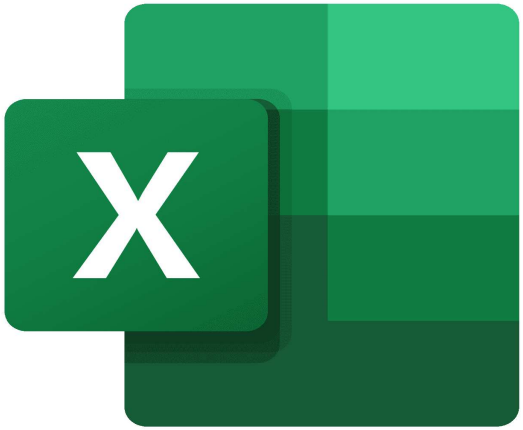


# Estadística con Excel

Héctor Manuel Garduño Castañeda

Octubre, 2021

scidata

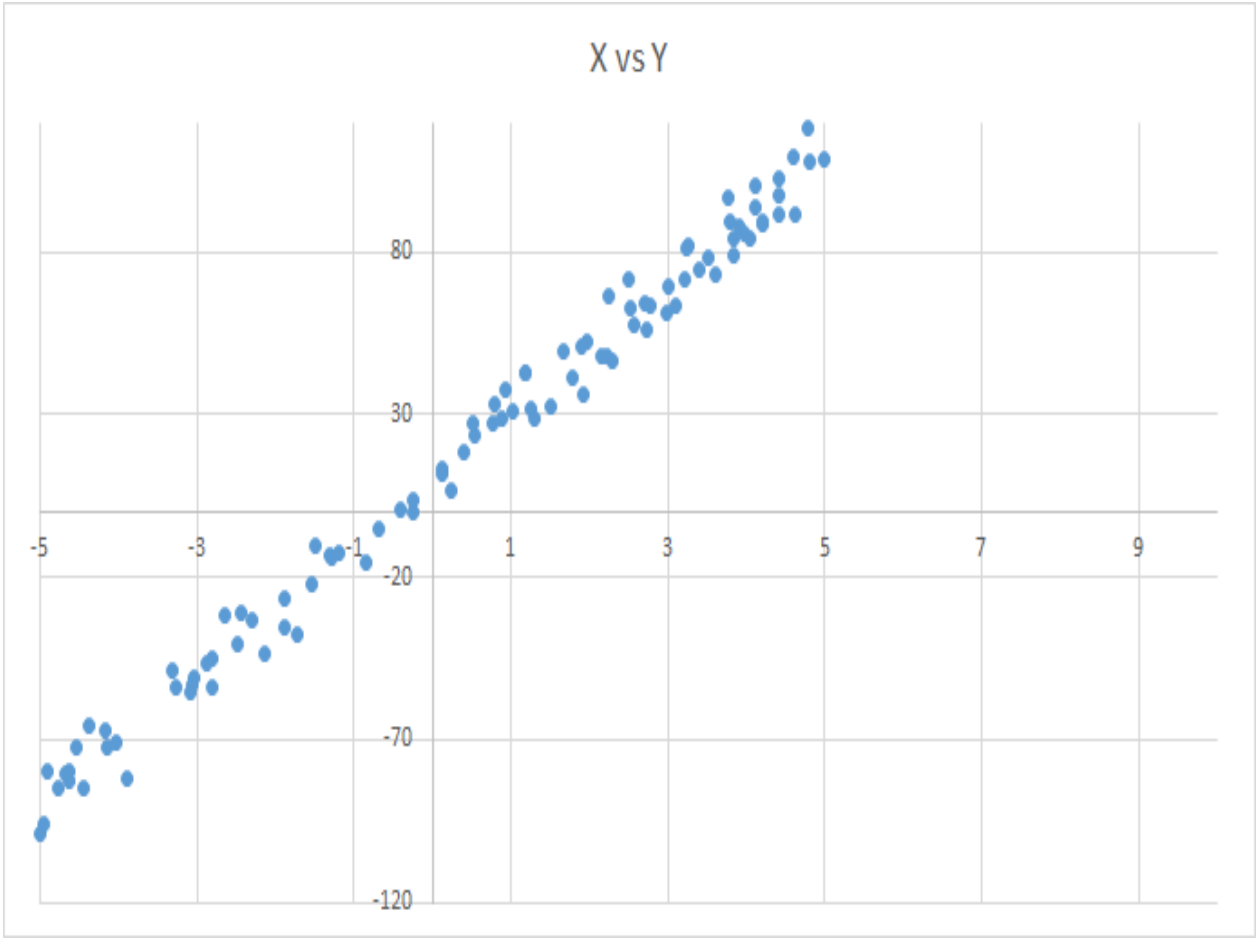
# Contenido

# Conceptos

Las regresiones son maneras de *modelar* una característica medida utilizando otras, también medidas, de la misma muestra con el objetivo de crear predicciones. Esto es: si $X_1, X_2, ..., X_n, X_{n+1}$ son algunas de las columnas de la tabla, encontrar una función $f$ tal que $X_{n+1} = f(X_1, X_2, ..., X_n)$. En cristiano: **¿será posible explicar el comportamiento de una de las características a través del conocimiento de otras?**

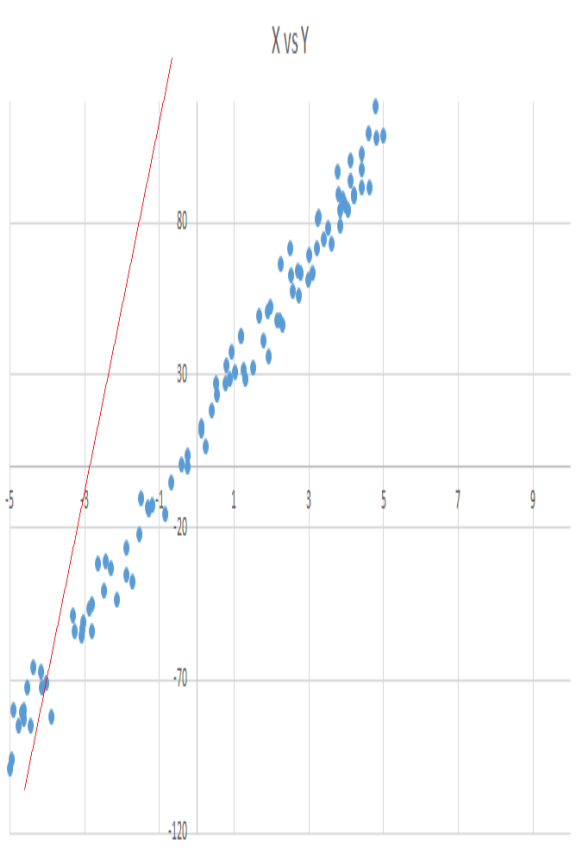
Bajo la idea anterior, decimos que las características $X_1, X_2, ..., X_n$ son **explicativas** y la característica $X_{n+1}$ es la variable objetivo.

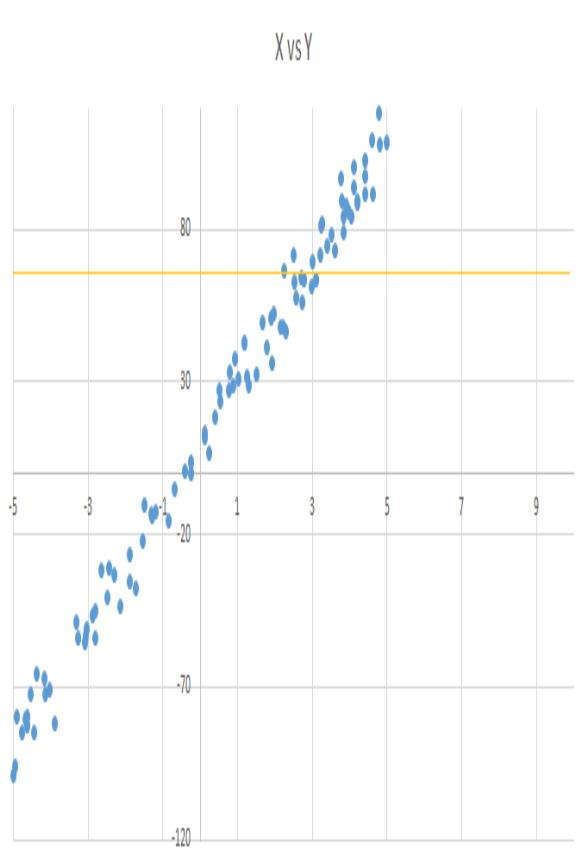
En este capítulo platicaremos de un problema de regresión muy sencillo conocido como **regresión lineal**. Observemos la siguiente nube de puntos:

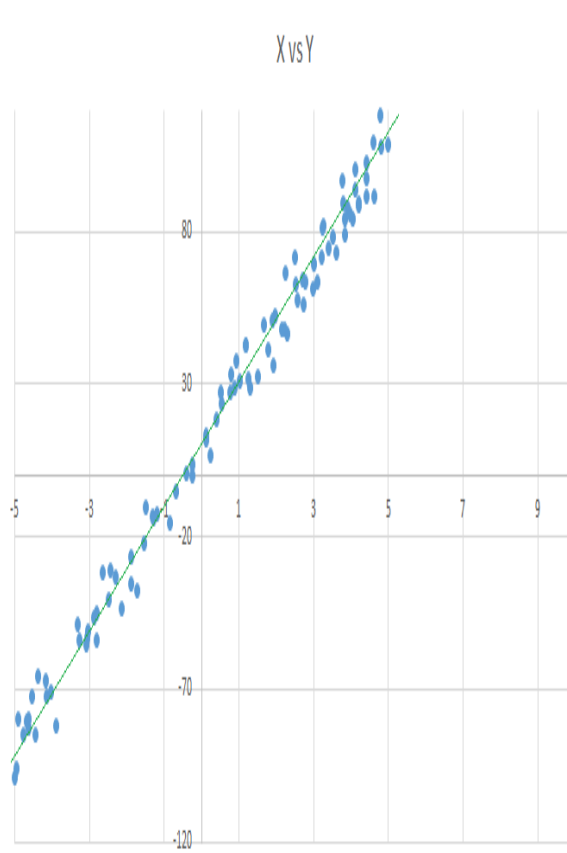



Debido a su forma, vale preguntarse cuál será la recta que mejor se aproxime, en algún sentido, a todos los puntos al mismo tiempo.

Observemos varias rectas graficadas con la nube de puntos. ¿Cuál dirías que es la que más se *ajusta* a todos los puntos al mismo tiempo?



Antes de avanzar, recordemos nuestros cursos de Geometría Analítica: toda recta en el plano es una ecuación de la forma

$$y = mx + b$$

donde

- $b$ se conoce como *ordenada al origen* y es el valor sobre el eje Y en que en la recta lo atraviesa. En regresión lineal se le llama **intercepto**.
- $m$ se conoce como *pendiente de la recta* y se identifica como la tangente inversa del ángulo que hace la recta con el eje X. En cristiano: $m$ mide la inclinación de la recta. También en regresión lineal se llama pendiente. Si $m > 0$ la recta va hacia arriba; si $m < 0$ la recta va hacia abajo; si $m = 0$, la recta es horizontal.

Por lo tanto, **hallar la ecuación de una recta equivale a hallar los valores de $b$ y $m$.**

# Terminología

Recordemos nuestro objetivo: tenemos una lista de parejas de puntos $(x_1, y_1), (x_2, y_2), ..., (x_d, y_d)$. Los graficamos como nube de puntos y buscamos la recta que mejor aproxime, en cierto sentido, a todos los puntos al mismo tiempo. Digamos que $y = mx + b$ es una recta que tomamos como aproximación. Por lo tanto, a cada $x_i$ se le asignan dos números: el $y_i$ (que es un valor que conocemos) y el $\hat{y}_i = mx_i + b$, que es el valor que nos da la recta para ese número $x_i$.

**Predicciones.** Son los valores $\hat{y}_1, \hat{y}_2, ..., \hat{y}_d$. Es decir, los valores que *la recta predice*.

**Residuos.** "'¿Qué tanto se equivocó la recta?"'. Recordemos: la recta le asigna a $x_i$ el valor $\hat{y}_i$. Pero el valor verdadero que acompaña a $x_i$ es $y_i$. Los residuos son los errores que la recta cometió: $\varepsilon_i = y_i - \hat{y}_i$.

Error estándar. Es la desviación estándar de los residuos. También se conoce como **error cuadrático medio** y se interpreta como el error cometido conjuntamente por la recta en todos los puntos.

Coeficiente de determinación. Es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, denotado por $R^2$, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.

Como dijimos anteriormente, se trata de encontrar la recta que mejor aproxime a los puntos en cierto sentido. Esto con el objetivo de poder realizar predicciones.

Es decir, supongamos que tenemos una tabla que contiene al menos dos columnas y $d$ renglones. Además, hay dos columnas, llamadas $X$ y $Y$, de las cuales obtenemos la regresión lineal, lo que significa que encontramos números $m$ y $b$ tales que la recta $y = mx + b$ es la que mejor aproxima a la nube de puntos.

Si *tenemos evidencia para afirmar que el modelo es bueno* (lo que significa que los errores cometidos fueron suficientemente pequeños), entonces al recibir un nuevo renglón $d+1$ del cual desconocemos su valor en la característica $Y$ pero conocemos su valor en la característica $X$, digamos $x_{d+1}$, entonces el valor en $Y$ deberá ser aproximadamente

$$mx_{d+1} + b.$$

Ahora bien... ¿qué significa que la recta sea la que mejor aproxime a todos los puntos al mismo tiempo? Tienes muchas rectas de dónde elegir!! Bueno, cada recta que das te genera una lista de residuos. La suma de todos los residuos te da el error cometido globalmente. Pero eso es precisamente el error estándar. O sea que buscas la recta que te consiga el menor error estándar posible.

Analíticamente, el problema anterior se escribe como: Hallar números $m$ y $b$ tales que la función $C(m,b) = \sqrt{\frac{1}{d}\sum(mx_i + b - y_i)^2}$ sea mínima. Esto es lo que se conoce como **Problema de los mínimos cuadrados**. Para el caso

de la regresión lineal simple, se sabe que si tomamos $m = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sum(x_i - \overline{X})^2}$ y $b = \overline{Y} - m\overline{X}$ entonces la recta $y = mx + b$ es la que mejor aproxima.

# En Excel

Para realizar la regresión lineal en Excel, se usa la aplicación **Regresión** del panel especial de **Análisis de datos** dentro del menú **Datos**.