Semana 4

Correlación Lineal *Valentín Vergara Hidd*

Si en el documento anterior revisamos cómo se puede establecer una **relación** entre dos variables, cuando las medimos en términos de un conteo y las reportamos como categorías, también podemos hacer un procedimiento similar para encontrar relaciones (lineales) entre variables que pertenecen al conjunto $\mathbb{R}$.

Además de la utilidad de identificar esta relación en términos exploratorios, es el primer paso para establecer modelos más utiles, siempre en términos lineales, pero que puede dejar de ser exploratorios y pasar a ser explicativos. Por tanto, antes de revisar la construcción de modelos de regresión lineal, veremos simplemente cómo identificar si entre dos variables hay indicios de una relación de este tipo.

---

1. Nubes de puntos

Si sólo consideramos dos variables, podemos revisar sus interacciones a través de una tabla de contingencia, como revisamos en el documento anterior. Sin embargo, esto supone que tenemos que utilizar variables que están agrupadas en categorías. Si queremos utilizar las variables en sus unidades originales, cuando son discretas o incluso continuas, podemos utilizar un gráfico de nube de puntos. La idea de este gráfico es que cada punto corresponde a las coordenadas (x, y) en un plano. Además, x es el valor de una de las variables y y el valor de la otra.

Una consideración importante es que en este caso no existen variables dependientes o independientes, por lo que seleccionamos de forma arbitraria cuál de las variables x y cuál es y. La figura 1 muestra un ejemplo.

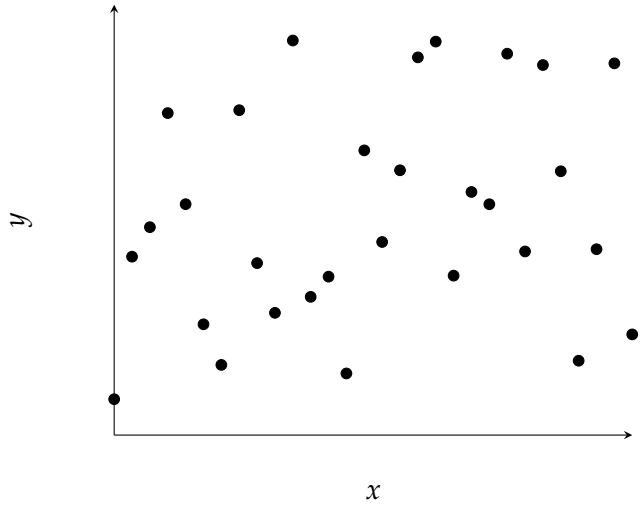

Figura 1: Nube de puntos para dos variables continuas

A veces notamos inmediatamente que la figura tiene *ruido*, es decir, que no se puede ver de

forma clara alguna tendencia. En este caso, podemos sospechar que no existe una correlación entre ambas variables. Otra forma de presentar esta idea es que x es independiente de y. Comparen la Figura 1 con la Figura 2: en 2 hay una tendencia muy identificable. Es más, podemos decir que al aumentar x aumenta también y.

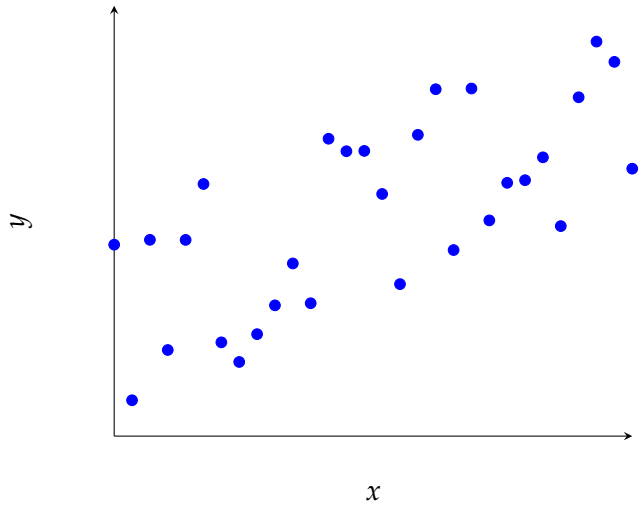

Figura 2: Otra nube de puntos

También podríamos ver la tendencia inversa: al aumentar una variable, la otra disminuye. Por ejemplo, la Figura 3 muestra esta relación.

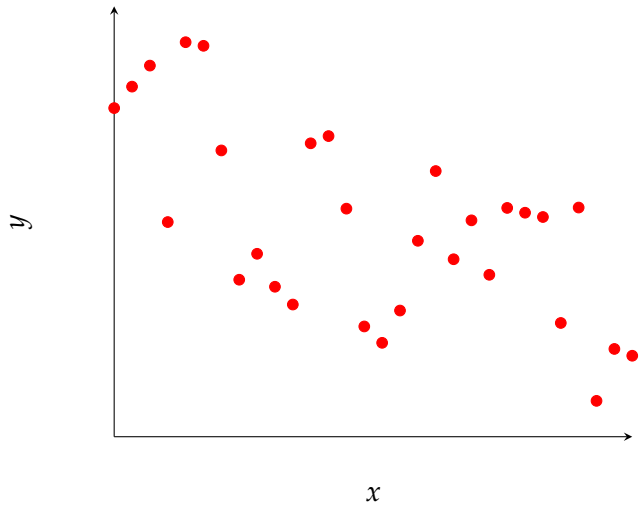

Figura 3: Nube de puntos con tendencia descendente

Hasta ahora, hemos visto todo esto únicamente a través de **inspección visual**, lo que no siempre es adecuado. Es muy sensible a cambios en las dimensiones de los ejes; y por ende, muy susceptible a manipulación. Es por esto que además de un gráfico de nube de puntos, muchas veces se reporta además un **coeficiente de correlación**

2. Coeficiente de correlación

Existen muchas medidas de correlación, pero todas operan más o menos de la misma forma. Se establece un rango en el que la correlación se puede mover, algo así como el **recorrido** de una función, que luego se interpreta en conjunto con el gráfico. Una de las medidas más populares fue presentada a principios del siglo pasado por Karl Pearson. Se basa en la covarianza entre ambas variables; y para normalizar la cantidad se utiliza la desviación estándar de cada variable.

Todo esto general que el **coeficiente de correlación de Pearson**, a veces denominado ρ*, sea una cantidad sin unidad de medida, que toma valores entre -1 y 1. El hecho que no tiene unidades de medida es muy importante, porque permite comparar variables que en un principio no tienen nada que ver entre sí. Formalmente, para dos variables x e y en n muestras, con promedios $\overline{x}$ y $\overline{y}$; y con desviaciones estándar s_x y s_y:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} \tag{1}$$

Noten que el numerador del lado derecho de 1 es la covarianza entre x e y. Por tanto, todos los términos utilizados deberían ser familiares. Otra forma de entender esto es utilizando una expresión equivalente (**?**)

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(Z_{x_i} Z_{y_i})}{n-1}, \tag{2}$$

donde Z_{x_i} es el **puntaje Z** de la variable x y Z_{y_i} el de la variable y. Si no recuerdan que es un puntaje Z, revisen los documentos anteriores, cuando hablamos de la distribución normal.