Una situación en la cual no podríamos esperar que el modelo lineal

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \ldots \quad (ML)$$

nos fuera útil para describir una relación entre Y y X, puede darse si por ejemplo el diagrama de dispersión no muestra una forma de asociación

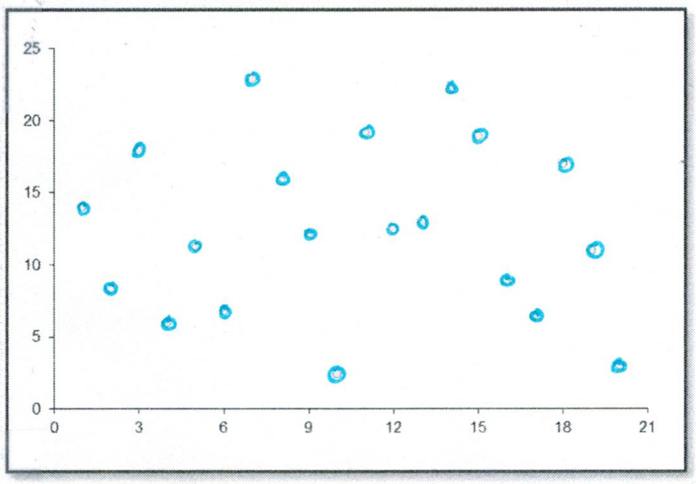


Figura 9. *Sin evidencia de asociación.*

En este caso, de hecho no podemos proponer una relación funcional entre X y Y. Otro caso en el que el modelo (ML) tampoco sería util, es cuando de existir alguna regla de asociación entre X y Y, esta no es lineal, pero si está dada a través de una relación más compleja, quizás un polinomio de orden mayor.

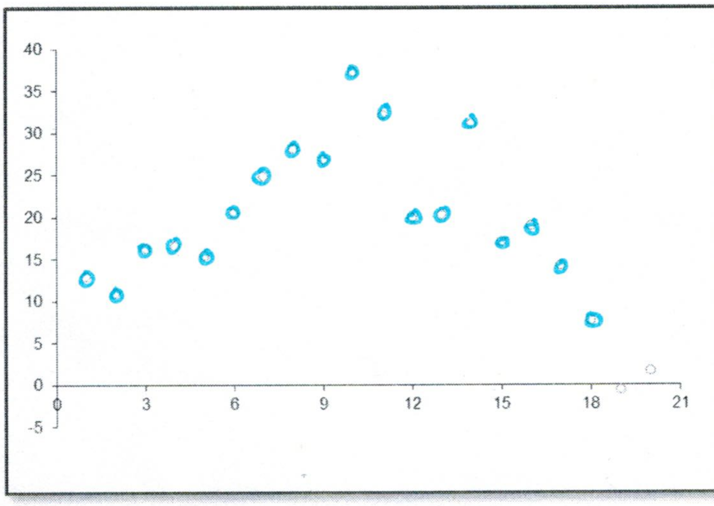


Figura 10. *Tendencia Compleja* (NO LINEAL)

Para este caso se puede intentar "el modelo lineal"(1)

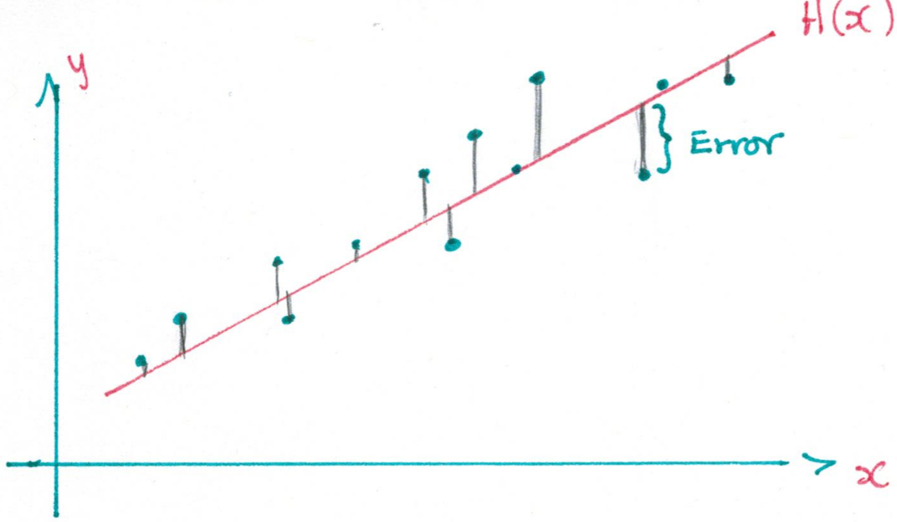
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3.$$

Una vez que se ha encontrado evidencia de una posible relación de asociación entre Y y X, el problema es producir una función que describa esta asociación en forma aproximada. Generalmente un conjunto de datos no se puede describir de manera perfecta con una función, ya que en muchos conjuntos de datos se presentan pares (x_1, y_1) y (x_2, y_2) en donde $x_1 = x_2$ pero $y_1 \neq y_2$, un ejemplo de lo anterior se da en los datos de estaturas y pesos

(1) Más adelante veremos es posible definir $X_1 \equiv X$; $X_2 \equiv X^2$ $X_3 \equiv X^3$ y escribir $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ una función lineal de las variables explicativas X_1, X_2, X_3

dos personas pueden tener la misma estatura pero su peso es diferente. De cualquier manera podemos pensar en encontrar "la mejor aproximación" a la relación entre X y Y usando una curva. Esta curva puede seleccionarse como "la más cercana" a los pares $(x_1, y_1), \ldots, (x_N, y_N)$, como nuestro objetivo es describir a Y para cada valor fijo de X, entonces la distancia apropiada a considerar es, para cada valor fijo x de X, la distancia que existe entre el valor y que le corresponde en los datos y el valor de $H(x)$. Así, si se cuenta con un conjunto de N datos $(x_1, y_1), \ldots, (x_N, y_N)$ y se busca describir que tan cerca se encuentra la función $H(x)$ a la nube de puntos, es posible evaluar la distancia de cada punto en el conjunto de datos a la curva descrita por la función

X	Y	$H(x)$	Error
x_1	y_1	$H(x_1)$	$y_1 - H(x_1)$
x_2	y_2	$H(x_2)$	$y_2 - H(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	y_N	$H(x_N)$	$y_N - H(x_N)$



El problema es hallar $H(x)$ tal que $\|\mathbf{e}\|$ sea lo más pequeña posible

$$\mathbf{e} = \begin{pmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_N - H(x_N) \end{pmatrix}$$

la norma $\|\cdot\|$ puede seleccionarse de entre varias normas posibles, aquí usaremos $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$ ya que este problema ha sido estudiado con esta elección y se han encontrado propiedades matemáticas que permiten un análisis más inmediato[1].

Para el caso particular en que $H(x) = \beta_0 + \beta_1 x$,

(1) Entre otras cosas debido a que se puede utilizar cálculo diferencial para minimizar $\|\mathbf{e}\|_2$

encontrar $H(x)$ que minimize $\|e\|_2$ corresponde con encontrar β_0 y β_1 tales que

$$\|e\|_2 = \sqrt{\sum_{i=1}^{N}(y_i - H(x_i))^2}$$

$$= \sqrt{\sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2} \equiv \sqrt{\Delta(\beta_0, \beta_1)}$$

sea lo más pequeña posible. En otras palabras la tarea es minimizar $\Delta(\beta_0, \beta_1)$ con respecto a β_0 y β_1 [(1)], ya que la raíz cuadrada es una función monótona en los reales no negativos, de donde encontrar β_0 y β_1 que minimizan $\|e\|_2$ equivale a encontrar β_0 y β_1 que minimizan $\Delta(\beta_0, \beta_1)$.

Para encontrar el mínimo de $\Delta(\beta_0, \beta_1)$ y los valores β_0 y β_1 donde este mínimo se alcanza, notemos que

(1) Conocido como el método de Mínimos Cuadrados, esta idea fue generada por Gauss y Legendre a principios del siglo 19

como función de β_0 y β_1, Δ es un polinomio de grado dos y en particular tiene derivadas en todo su dominio (en todo $\mathbb{R}^2$). Entonces podemos usar cálculo diferencial y calcular derivadas parciales con respecto a β_0 y β_1 para encontrar los puntos críticos de $\Delta(\beta_0, \beta_1)$:

$$\frac{\partial \Delta(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$= -2\left\{\sum_{i=1}^{n} y_i - \beta_0 N - \beta_1 \sum_{i=1}^{N} x_i\right\}$$

$$\frac{\partial \Delta(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$= -2\left\{\sum_{i=1}^{N} y_i x_i - \beta_0 \sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2\right\}$$

A continuación debemos resolver el sistema de ecuaciones simultáneas

$$\frac{\partial \Delta(\beta_0,\beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial \Delta(\beta_0,\beta_1)}{\partial \beta_1} = 0$$

Este sistema de ecuaciones es

$$\sum_{i=1}^{N} y_i - \beta_0 N - \beta_1 \sum_{i=1}^{N} x_i = 0$$

$$\sum_{i=1}^{N} y_i x_i - \beta_0 \sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0$$

ó equivalentemente

$$\begin{aligned} \beta_0 N + \beta_1 \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} y_i \\ \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 &= \sum_{i=1}^{N} x_i y_i \end{aligned} \quad \ldots\ldots \quad (EN)$$

El sistema de ecuaciones es conocido como Sistema Normal de ecuaciones (ó ecuaciones Normales) y es un sistema de ecuaciones lineales en β_0 y β_1. Este sistema tiene solución siempre que el discriminante

$$D = \begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix} = N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2$$

$$= N \left(\sum_{i=1}^{N} x_i^2 - N \bar{x}^2 \right)$$

sea distinto de 0. Notemos que D es una función de los datos $x_1, \ldots, x_N$, pero <u>NO</u> depende de los valores $y_1, \ldots, y_N$