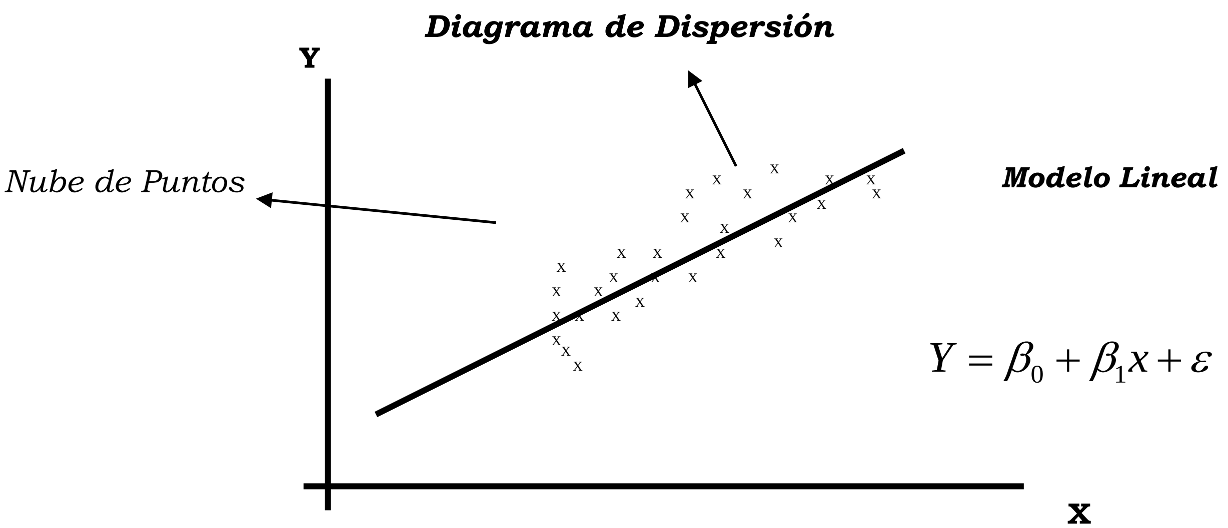
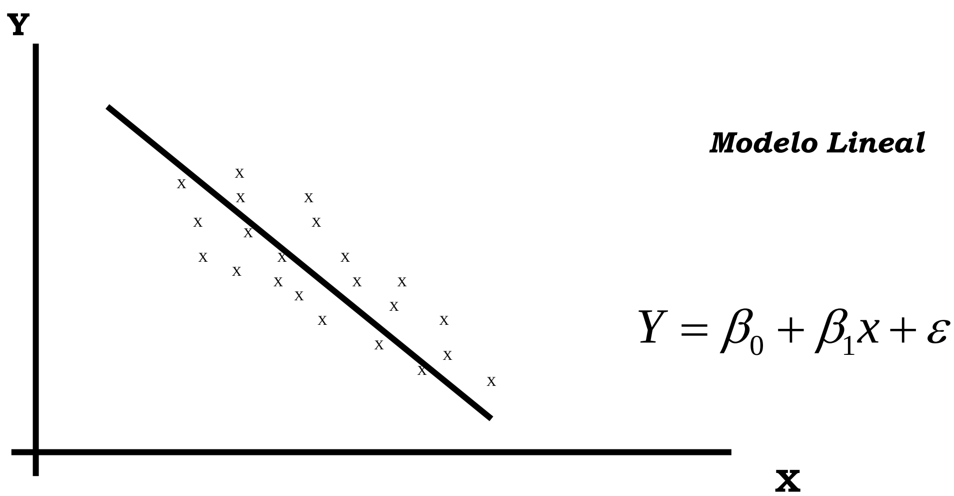
"Análisis de Regresión y Correlación"

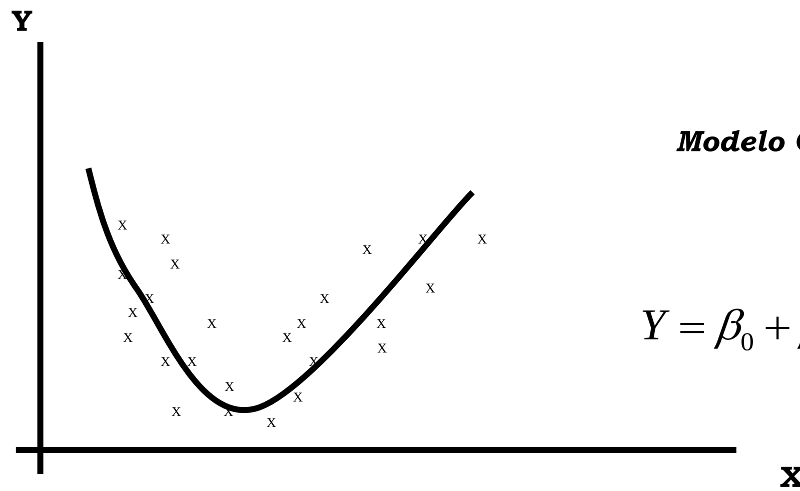
"Análisis de Regresión"

Cuando se tiene una variable respuesta "**Y**" que depende de una variable independiente "**X**"; entonces es posible establecer una ecuación matemática del tipo **Y = f(X)** que explica esta relación de dependencia. Para ello, el investigador fija valores para "**X**" y mide los valores correspondientes de la respuesta "**Y**" obteniéndose así, un conjunto de pares ordenados ($\mathbf{x}_i,\mathbf{y}_i$) donde **i = 1, 2, 3, . . . ,** ***n***; que se procede a graficar.

Datos Experimentales

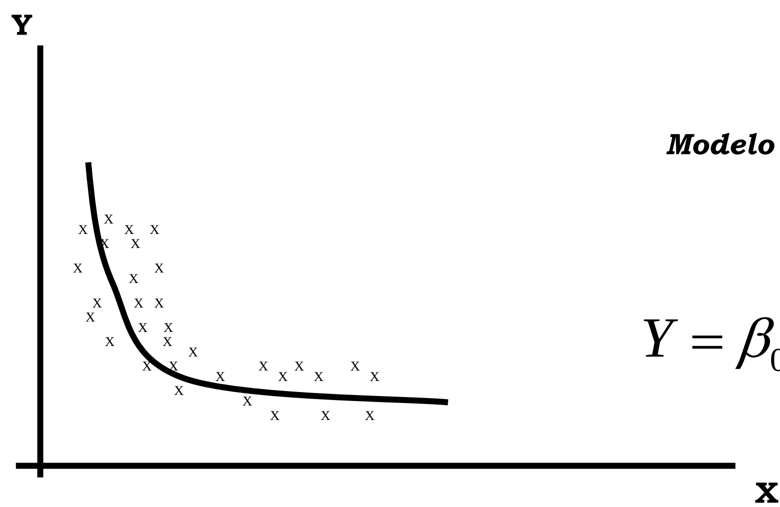






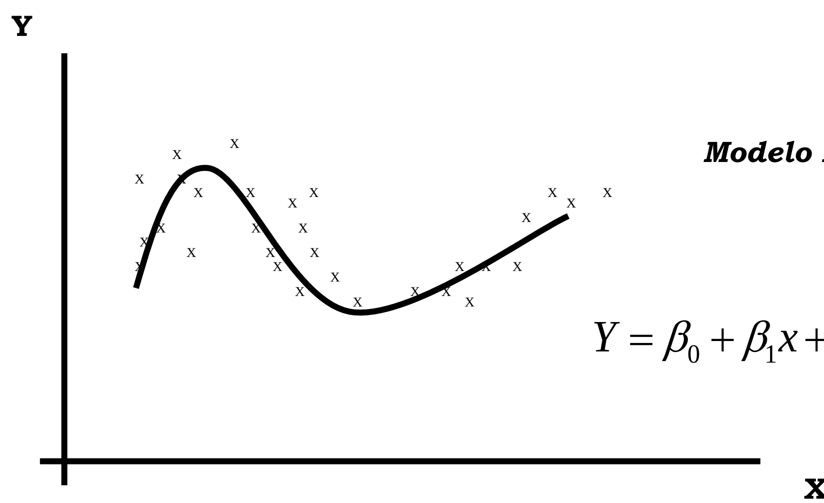
Modelo Cuadrático

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$



Modelo Exponencial

$$Y = \beta_0 e^{\beta_1 x} + \varepsilon$$



Modelo Polinomial

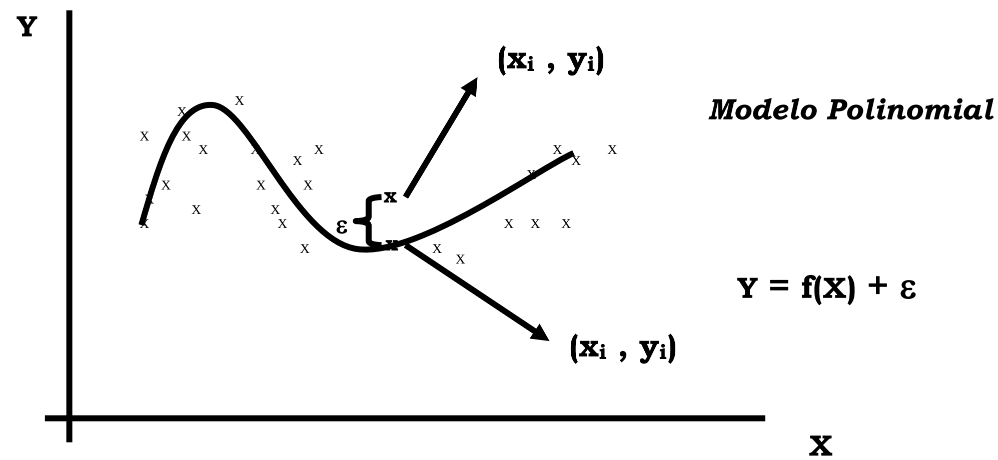
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

¿*Cuál será el comportamiento promedio descrito por la nube de puntos***?**

La forma de elegir y encontrar el modelo matemático (Fórmula Matemática) adecuado que describe el comportamiento promedio descrito por la nube de puntos; se conoce con el nombre de ***"Análisis de Regresión"***.

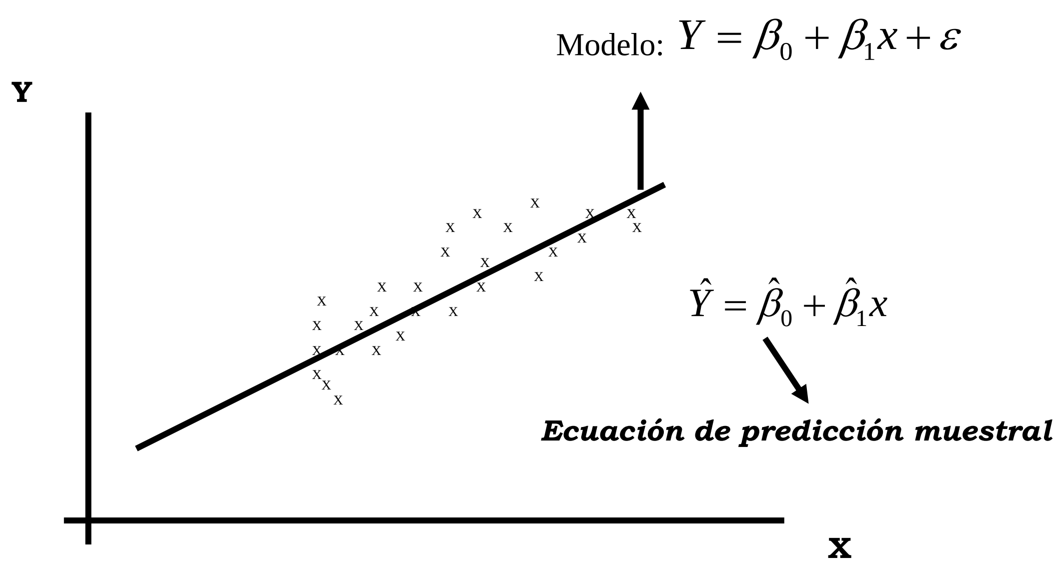
"Método de los mínimos cuadrados"

En el Análisis de Regresión se plantea la interrogante: **¿**Cómo ajustar un modelo matemático a un conjunto de datos**?**

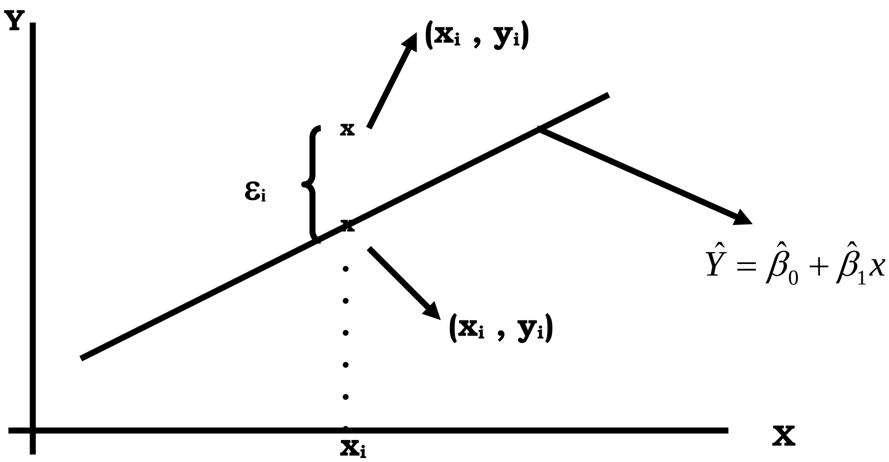


El ajuste será bueno si los errores son mínimos, de manera que el método de los mínimos cuadrados consiste en minimizar a $\sum \varepsilon_i^2$, para ello se deriva e iguala a cero.

"Modelo de Regresión Lineal Simple"



Considérese un solo punto $(\mathbf{x_i}, \mathbf{y_i})$



Luego: $\boldsymbol{\varepsilon_i} = \mathbf{Y_{i\ Observado}} - \mathbf{Y_{i\ Estimado}}$. Es decir, $\varepsilon_i = Y_{i(Observ)} - Y_{i(Estim)}$

$$\therefore \sum \varepsilon_i^2 = \sum \left(y_{i(Observ)} - y_{i(Estim)}\right)^2 \Rightarrow \quad f(\hat{\beta}_0, \hat{\beta}_1) = \sum \varepsilon_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Luego: $$\frac{\vartheta f(\hat{\beta}_0, \hat{\beta}_1)}{\vartheta \hat{\beta}_0} = 2\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \times (-1) \qquad \textbf{(1)}$$

$$\frac{\vartheta f(\hat{\beta}_0, \hat{\beta}_1)}{\vartheta \hat{\beta}_1} = 2\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \times (-x_i) \qquad \textbf{(2)}$$

De ahí que al igualar a cero se tiene:

$$\sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_i = 0 \qquad (1) \quad \text{Nota: } \sum \hat{\beta}_0 = n\beta_0$$

$$\sum x_i y_i - \sum \hat{\beta}_0 x_i - \sum \hat{\beta}_1 x_i^2 = 0 \quad (2)$$

Luego el sistema de ecuaciones para $\hat{\beta}_0$ y $\hat{\beta}_1$ queda:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i & \textbf{(1)} \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i & \textbf{(2)} \end{cases}$$

“Ejemplo 9”: Se quiere analizar la relación existente entre las ventas **“Y”** y el área de exposición **“X”** de los alimentos para niños. Para ello el encargado de mercadeo tomo una muestra de **12** tiendas y fijó áreas de **5**, **10**, **15** y **20 m^2** y midió las ventas semanales. Los datos obtenidos fueron:

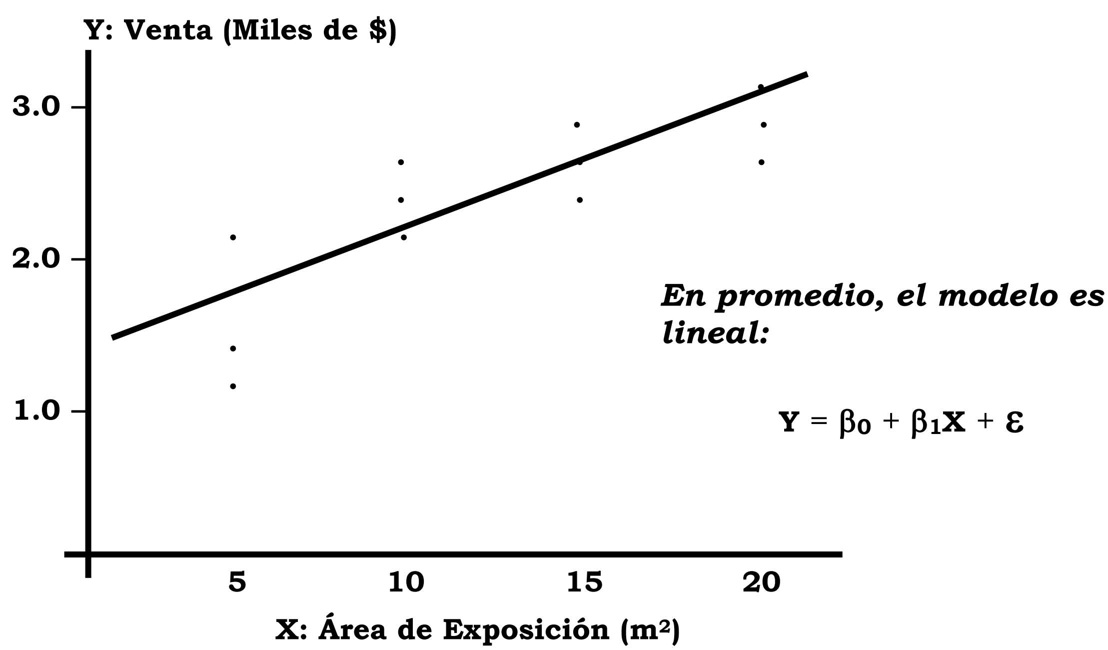
Tiendas	Area: X(m^2)	Ventas: Y (Miles $)
1	5	1.6
2	5	2.2
3	5	1.4
4	10	1.9
5	10	2.4
6	10	2.6
7	15	2.3
8	15	2.7
9	15	2.8
10	20	2.6
11	20	2.9
12	20	3.1

Realice lo siguiente:

- Construya un diagrama de puntos de la situación
- Encuentre la ecuación de predicción muestral $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Predecir las ventas para un área de exhibición de **13m^2** y de **100m^2**.

“Desarrollo”

1.) Diagrama de Dispersión o Nube de Puntos



Se tiene que:

$$n = 12,\quad \sum_{i=1}^{n} x_i = 150,\quad \sum_{i=1}^{n} y_i = 28.5,\quad \sum_{i=1}^{n} x_i^2 = 2250 \quad \text{y} \quad \sum_{i=1}^{n} x_i y_i = 384$$

2.) Ecuación de Predicción Muestral $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{12} x_i = \sum_{i=1}^{12} y_i & \textbf{(1)} \\ \hat{\beta}_0 \sum_{i=1}^{12} x_i + \hat{\beta}_1 \sum_{i=1}^{12} x_i^2 = \sum_{i=1}^{12} x_i y_i & \textbf{(2)} \end{cases}$$

Sustituyendo se tiene:

$$\begin{cases} 12\,\hat{\beta}_0 + 150\,\hat{\beta}_1 = 28.5 \\ 150\,\hat{\beta}_0 + 2250\,\hat{\beta}_1 = 384 \end{cases} \Rightarrow \quad \begin{matrix} \hat{\beta}_1 = 0.074 \\ \hat{\beta}_0 = 1.45 \end{matrix}$$

$\therefore\quad \hat{Y} = 1{,}45 + 0{,}074x$ *Ecuación de Regresión Muestral*

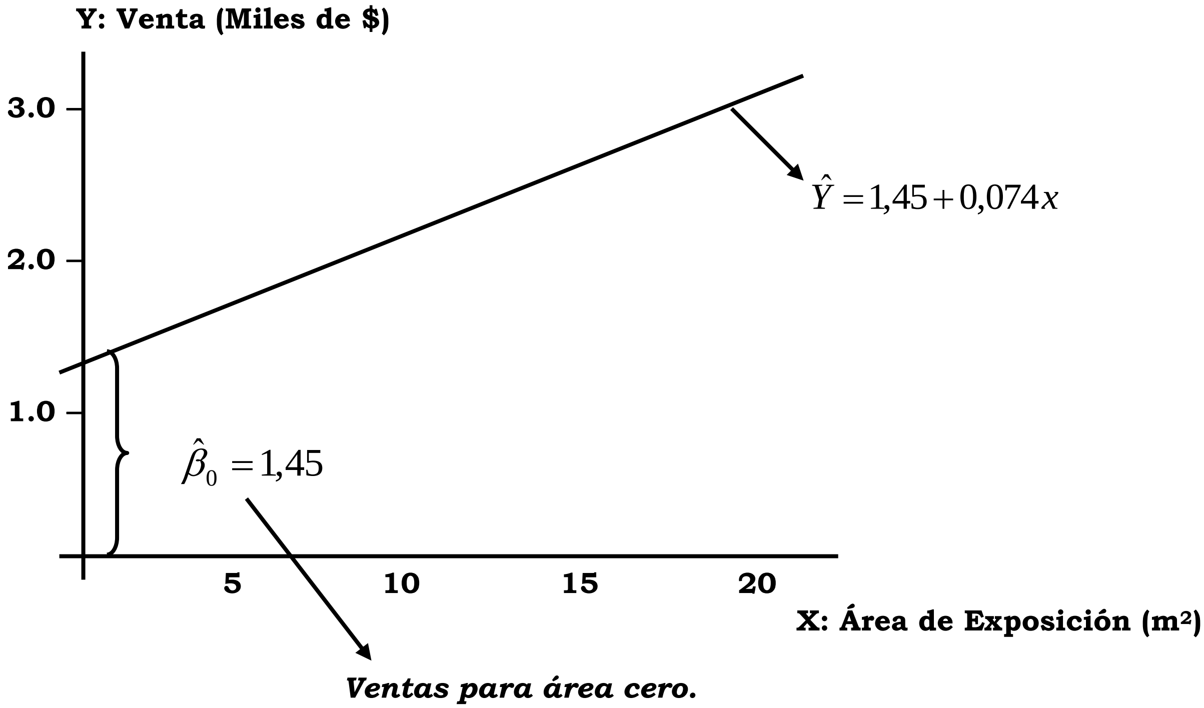
3.) Predicción de las ventas para $\mathbf{x = 13\ m^2}$ y $\mathbf{x = 100\ m^2}$.

Para $\quad \mathbf{Y = 1.45 + 0.074\ (13)}$
$\quad\quad \mathbf{= 2.412}$ ***Miles de $***

Para $\quad \mathbf{Y = 1.45 + 0.074\ (100)}$
$\quad\quad \mathbf{= 8.850}$ ***Miles de $***

Pero $\mathbf{X = 100\ m^2}$ esta fuera del rango de trabajo.

"Un modelo de regresión sólo debe usarse para predecir dentro de un rango de trabajo"



Las ventas para **x = 0** no tienen sentido práctico. El $\hat{\beta}_1 = 0,074$ indica que las ventas aumentan en **0.074** por cada metro que se aumente el área de exposición.

"Coeficiente de Correlación Lineal"

El estudio de la relación de dependencia entre las distintas variables que influyen en una respuesta, se conoce con el nombre de ***"Análisis de Correlación"***. Cuando se quiere calcular el coeficiente de correlación lineal; se usa la fórmula:

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

- Si $r_{XY} \longrightarrow 0$, no hay correlación de tipo lineal entre los tratamientos.
- Si $r_{XY} \longrightarrow \pm 0{,}5$ hay una correlación de tipo lineal media entre los tratamientos.
- Si $r_{XY} \longrightarrow \pm 0{,}7$ hay una correlación de tipo lineal aceptable entre los tratamientos.
- Si $r_{XY} \longrightarrow \pm 0{,}80$ hay una correlación de tipo lineal alta entre los tratamientos.
- Si $r_{XY} \longrightarrow \pm 1$, el modelo lo explica todo; el modelo se acopla en forma casi perfecta.

"Ejemplo 10": Aplicar la fórmula de coeficiente de correlación lineal para determinar si hay o no relación de tipo lineal entre los tratamientos del ***"Ejemplo 9" (Ejemplo anterior).***