Regresión múltiple

I.- INTRODUCCIÓN

Como la Estadística Inferencial nos permite trabajar con una variable a nivel de intervalo o razón, así también se puede comprender la relación de dos o más variables y nos permitirá relacionar mediante ecuaciones, una variable en relación de la otra variable llamándose Regresión Lineal y una variable en relación a otras variables llamándose Regresión múltiple.
Casi constantemente en la practica de la investigación estadística, se encuentran variables que de alguna manera están relacionados entre si, por lo que es posible que una de las variables puedan relacionarse matemáticamente en función de otra u otras variables.

II.- MARCO TEORICO

Se define como un procedimiento mediante el cual se trata de determinar si existe o no relación de dependencia entre dos o más variables. Es decir, conociendo los valores de una variable independiente, se trata de estimar los valores, de una o más variables dependientes.
La regresión en forma grafica, trata de lograr que una dispersión de las frecuencias sea ajustada a una línea recta o curva.

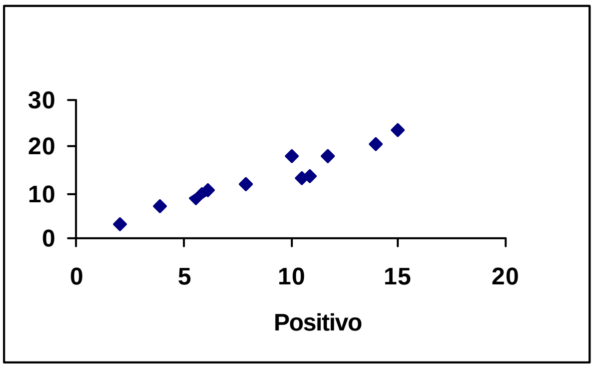
Clases de Regresión
La regresión puede ser Lineal y Curvilínea o no lineal, ambos tipos de regresión pueden ser a su vez:

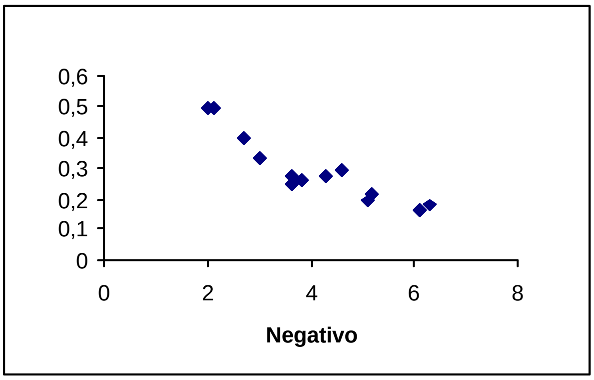
a) **Regresión Simple**: Este tipo se presenta cuando una variable independiente ejerce influencia sobre otra variable dependiente. Ejemplo: Y = f(x)
Esta regresión se utiliza con mayor frecuencia en las ciencias económicas, y sus disciplinas tecnológicas. Cualquier función no lineal, es transformada en lineal para su estudio y efectos.
Objetivo: Se utiliza la regresión lineal simple para:
1.- Determinar la relación de dependencia que tiene una variable respecto a otra.
2.- Ajustar la distribución de frecuencias de una línea, es decir, determinar la forma de la línea de regresión.
3.- Predecir un dato desconocido de una variable partiendo de los datos conocidos de otra variable.
Por ejemplo:
En una empresa de servicio de Internet busca relacionar las ganancias que obtiene cada computadora con el numero de usuarios que ingresan a dicha cabina diariamente. En la tabla representa Y (Ganancias S/.) e X (Numero de usuarios)

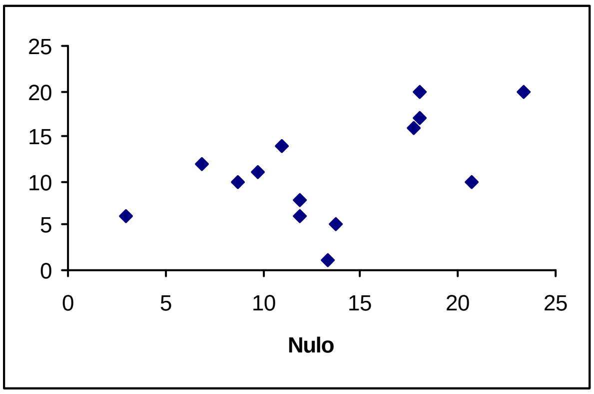
Y	100	98	99	102	102	111	97	104	102	96
X	116	96	110	105	99	106	100	109	98	108

Coeficiente de Regresión
Indica el número de unidades en que se modifica la variable dependiente "Y" por efecto del cambio de la variable independiente "X" o viceversa en una unidad de medida.

Clases de coeficiente de Regresión:
El coeficiente de regresión puede ser: Positivo, Negativo y Nulo.
Es positivo cuando las variaciones de la variable independiente X son directamente proporcionales a las variaciones de la variable dependiente "Y"
Es negativo, cuando las variaciones de la variable independiente "X" son inversamente proporcionales a las variaciones de las variables dependientes "Y"
Es nulo o cero, cuando entre las variables dependientes "Y" e independientes "X" no existen relación alguna.







Procedimiento para hallar el Coeficiente de Regresión

Para determinar el valor del coeficiente de regresión de una manera fácil y exacta es utilizando el método de los Mínimos Cuadrados de dos maneras:

1.- Forma Directa

De la ecuación de la recta:

$$\boxed{Y = a_0 + a_1 x}$$

Si a_0 y a_1 , se obtienen a partir de las ecuaciones normales:

$$\sum y = a_0 N + a_1 \sum x$$

$$\sum xy = a_0 \sum x + a_1 \sum x^2$$

Aplicando normales Y sobre X tenemos:

$$a_1 = \frac{n\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2}$$

El Coeficiente de Regresión es

$$= a_1$$

De la misma manera la recta de regresión de "X" sobre "Y" será dada de la siguiente manera:

$$\boxed{X = b_0 + b_1 y}$$

Donde: b_0 y b_1 se obtienen a partir de las ecuaciones normales:

$$\sum x = b_0 N + b_1 \sum y$$

$$\sum xy = b_0 \sum y + b_1 \sum y^2$$

Aplicando normales X sobre Y tenemos:

$$b_0 = \frac{\sum x \sum y^2 - \sum y \sum xy}{N \sum y^2 - \left(\sum y\right)^2}$$

$$b_1 = \frac{N \sum xy - \sum x \sum y}{N \sum y^2 - \left(\sum y\right)^2}$$

El Coeficiente de Regresión es $= b_1$

2.- Forma Indirecta del Método de los Mínimos Cuadrados.

El fundamento de este método es de las desviaciones de X respecto a su media aritmética. X

$$y = \left[\frac{\sum xy}{\sum x^2}\right] x \qquad\qquad x = \left[\frac{\sum xy}{\sum y^2}\right] y$$

Ecuación de ***y*** sobre ***x*** Ecuación de ***y*** sobre ***x***

Ecuación de ***y*** sobre ***x*** Ecuación de ***y*** sobre ***x***

Donde:

$x = X - \overline{X}$

$y = Y - \overline{Y}$

x, y = desviaciones

$\overline{X}$ = media aritmética

—

Y = media aritmética

b) **Regresión Múltiple**: Este tipo se presenta cuando dos o más variables independientes influyen sobre una variable dependiente. Ejemplo: Y = f(x, w, z).

Por ejemplo:
Una Empresa de desarrollo de software establece relacionar sus Ventas en función del numero de pedidos de los tipos de software que desarrolla (Sistemas, Educativos y Automatizaciones Empresariales), para atender 10 proyectos en el presente año.
En la Tabla representa Y (Ventas miles de S/.) e X (Nº pedidos de sistemas), W (Nº de pedidos de Aplicaciones Educativas) y Z (Nº de pedidos de Automatizaciones empresariales).

Y	440	455	470	510	506	480	460	500	490	450
X	50	40	35	45	51	55	53	48	38	44
W	105	140	110	130	125	115	100	103	118	98
Z	75	68	70	64	67	72	70	73	69	74

Objetivo: Se presentara primero el análisis de regresión múltiple al desarrollar y explicar el uso de la ecuación de regresión múltiple, así como el error estándar múltiple de estimación. Después se medirá la fuerza de la relación entre las variables independientes, utilizando los coeficientes múltiples de determinación.

Análisis de Regresión Múltiple
Dispone de una ecuación con dos variables independientes adicionales:

$$\boxed{Y' = a' + b_1x_1 + b_2x_2}$$

Se puede ampliar para cualquier número “m” de variables independientes:

$$\boxed{Y' = a' + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots + b_mx_m}$$

Para poder resolver y obtener a, b_1 y b_2 en una ecuación de regresión múltiple el cálculo se presenta muy tediosa porque se tiene atender 3 ecuaciones que se generan por el método de mínimo de cuadrados:

$$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2$$

$$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^{\ 2} + b_2 \sum x_1 x_2$$

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^{\ 2}$$

Para poder resolver se puede utilizar programas informáticos como AD+, SPSS y Minitab y Excel.

El error estándar de la regresión múltiple (S_{xy})

Es una medida de dispersión la estimación se hace más precisa conforme el grado de dispersión alrededor del plano de regresión se hace mas pequeño.
Para medirla se utiliza la formula:

$$\boxed{S_{xy} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - m - 1}}}$$

Y : Valores observados en la muestra
$\hat{Y}$: Valores estimados a partir a partir de la ecuación de regresión
n : Número de datos
m : Número de variables independientes

El coeficiente de determinación múltiple $\left(r^2\right)$

Mide la tasa porcentual de los cambios de Y que pueden ser explicados por x_1, x_2 y x_3 simultáneamente.

$$r^2 = \frac{SCregresiòn}{SCTotal}$$

III. APLICACION DE REGRESION MULTIPLE

Mediante el siguiente problema podremos ilustrar la aplicación de Regresión Múltiple:
En la Facultad de Ingeniería de Sistemas se quiere entender los factores de aprendizaje de los alumnos que cursan la asignatura de PHP, para lo cual se escoge al azar una muestra de 15 alumnos y ellos registran notas promedios en las asignaturas de Algoritmos, Base de Datos y Programación como se muestran en el siguiente cuadro.

Alumno	PHP	Algoritmos	Base de Datos	Programación
1	13	15	15	13
2	13	14	13	12
3	13	16	13	14
4	15	20	14	16
5	16	18	18	17
6	15	16	17	15
7	12	13	15	11
8	13	16	14	15
9	13	15	14	13
10	13	14	13	10
11	11	12	12	10
12	14	16	11	14
13	15	17	16	15
14	15	19	14	16
15	15	13	15	10

Lo que buscamos es construir un modelo para determinar la dependencia que exista de aprendizaje reflejada en las notas de la asignatura de PHP, conociendo las notas de las asignaturas Algoritmos, Base de Datos y Programación.
Se presentara la siguiente ecuación a resolver:

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3$$

Utilizando las formulas de las ecuaciones normales a los datos obtendremos los coeficientes de regresión o utilizando Regresión de Análisis de datos, en la Hoja de Calculo de Excel podemos calcular también los coeficientes de regresión:

	Coeficientes	*Error típico*	*Estadístico t*	*p-valor*	*Inferior 95%*	*Superior 95%*
Intercepción	2.551	2.369	1.077	0.305	-2.663	7.766
Algoritmos	0.583	0.267	2.186	0.051	-0.004	1.169
Base de Datos	0.373	0.144	2.589	0.025	0.056	0.691
Programación	-0.242	0.270	-0.893	0.391	-0.837	0.354

Por lo tanto podemos construir la ecuación de regresión que buscamos:

$$Y = 2.551 + 0.583x_1 + 0.373x_2 - 0.242x_3$$

El Error Estándar de Regresión Múltiple $(S_{x,y})$

Mediante esta medida de dispersión se hace más preciso el grado de dispersión alrededor del plano de regresión, se hace más pequeño.
Para calcularla se utiliza la formula siguiente:

$$S_{yx} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - m - 1}}$$

En los resultados de Excel se llama ***error típico*** y para explicar la relación del aprendizaje de PHP que se viene desarrollando es de ***0.861***

El coeficiente de determinación múltiple (r2)
Utilizaremos para determinar la tasa porcentual de Y para ser explicados las variables múltiples, utilizando la siguiente formula:

$$r^2 = \frac{SCregresiòn}{SCTotal}$$

$$r^2 = \frac{18.7737874}{26.9333333} = 0.69704656$$

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.83489314
Coeficiente de determinación R^2	0.69704656
R^2 ajustado	0.6144229
Error típico	0.86126471
Observaciones	15

IV.- CONCLUSIONES

El 69.70% del aprendizaje del Curso de PHP puede ser explicado mediante las notas obtenidas por las asignaturas de Algoritmos, Base de Datos y Programación.

V.- BIBLIOGRAFIA DE REGRESION

- Galdos Cálculo y Estadística III Edición Unica. Grupo La Republica. Lima Perú; 2005.
- Cannavos G. Probabilidad y Estadística Aplicación y métodos. Ed. en español Mc GRAW- HILL/INTERAMERICANA DE MEXICO.1995.