Regresión Lineal Simple

Cuando la relación funcional entre las variables dependiente (Y) e independiente (X) es una línea recta, se tiene una regresión lineal simple, dada por la ecuación

$$Y = ßo + ß1X + \varepsilon$$

donde:

ßo : El valor de la ordenada donde la línea de regresión se intercepta al eje Y.
ß1 : El coeficiente de regresión poblacional (pendiente de la línea recta)
ε : El error.

Suposiciones de la regresión lineal

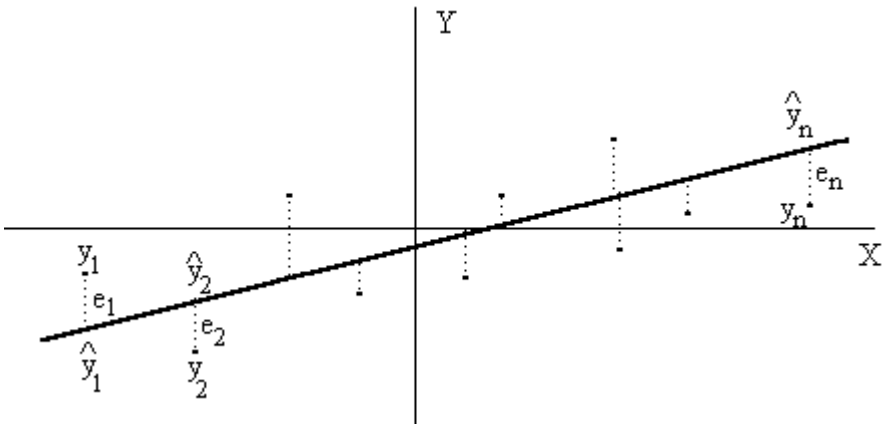
1. Los valores de la variable independiente X son "fijos".
2. La variable X se mide sin error (se desprecia el error de medición en X)
3. Los errores son aleatorios, que se distribuyen normalmente con media cero y variancia σ^2.

Estimación de parámetros

La función de regresión lineal simple es expresado como:

$$Y = ßo + ß1X + \varepsilon$$

Minimizando la suma de cuadrados de los errores, se determinan los valores de bo y b1, así:



$$Q = \sum e_i^2 = \sum \left(y_i - \beta_0 - \beta_1 x \right)^2$$

$$b_o = \overline{y} - b_1 \overline{x}$$

$$b_1 = \frac{spxy}{scx}$$

b0 : es el valor que representa (estimador) a ß0 constituye el intercepto cuando X=0;

b1 : es el valor que representa (estimador) a ß1.

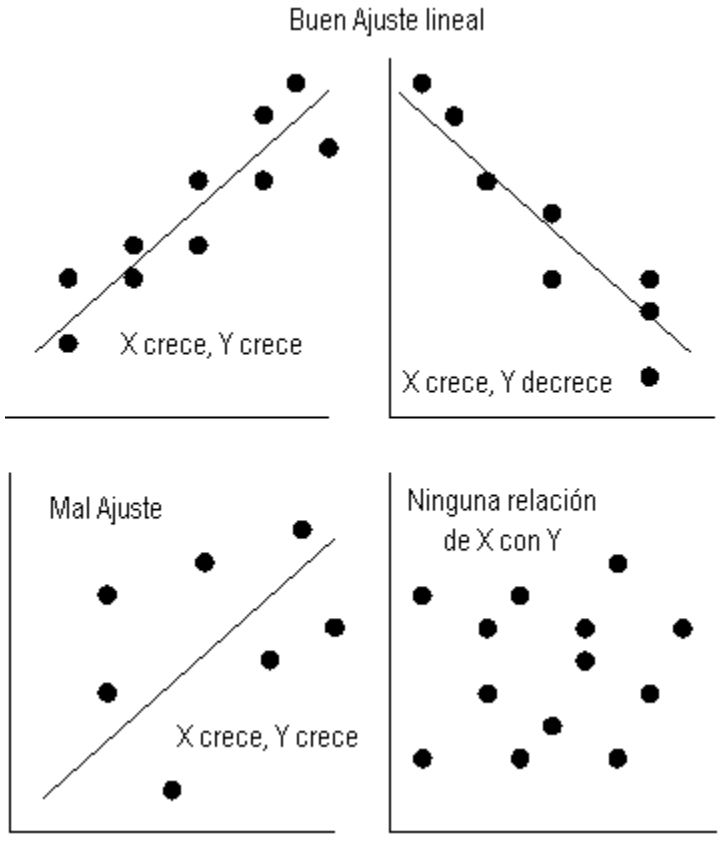
Sus desviaciones estándares respectivas son:

$$Sb0 = \sqrt{\frac{CMresidual.\sum X_i^2}{n.SCX}} \qquad Sb1 = \sqrt{\frac{CMresidual}{SCX}}$$

Luego, la ecuación de regresión es: y = bo + b1X

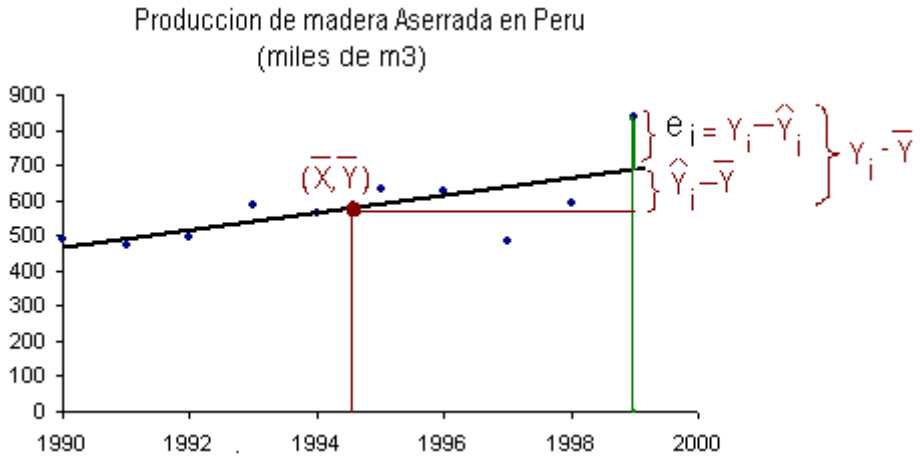
El coeficiente de regresión (b1) .- pendiente de la recta de regresión, representa la tasa de cambio de la respuesta Y al cambio de una unidad en X.

Si b1=0, se dice que no existe relación lineal entre las dos variables.



Fuentes de variación en la regresión lineal

Los cálculos de regresión pueden ser vistos como un proceso de partición de la suma total de cuadrados; así, gráficamente se tiene:



Análisis de Variancia para la regresión lineal simple

Cuadro del ANVA.

Fuentes	Grados de Libertad	Suma de Cuadrados (SC)	Cuadrados Medios (CM)	Fc
Regresión	1	b1.SPXY	b1.SPXY	CM(regresión)/ CM(residual)
Residual: Error	n-2	Diferencia	SC(residual) / (n-2)	
Total	n-1	SC Y		

La prueba estadística “F” evalúa las hipótesis:

Hp: ß1 = 0. No existe una regresión lineal entre X e Y.
Ha: ß1 ≠ 0. Existe regresión lineal de Y en función de X.

Para el ejemplo del grafico (año base 1990 = 0)

Años (X)	0	1	2	3	4	5	6	7	8	9
Madera Aserrada (Y)	489.25	475.24	495.72	585.2	565.78	630.22	624.92	482.27	590.27	834.67

Hp: $\beta = 0$
Ha: $\beta \neq 0$
α=0.05

Mediante el análisis de regresión, se encuentra el siguiente cuadro del Análisis de varianza.

	Gl	*SC*	*CM*	*F*	*F0.05*	*Pr>F*
Regresión	1	49223	492236,9941	5,31	0,0295	
Residual	8	563037037.8				
Total	9	105526				

Si el valor F calculado es mayor o igual al valor tabular; entonces, se rechaza la Hipótesis planteada (Hp), caso contrario se acepta.

Para el ejemplo, Fc = 6,99 es superior a F0.05 = 5.31; entonces, rechazamos la Hp, se concluye que existe una relación lineal entre la producción aserrada entre los años de 1990 a 1999.

Modelo de regresión estimado:

Total de Madera aserrada (miles de m3) = 467,42 + 24,42 X

X = El periodo.

R^2 = (49223 / 105526) *100% = 46%

Intercepto = 467,42
Tasa = 24,42

Significa que el crecimiento anual es de 24 mil metros cúbicos.

Modelos No Lineales.

Se consideran a todos los modelos cuya función es no lineal en los parámetros, por ejemplo:

Modelo exponencial: $Y = \alpha e^{\beta x}$, ln(Y) =Ln(α) + βX;
y1 = a + βX

Modelo Potencial: $Y = \alpha X^{\beta}$, Ln(Y) = Ln(α) + β Ln(X);
y1 = a + β x1

Modelo Logístico:

$$Y = \frac{C}{1 + e^{\alpha + \beta X}};$$

C es el umbral, α y β son parámetros para estimar.

Modelo logístico linealizado : $Ln\left(\frac{C-Y}{Y}\right)=\alpha+\beta X$;

y1 = α+ β X

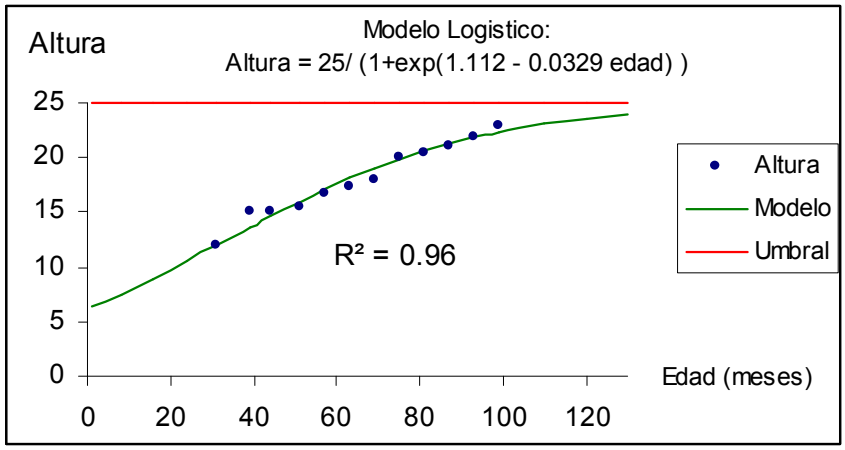
En el siguiente ejemplo se dispone de altura de árboles de Bolaina y la edad en meses desde los 31 meses hasta los 99 meses, la altura en metros.

Se desea ajustar un modelo logístico, bajo el umbral de 25 metros.

Edad	Altura	LN((25-Altura)/Altura)	Modelo
1			6.3
8			7.5
16			8.9
24			10.5
31	12	0.080042708	11.9
39	15	-0.405465108	13.6
44	15.1	-0.422159987	14.6
51	15.4	-0.472604411	15.9
57	16.7	-0.699153205	17.1
63	17.4	-0.828321959	18.1
69	17.9	-0.924705929	19.0
75	20.1	-1.41148461	19.9
81	20.5	-1.516347489	20.6
87	21	-1.658228077	21.3
93	21.8	-1.91875916	21.9
99	23	-2.442347035	22.4
110			23.1
120			23.6
130			24.0

Estimación del modelo linealizado por regresión.

Los estimados son a = 1.11242 y b = - 0.03291
El modelo sería: $Altura = \frac{25}{1 + e^{1.11242 - 0.03291 Edad}}$



Coeficiente de correlación Lineal Simple (r).

Es un número que indica el grado o intensidad de asociación entre las variables X e Y. Su valor varía entre -1 y +1; esto es:

$-1 \leq r \leq 1$.

Si r = -1, la asociación es perfecta pero inversa; es decir, a valores altos de una variable le corresponde valores bajos a la otra variable, y viceversa.

Si r=+1, también la asociación es perfecta pero directa.

Si r=0, no existe asociación entre las dos variables.

Luego puede verse que a medida que r se aproxime a -1 ó +1 la asociación es mayor, y cuando se aproxima a cero la asociación disminuye o desaparece.

El coeficiente de correlación está dada por:

$$r = \frac{SPXY}{\sqrt{SCX.SCY}}$$

Para los datos de la producción de madera aserrada total entre los años 1990 a 1999, existe una asociación de 0.68.

$$r = \frac{2015,17}{\sqrt{(105525,86)(82,5)}} = 0.68$$

Coeficiente de Determinación (R^2)

Mide el porcentaje de variación en la variable respuesta, explicada por la variable independiente.

R^2 = SC regresión / SC total

$0 \leq R^2 \leq 1$.

Interpretación de R^2:

Se interpreta como una medida de ajuste de los datos observados y proporciona el porcentaje de la variación total explicada por la regresión.

R^2 es un valor positivo, expresado en porcentaje es menor de 100.

También, se puede obtener el R^2 ajustado que es la relación entre cuadrados medios, así:

R^2 ajustado = 1 – CME / CM Total;

Este valor podría ser negativo en algunos casos.

Lo que se espera que ambos R^2, resulten similares, para dar una confianza al coeficiente de determinación.

Para el ejemplo, resulta:

R^2 ajustado = 1 – 70378 / (105526 / 9) = 0,39 y R^2 = 1 – 56302,7 / 105525,86 = 0,46