de determinación múltiple son:

1. **Se representa por una letra *R* mayúscula al cuadrado.** En otras palabras, se escribe como R^2 debido a que se comporta como el cuadrado de un coeficiente de Correlación.
2. **Puede variar de 0 a 1.** Un valor cercano a 0 indica poca asociación entre el conjunto de variables independientes y la variable dependiente. Un valor cercano a 1 significa Una asociación fuerte.
3. **No puede adoptar valores negativos.** Ningún número que se eleve al cuadrado o se eleve a la segunda potencia puede ser negativo.
4. **Es fácil de interpretar.** Como *R*2 es un valor entre 0 y 1 es fácil de interpretar, comparar y comprender.

El coeficiente de determinación se calcula a partir de la información determinada en la Tabla ANOVA. Se utiliza la columna de suma de cuadrados normal la suma de cuadrados de la regresión.

$$R^2 = \frac{VT}{VT+VA}$$

UNIVERSIDAD DE SAN PEDRO SULA

BIBLIOGRAFÍA

- Lind, D.A., Marchal, W.G., Wathen, S.A. (15). (2012). *Estadística Aplicada a los Negocios y la Economía*. México: McGraw-Hill

- David M. Levine, Timothy C. Krehbiel, Mark L. Berenson. 2006. *Estadística para Administración*. (4° edición). Naucalpan de Juárez, México.: Pearson Prentice Hall