Semana 4

Nubes de puntos y lineas rectas *Valentín Vergara Hidd*

Este documento funcionará como la introducción a los modelos lineales, particularmente el de regresión. Construiremos algunos conceptos a partir de la nube de puntos del documento anterior, además de incluir la idea de función y el estudios de rectas. Esta es una forma más intuitiva de entender un modelo, que posteriormente analizaremos utilizando algunas herramientas de álgebra lineal.

---

1. De vuelta a la nube de puntos

Pensemos en una nube de puntos, por ahora simplemente considerando variables X e Y en los ejes homónimos.

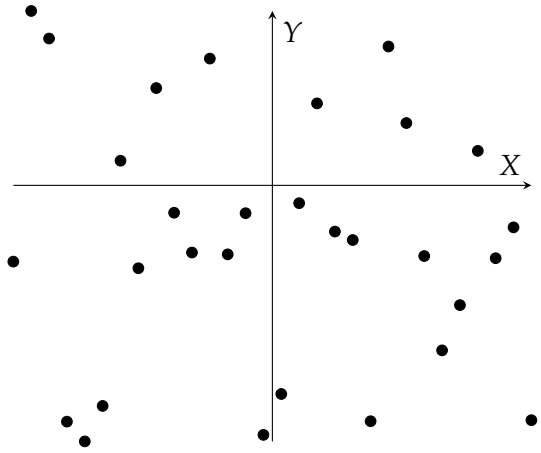

Figura 1: Nube de puntos. Cada punto posee coordenadas (x, y)

Como pueden ver, aparentemente no hay relación entre X e Y. Retomando lo que vimos en el documento anterior, la correlación entre X e Y debería ser cero. ¿Qué pasa si vemos una tendencia en los datos? Otro ejemplo puede mostrar esta tendencia

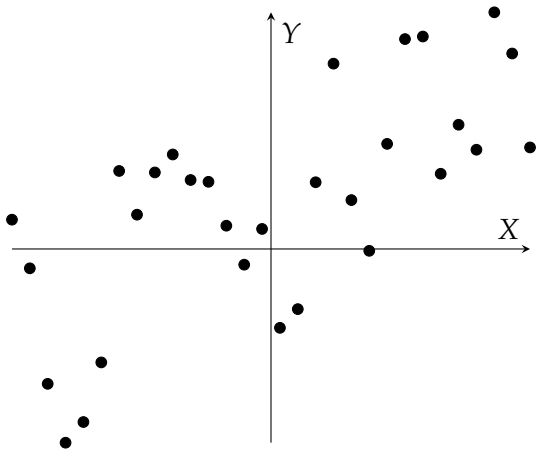

Figura 2: Otra nube de puntos, esta vez se puede ver una tendencia.

Toda la información contenida en la Figura 2 puede ser *resumida* utilizando una línea recta. Veamos algunos ejemplos de lineas que se podrían agregar.

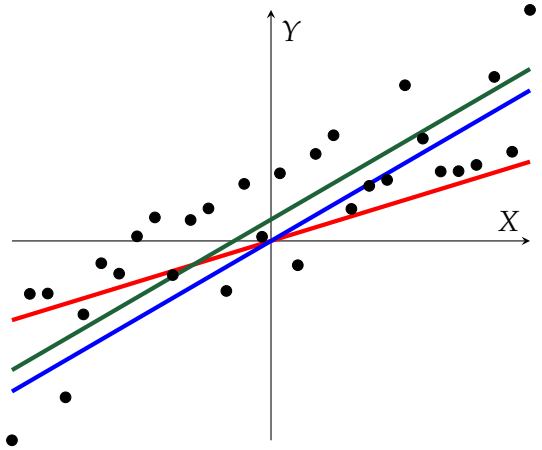

Figura 3: Otra nube de puntos, esta vez se puede ver una tendencia.

Obviamente, la línea roja no es una muy buena representación de la nube de puntos. Lo mismo se podría decir de la línea azul, particularmente porque no pasa a una distancia similar de la mayoría de los puntos. La línea verde cumple este requisito, por lo que podriamos resumir la relación entre X e Y a través de ella. Esto es fundamental, porque permite conocer más acerca de la relación entre ambas variables. Para desarrollar este punto, otro ejemplo: supongamos que tenemos las siguientes nubes de puntos.

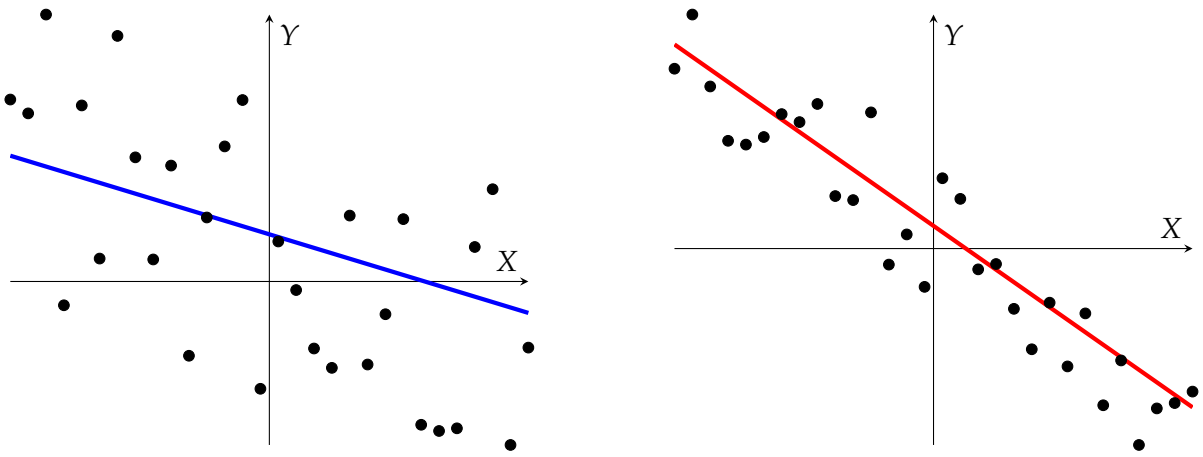

Figura 4: Dos nubes de puntos con una línea que atraviesa a distancia mínima de ellos

Obviamente, la relación es más estrecha en el gráfico de la derecha que en el de la izquierda. Dicho de otra forma, el coeficiente de correlación es negativo en ambos, pero en el gráfico de la derecha está más cerca de -1. Otro detalle importante es que el **valor absoluto de la pendiente** de la línea roja es mayor que la de la línea azul. Sólo para recordar, si se tienen dos puntos de una línea que se debuja en un plano y que identificaremos con las coordenadas (x_1, y_1) y (x_2, y_2), la pendiente β de la recta que se forma al unir los puntos es:

$$\beta = \frac{y_2 - y_1}{x_2 - x_1} \tag{1}$$

Luego, la recta se puede describir con la función:

$$f(x) = y = \iota + \beta x, \tag{2}$$

donde ι es $f(0)$
Como una guía, se puede ver la Figura 5

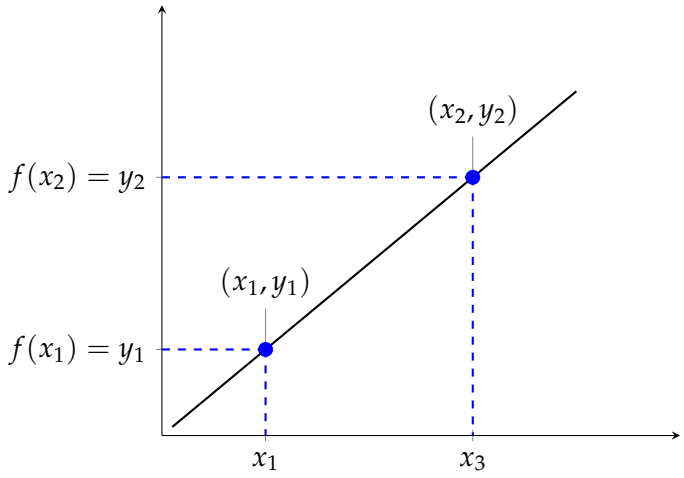

Figura 5: A partir de dos puntos de una recta, se puede obtener su pendiente.

Por tanto, al utilizar la función 2 es posible conocer algunas cosas sobre la relación entre variables. Otro ejemplo, si el valor absoluto de la pendiente es igual o cercano a cero, significa que

no hay correlación entre las variables o que ésta es despreciable. Otro elemento que utilizaremos más adelante es la constante ι, que adquiere relevancia al analizar un modelo lineal. Debido a que $f(0) = \iota$, podemos saber exactamente el punto en que la recta toca el eje vertical, independiente de qué valores se encuentren en la nube de puntos. Sin embargo, es importante contar con la información respecto de la recta para poder llegar a esta conclusión.

En síntesis, podemos conocer bastante de la relaxción entre dos variables, simplemente al trazar una recta que minimice la distancia respecto de los puntos. Luego, sobre esta recta podemos también inferir algunas cosas de la relación entre las variables al mirar los dos parámetros que hemos definido para una línea recta: ι y β. Un punto importantísimo en todo este posible análisis es la forma en que calculamos la recta que pasa a distancia mínima de los puntos. Tal como en la Figura 3, puede haber muchas *posibles* rectas, pero sólo hay una en la que se minimizan la distancia con los puntos. Gran parte del análisis estadístico requerido para un modelo líneal descansa sobre esta idea y efectivamente busca los mejores parámetros ι y β. Este tipo de estimación es la que seguiremos viendo en el próximo documento.