ESTADÍSTICA DESCRIPTIVA: DOS VARIABLES

Julián de la Horra
Departamento de Matemáticas U.A.M.

1 Introducción

En muchos casos estaremos interesados en hacer un estudio conjunto de varias características de una población. Para fijar ideas y para no complicar la notación supongamos que deseamos estudiar dos características cuantitativas X e Y de una población (consideramos variables cuantitativas porque los conceptos que se van a definir, sólo tienen sentido para ellas). X e Y pueden ser la longitud y la anchura de una especie de insectos, la tasa de inflación y la tasa de desempleo de un país a lo largo de una serie de años, etc.

El objetivo fundamental en este capítulo será encontrar una función lo más sencilla posible que exprese (de manera resumida) la relación que se observe entre X e Y a partir de los datos obtenidos. Nos centraremos en el caso en que esta relación sea de tipo lineal y pueda expresarse razonablemente bien mediante la recta de regresión de Y sobre X. Esta recta de regresión es muy útil porque puede ser utilizada para muchas relaciones no lineales, mediante sencillos cambios de las variables originales.

2 Conceptos básicos y planteamiento

Para hacer el estudio conjunto de las variables cuantitativas X e Y, supondremos que disponemos de una muestra de n pares de observaciones de X e Y:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Es decir, para el elemento i-ésimo de la muestra observamos lo que valen las variables X e Y. Esto es fundamental para poder decir algo sensato sobre la posible relación entre las variables. Igual que en el capítulo dedicado a la Estadística Descriptiva de una variable, no se hará ninguna mención sobre cómo se ha obtenido la muestra. Tenemos en mente la idea de que representa a la población total (de alguna forma), pero esta idea ni se precisará ni se necesitará (de momento).

Por supuesto, se puede hacer un estudio de cada variable por separado y calcular, en particular, medidas de centralización y de dispersión como $\bar{x}$, v_x, $\bar{y}$, v_y. Además, estos valores los necesitaremos más adelante. Pero, como ya hemos indicado, no es éste el objetivo fundamental.

Antes de hacer cualquier cálculo, conviene representar en el plano los pares de valores obtenidos. Con esto obtenemos un diagrama de dispersión con una nube de puntos, que nos puede dar una idea visual de las posibles relaciones existentes.

Además de los conceptos ya estudiados de media y varianza, vamos a necesitar en nuestro estudio el concepto de covarianza; este concepto utiliza las dos variables a la vez.

Definición.- La **covarianza muestral** entre las observaciones de X e Y se define como

$$cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \bullet$$

La manera más sencilla de calcular la covarianza es haciendo un desarrollo similar al de la varianza:

$$\begin{aligned} cov_{x,y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \end{aligned}$$

La covarianza va a aparecer de manera natural al obtener rectas de regresión (un poco más adelante). De momento, es fácil ver que existe una relación entre el signo de la covarianza y el tipo de asociación que hay entre X e Y:

1. Cuando los valores de Y tienden a crecer al crecer los valores de X, decimos que hay una **asociación positiva entre X e Y**. Es fácil razonar gráficamente a partir de la definición de covarianza para ver que, en este caso, la **covarianza será positiva**.

2. Cuando los valores de Y tienden a disminuir al crecer los valores de X, decimos que hay una **asociación negativa entre X e Y**. Es fácil razonar gráficamente a partir de la definición de covarianza para ver que, en este caso, la **covarianza será negativa**.

3. Finalmente, cuando **no parece haber una influencia clara de X sobre Y** (es decir, cuando los valares de X aumentan, no se aprecia ni aumento ni disminución de los valores de Y), también es fácil ver que, en este caso, el valor de la **covarianza será próximo a cero**.

3 Modelo de regresión lineal

Supongamos que la nube de puntos obtenida en el diagrama de dispersión de Y sobre X sugiere una relación lineal entre las variables, bien con una asociación positiva, bien con una asociación negativa entre ellas.

En estos casos, parece bastante razonable intentar resumir toda la nube de puntos mediante una recta; con esta recta se trataría de formalizar la idea de que existe una cierta relación lineal entre los valores de X e Y. Una de las variables jugará el papel de variable independiente (X) y la otra desempeñará el papel de variable respuesta (Y) o variable dependiente de la primera . Esta sección está dedicada a obtener la recta de regresión de Y sobre X.

Definición.- La **recta de regresión de** Y **sobre** X es la recta $y = a+bx$ que minimiza el error cuadrático medio (E.C.M.):

$$E.C.M. = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \qquad \bullet$$

La idea de la recta de regresión es sencilla: intentamos encontrar la recta que mejor representa a la nube de puntos, en el sentido de minimizar la media de los cuadrados de las distancias verticales de los diferentes puntos de la nube a la recta.

El problema de hallar esta recta de regresión se reduce al problema técnico de minimizar una función (E.C.M.) de dos variables (a y b). Eso es lo que hacemos a continuación:

$$\begin{aligned} E.C.M. &= \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 + na^2 + b^2\sum_{i=1}^{n} x_i^2 - 2a\sum_{i=1}^{n} y_i - 2b\sum_{i=1}^{n} x_i y_i + 2ab\sum_{i=1}^{n} x_i\right) \end{aligned}$$

Derivando con respecto a cada variable e igualando a cero, obtenemos el siguiente sistema de ecuaciones:

$$\begin{aligned} \frac{\partial(E.C.M.)}{\partial a} &= \frac{1}{n}\left(2na - 2\sum_{i=1}^{n} y_i + 2b\sum_{i=1}^{n} x_i\right) = 0 \\ \frac{\partial(E.C.M.)}{\partial b} &= \frac{1}{n}\left(2b\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i\right) = 0 \end{aligned}$$

La solución del sistema anterior se obtiene de manera inmediata:

$$a = \bar{y} - \frac{cov_{x,y}}{v_x}\bar{x} \qquad ; \qquad b = \frac{cov_{x,y}}{v_x}$$

Se puede comprobar (pero no lo haremos) que esta solución corresponde a un mínimo de la función. Por tanto, la recta de regresión de Y sobre X es:

$$y = a + bx = \bar{y} - \frac{cov_{x,y}}{v_x}\bar{x} + \frac{cov_{x,y}}{v_x}x$$

En definitiva, la **recta de regresión de** Y **sobre** X se puede escribir de la siguiente forma:

$$y - \bar{y} = \frac{cov_{x,y}}{v_x}(x - \bar{x})$$

4 Evaluación del ajuste

La recta de regresión de Y sobre X que acabamos de estudiar se puede obtener para cualquier conjunto de datos pero, obviamente, en unos casos, esta recta resumirá muy bien la nube de puntos (buen ajuste), y en otros casos, la resumirá peor (mal ajuste). La herramienta numérica que se suele utilizar para evaluar la bondad de este ajuste es el coeficiente de correlación lineal, que se define a continuación.

Definición.- El **coeficiente de correlación lineal entre** X **e** Y se define como:

$$r = \frac{cov_{x,y}}{\sqrt{v_x v_y}} \qquad \bullet$$

El problema inicial del coeficiente de correlación es que, a partir de la definición, no se sabe cuál es su significado. Este significado quedará muy claro en cuanto veamos que el error cuadrático medio cometido con la recta de regresión de Y sobre X se puede expresar en función del coeficiente de correlación lineal:

$$\begin{aligned}
& \text{“Error cuadrático medio cometido con la recta de regresión”} \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y} + \frac{cov_{x,y}}{v_x}\bar{x} - \frac{cov_{x,y}}{v_x}x_i\right)^2 \\
= \;& \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + \left(\frac{cov_{x,y}}{v_x}\right)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\frac{cov_{x,y}}{v_x}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) \\
= \;& v_y - \frac{(cov_{x,y})^2}{v_x} = v_y\left[1 - \frac{(cov_{x,y})^2}{v_x v_y}\right] = v_y(1 - r^2)
\end{aligned}$$

Ahora es fácil decir varias cosas sobre el significado de r, y sobre sus posibles valores:

1. El **coeficiente de correlacion lineal** toma siempre un **valor entre -1 y +1** (ya que el E.C.M., al ser una suma de cuadrados, no puede ser negativo).

2. Cuando el **valor de r es próximo a +1**, el error cuadrático medio cometido con la recta de regresión es próximo a cero y, por tanto, **el ajuste es bueno**. Además, tendremos una **asociación positiva** entre X e Y, ya que la covarianza es positiva (por ser r positivo).

3. Cuando el **valor de r es próximo a -1**, el error cuadrático medio cometido con la recta de regresión es nuevamente próximo a cero y, por tanto, **el ajuste es bueno**. Además, tendremos una **asociación negativa** entre X e Y, ya que la covarianza es negativa (por ser r negativo).

4. Cuando el **valor de r es próximo a cero**, el error cuadrático medio cometido con la recta de regresión se hace mayor y, por tanto, **el ajuste es malo**. Además, observemos que, en este caso, no habrá una clara influencia de X sobre Y, ya que el valor de la covarianza es próximo a cero (por ser r próximo a cero).

5. Finalmente, señalemos que el valor de r siempre hay que tomarlo con precaución ya que resume en un sólo número toda la riqueza de la nube de puntos.