

1. INTRODUCCIÓN

Si sabemos que existe una relación entre una variable denominada dependiente y otras denominadas independientes (como por ejemplo las existentes entre: la experiencia profesional de los trabajadores y sus respectivos sueldos, las estaturas y pesos de personas, la **producción** agraria y la cantidad de fertilizantes utilizados, etc.), puede darse el problema de que la dependiente asuma múltiples **valores** para una combinación de valores de las independientes.

La dependencia a la que hacemos referencia es relacional **matemática** y no necesariamente de causalidad. Así, para un mismo número de unidades producidas, pueden existir niveles de **costo**, que varían **empresa** a empresa.

Si se da ese tipo de relaciones, se suele recurrir a los estudios de regresión en los cuales se obtiene una nueva relación pero de un tipo especial denominado **función**, en la cual la variable independiente se asocia con un indicador de tendencia central de la variable dependiente. Cabe recordar que en términos generales, una función es un tipo de relación en la cual para cada **valor** de la variable independiente le corresponde uno y sólo un valor de la variable dependiente.

2. ASPECTOS TEÓRICOS

REGRESIÓN SIMPLE Y CORRELACIÓN

La Regresión y la correlación son dos **técnicas estadísticas** que se pueden utilizar para solucionar **problemas** comunes en los **negocios**.

Muchos estudios se basan en la creencia de que es posible identificar y cuantificar alguna Relación Funcional entre dos o más **variables**, donde una variable depende de la otra variable.

Se puede decir que Y depende de X, en donde Y y X son dos variables cualquiera en un **modelo** de Regresión Simple.

"Y es una función de X"

$$Y = f(X)$$

Como Y depende de X, Y es la variable dependiente, y X es la variable independiente.

En el Modelo de Regresión es muy importante identificar cuál es la variable dependiente y cuál es la variable independiente.

En el Modelo de **Regresión Simple** se establece que Y es una función de sólo una variable independiente, razón por la cual se le denomina también Regresión Divariada porque sólo hay dos variables, una dependiente y otra independiente y se representa así:

$$Y = f(X)$$

"Y está regresando por X"

La variable dependiente es la variable que se desea explicar, predecir. También se le llama REGRESANDO ó VARIABLE DE RESPUESTA.

La variable Independiente X se le denomina VARIABLE EXPLICATIVA ó REGRESOR y se le utiliza para EXPLICAR Y.

En el estudio de la relación funcional entre dos variables poblacionales, una variable X, llamada independiente, explicativa o de predicción y una variable Y, llamada dependiente o variable respuesta, presenta la siguiente notación:

$$Y = a.X + b + e$$

Donde:

a es el coeficiente de regresión poblacional (pendiente de la línea recta)

b es el valor de la ordenada donde la línea de regresión se intercepta con el eje Y.

e es el error

SUPOSICIONES DE LA REGRESIÓN LINEAL

1. Los valores de la variable independiente X son fijos, medidos sin error.
2. La variable Y es aleatoria
3. Para cada valor de X, existe una **distribución** normal de valores de Y (subpoblaciones Y)
4. Las variancias de las subpoblaciones Y son todas iguales.
5. Todas las medias de las subpoblaciones de Y están sobre la recta.
6. Los valores de Y están normalmente distribuidos y son estadísticamente independientes.

ESTIMACIÓN DE LA ECUACIÓN DE REGRESIÓN MUESTRAL

Consiste en determinar los **valores** de "a" y "b" a partir de la **muestra**, es decir, encontrar los valores de a y b con los **datos** observados de la muestra. El **método** de estimación es el de **Mínimos Cuadrados**, mediante el cual se obtiene:

$$b = \bar{Y} - a \cdot \bar{X}$$

$$a = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2}$$

Luego, la ecuación de regresión muestral estimada es

$$Y_{est.} = a \cdot X + b$$

3. ANTECEDENTES DEL PROBLEMA

Los datos de la siguiente tabla representan las estaturas (X, cm) y los pesos (Y, kg) de una muestra de 12 hombres adultos. Para cada estatura fijada previamente se observó el peso de una **persona** seleccionada de entre el **grupo** con dicha estatura, resultando:

X	152	155	152	155	157	152	157	165	162	178	183	178
Y	50	61.5	54.5	57.5	63.5	59	61	72	66	72	84	82

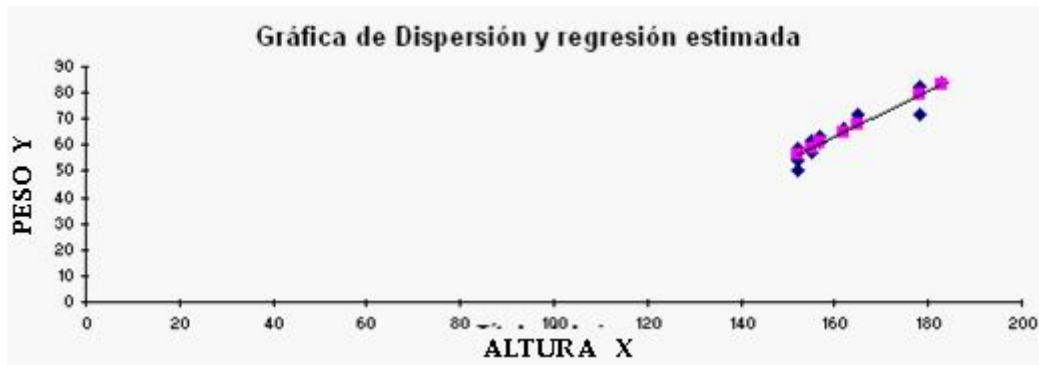
Con estos datos vamos a plantear una ecuación de regresión simple que nos permita pronosticar los pesos conociendo las tallas.

4. DESARROLLO

Representación matemática y gráfica de los datos:

	estatura	pesos	Regresión Lineal				I.C. para la media		I. C. individual		
datos	x	y	x ^2	y ^2	xy	y est.	Residual	L. I.	L. S.	L. I.	L. S.
1	152	50	23104	2500	7600	56.43	-6.43	53.07	59.79	47.30	65.56
2	155	61.5	24025	3782.3	9532.5	59.03	2.47	56.09	61.97	50.05	68.02
3	152	54.5	23104	2970.3	8284	56.43	-1.93	53.07	59.79	47.30	65.56
4	155	57.5	24025	3306.3	8912.5	59.03	-1.53	56.09	61.97	50.05	68.02
5	157	63.5	24649	4032.3	9969.5	60.77	2.73	58.05	63.48	51.85	69.68
6	152	59	23104	3481	8968	56.43	2.57	53.07	59.79	47.30	65.56
7	157	61	24649	3721	9577	60.77	0.23	58.05	63.48	51.85	69.68
8	165	72	27225	5184	11880	67.71	4.29	65.17	70.24	58.85	76.57
9	162	66	26244	4356	10692	65.11	0.89	62.65	67.56	56.27	73.94
10	178	72	31684	5184	12816	78.99	-6.99	74.65	83.33	69.45	88.52
11	183	84	33489	7056	15372	83.32	0.68	78.01	88.64	73.31	93.34
12	178	82	31684	6724	14596	78.99	3.01	74.65	83.33	69.45	88.52

Representación Gráfica



- De acuerdo al **desarrollo** matemático hemos obtenido los siguientes cálculos:

$$n = 12 \quad \bar{X} = 162.167 \quad \bar{Y} = 65.25$$

$$SC(X) = \sum X^2 - n(\bar{X})^2 = 316986 - 12(162.167)^2 = 1409.667$$

$$SP(XY) = \sum XY - n(\bar{X})(\bar{Y}) = 128199.5 - 12(162.167)(65.25) = 1223$$

Lo que nos permite obtener los coeficientes a y b.

Luego,

$$a = 1223 / 1409.667 = 0.8676$$

$$b = 65.25 - (0.8676)(162.167) = -75.446$$

5. INTERPRETACIÓN

- La ecuación de regresión estimada es: $\hat{Y} = -75.446 + 0.8676 X$

Coefficiente de correlación: R= 0.9379

Coefficiente de determinación: R²=0.8796

El valor de a = 0.8676 indica el incremento del peso en kilogramos, en promedio, por cada centímetro de aumento en la estatura de los hombres adultos.

El valor de b, no tiene **interpretación** práctica en el ejemplo, se interpretaría como el valor obtenido, en promedio, para el peso Y, cuando la estatura es 0.

Utilizando la ecuación de regresión para estimar o predecir valores de la variable Y: Para una talla de 180 se obtiene un peso de 80.7 kg.

¿Cuánto se espera que pese (en promedio) una persona que mide 1.60 m?

Sustituyendo el valor de **interés** en la ecuación:

$$\hat{Y} = -75.446 + 0.8676 X \quad \text{Se obtiene:} \quad \hat{Y} = -75.446 + 0.8676 (160) = 63.37 \text{ kg.}$$

6. CONCLUSIÓN

La ecuación de Regresión Lineal estimada para las variables estatura y peso muestran, de acuerdo a la prueba F, relación.

Esta relación se ha estimado en un R = 93.7, que indica una fuerte relación positiva.

Además si consideramos el coeficiente de determinación R² = 87.9 podemos indicar que el 87.9% de las variaciones que ocurren en el peso se explicarían por las variaciones en la variable estatura.

7. ERRORES COMETIDOS POR EL METODO:

Ya que nosotros estamos estimando mediante una recta una nube de valores reales, estamos cometiendo un error, ya que rara vez la recta se ajustará 100% a la realidad. Si fuera éste el caso, se podría decir que el proceso es lineal y 100% dependiente y ajustable por la recta obtenida ($Y = a.X + b$)

Para conocer el promedio del error se puede utilizar:

$$\overline{Err} = \frac{\sqrt{\sum (Y_{calc_i} - Y_{real_i})^2}}{n} \text{ o bien } \overline{Err} = \frac{|Y_{calc_i} - Y_{real_i}|}{n}$$

Es decir, el promedio de los errores (diferencias entre el valor calculado con la ec. de la recta y el valor registrado en la realidad). Usaremos el cuadrado de cada diferencia y la raíz de la suma, o bien el valor absoluto, para evitar que se anulen los positivos y negativos entre sí.

Por último necesitamos saber qué tanto se aleja nuestra estimación de la realidad, es decir, un coeficiente que nos indique qué tan buena es la estimación global.

Definición. En el contexto del análisis de la regresión lineal simple el **coeficiente de correlación múltiple** establece una medida del grado de asociación lineal entre la variable respuesta y la variable predictora, concretamente entre la variable respuesta y la recta de regresión estimada. Se define, a partir de los n pares de observaciones, mediante

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

verificandose que $-1 \leq R \leq 1$

Su cuadrado, R^2 , denominado **coeficiente de determinación múltiple**, puede interpretarse como **el porcentaje de variabilidad de Y explicada o debida a la recta de regresión**, en tanto que puede comprobarse que

$$1 - R^2 = \frac{SSE}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Cuando todos los puntos se encuentran sobre la recta de regresión estimada, es decir, "el ajuste es perfecto", la suma de cuadrados de residuos, SSE, toma el valor cero y, por tanto, $R^2 = 1$. El denominador de la última expresión es una medida de la variabilidad total de las n observaciones de la variable respuesta.