

Capítulo 1

Recordatorio: Estadística descriptiva básica

| | |
|---|-----------|
| 1.1. Representación de datos | 2 |
| 1.2. Medidas descriptivas. Resúmenes cuantitativos | 5 |
| 1.2.1. Medidas de centralización | 5 |
| 1.2.2. Medidas de dispersión (respecto de la media) | 6 |
| 1.2.3. Medidas de asimetría (respecto de la media) | 12 |
| 1.3. Datos multidimensionales. Regresión lineal | 13 |
| 1.3.1. Medidas de dependencia (lineal) | 13 |
| 1.3.2. Ajuste lineal: recta de regresión | 16 |
| 1.3.3. Más de dos variables: regresiones múltiples | 25 |
| 1.4. Comandos de Excel | 30 |
| 1.4.1. Una variable | 30 |
| 1.4.2. Dos variables | 31 |

La estadística descriptiva que en este capítulo se presenta es material básico, que el alumno, en gran medida, ya conoce al menos desde el bachillerato¹.

Los dos primeros apartados 1.1 y 1.2 de este capítulo están dedicados a unificar lenguaje y formalizar notaciones y conceptos (elementales) que permiten describir y resumir datos. En el apartado 1.3 se analiza el modelo de *regresión lineal* para datos bidimensionales (o multidimensionales). Finalmente, el apartado 1.4 contiene algunas instrucciones útiles de Excel.

¹Como quiera que se denomine ahora.

1.1. Representación de datos

Notaciones

De los miembros de una **población** nos interesa una cierta cantidad. Esa cantidad de referencia es la **variable** de estudio, que genéricamente denotamos por X .

Disponemos de un conjunto de valores (que llamamos **datos**) de esa variable en esa población. A ese conjunto de valores nos referiremos como una **muestra**.

Supondremos en lo que sigue que tenemos una *muestra* de una determinada *variable* de nuestro interés en una *población* específica. Por ejemplo,

- saldo medio de un determinado año en cuentas corrientes de 5000 personas físicas residentes en España entre 25 y 45 años,
- notas en examen de entrada en la universidad,
- conexiones diarias a páginas web,
- rendimientos diarios del Ibex35,
- ...

Estas de arriba son **variable cuantitativas**: sus valores son *números* reales con un número determinado de decimales y a veces simplemente números enteros –edad en años–, ¡cuestión de unidades!

Por el contrario, para una **variable categórica** los valores son *categorías*; por ejemplo, grupo sanguíneo, color de ojos, provincia de nacimiento, ...

Notación sistemática. Usaremos sistemáticamente

$$x_1, x_2, \dots, x_n$$

para denotar los datos de una muestra. El **tamaño de la muestra** es el número n de datos.

El orden en que aparecen los datos no será relevante en lo que sigue.

A esos mismos datos ordenados de menor a mayor los denotamos por

$$x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_n},$$

para indicar que se ha utilizado una permutación de $\{1, \dots, n\}$ para ordenar los datos.

Representación de datos

Nos centramos, en lo que sigue, en variables cuantitativas. Digamos que la muestra en cuestión consta de los datos x_1, \dots, x_n . Si el tamaño n de la muestra es grande, suele ser conveniente, para su representación y análisis, agrupar los datos de la muestra en **clases**. De hecho, es frecuente que, en lugar de disponer de los datos en bruto, éstos vengan ya *agrupados* en clases.

La estructura de las clases es discrecional. Si por ejemplo la muestra consiste únicamente de números enteros, por ejemplo ceros y unos, las clases serán, cómo no, la clase del cero y la clase del uno. Si por el contrario los datos son números reales (presentados con una determinada precisión), lo habitual es tomar como clases a intervalos consecutivos de una determinada longitud fija (el **paso o ancho** de la clase), con quizás una primera y una última clases especiales, reservadas para agrupar datos pequeños y datos grandes, respectivamente.

En general, denotaremos las clases por C_1, \dots, C_k , donde k es el número de clases consideradas. Todos los datos de la muestra han de caer en alguna de estas clases.

EJEMPLO 1.1.1. *Muestra de tamaño $n = 10\,000$ de la variable “altura en cm”.*

Fijamos clases de ancho 5 cm,

$$C_1 = [0, 150), \quad C_2 = [150, 155), \dots, \quad C_{14} = [210, 215), \quad C_{15} = [215, \infty),$$

excepto la primera y la última, que son especiales. ♣

Dados unos datos x_1, \dots, x_n y unas clases C_1, \dots, C_k , llamamos

- **frecuencia absoluta** de una clase al número de datos en la clase. Denotamos estas frecuencias absolutas por n_1, \dots, n_k . Así que

$$\boxed{\sum_{j=1}^k n_j = n}$$

- La **frecuencia relativa** (o simplemente frecuencia) de la clase C_j es la proporción $f_j = n_j/n$ que los datos de esa clase C_j representan del total de datos. Así que

$$\boxed{\sum_{j=1}^k f_j = \sum_{j=1}^k \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^k n_j = 1 \quad \Rightarrow \quad \sum_{j=1}^k f_j = 1}$$

- Por **marca de clase** se entiende un valor *representativo* de esa clase: típicamente se escoge el valor medio del intervalo, aunque en los intervalos extremos se suelen elegir marcas en función de cuán dispersos están los datos en esas clases. Usaremos la notación

$$y_1, y_2, \dots, y_k$$

para las marcas de clases, que en el ejemplo 1.1.1 anterior podrían ser

$$y_2 = 152.5, \quad y_3 = 157.5, \dots, \quad y_{14} = 212.5$$

y para las clases de los extremos $y_1 = 145$ e $y_{15} = 220$, por ejemplo.

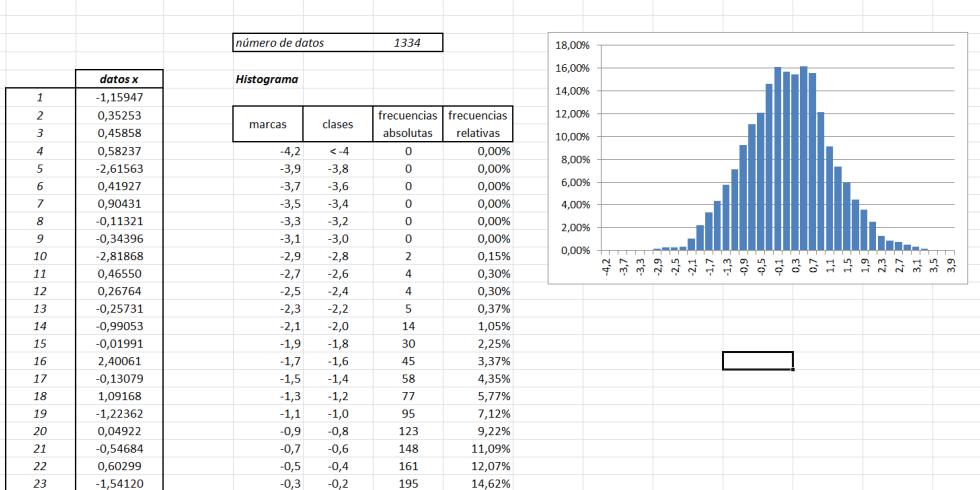
Como ya se ha mencionado más arriba, es frecuente que los datos de la muestra de partida vengan ya agrupados en clases, es decir, que se disponga sólo de

- las clases C_1, \dots, C_k ,
- las marcas de clase y_1, \dots, y_k ,
- y las frecuencias absolutas n_1, \dots, n_k (y por ende², el tamaño n de la muestra y las frecuencias relativas f_1, \dots, f_k);

y no de los datos en bruto x_1, \dots, x_n .

Representación gráfica de los datos. Para datos ya agrupados en clases dadas por intervalos, el **histograma** es un diagrama de barras verticales. Para cada clase, se dibuja una barra con base el intervalo de la clase y con altura la frecuencia de la clase (preferiblemente, frecuencias relativas).

Si los datos no vienen agrupados, conviene agruparlos en clases, tomando por ejemplo un cierto ancho de clase común. Para que el histograma resultante sea informativo, el ancho no debe ser muy pequeño (habría muchas clases, pero probablemente contendrían muy pocos elementos) ni demasiado grande (quedarían pocas clases).



Nota 1.1.1. En un histograma se suele requerir que el área total encerrada por los rectángulos sea 1. Las alturas de los rectángulos, que son las frecuencias relativas f_j , suman 1, pero falta tener en cuenta las longitudes de la bases de los rectángulos. Para por ejemplo comparar (gráficamente) con funciones de densidad, los f_j deberían ser reescalados a $\tilde{f}_j = f_j/b_j$, dividiendo por las longitudes b_j de las clases, de manera que

$$\text{área} = \sum_{j=1}^k \tilde{f}_j b_j = \sum_{j=1}^k f_j = 1.$$

Si, como es habitual, el ancho de cada clase es fijo, $b_j = b$ para cada $j = 1, \dots, k$, se trata sólo de un cambio de escala en la figura.

²Si los datos que se suministran son las frecuencias relativas f_1, \dots, f_k , entonces perdemos la información sobre el tamaño n de la muestra original.



Nota 1.1.2. Suavizado de histogramas. Por razones varias (diluir papel del azar o de errores), a veces se “suaviza” el histograma. Véase la sección 9.2 para más detalles.

Pongamos que las clases son $C_j = [a + (j-1)h, a + jh)$ para $j \in \mathbb{Z}$. El paso es h . Nótese que hay infinitas clases cubriendo todo \mathbb{R} . Para ciertos j_1 y j_2 , las frecuencias de las clases C_j para $j \leq j_1$ y para $j \geq j_2$ serán todas 0. Sea f_j la frecuencia de la clase C_j .

Para suavizar el histograma, podemos tomar, por ejemplo, en lugar de la frecuencia observada f_j , la *frecuencia suavizada*

$$\tilde{f}_j = 0.25f_{j-1} + 0.5f_j + 0.25f_{j+1}.$$

De que $\sum_{j \in \mathbb{Z}} f_j = 1$ se sigue que $\sum_{j \in \mathbb{Z}} \tilde{f}_j = 1$. Este procedimiento asignará frecuencia positiva a algunas clases (a la izquierda de j_1 y a la derecha de j_2) que antes tenían frecuencia 0.

1.2. Medidas descriptivas. Resúmenes cuantitativos

La que sigue es una clasificación (un tanto bizantina, y hasta maniquea) de las distintas medidas cuantitativas que se usan para *resumir* una muestra.

| Medidas de centralización | Medidas de dispersión | Medidas de simetría |
|---------------------------|---|--------------------------|
| media | varianza y cuasivarianza | coeficiente de asimetría |
| mediana | (cuasi)desviación típica rango y cuartiles | |

1.2.1. Medidas de centralización

A. Media

Denotamos la media (aritmética) de los datos de la muestra como

$$(1.1) \quad \boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i}$$



Nota 1.2.1. Si no disponemos de los datos en bruto, sino que ya vienen agrupados en k clases (con frecuencias f_1, \dots, f_k y marcas de clase y_1, \dots, y_k), la media se calcularía como

$$(1.2) \quad \bar{y} = \sum_{j=1}^k f_j y_j.$$

En la fórmula (1.1) se promedian los datos x_i todos con el mismo peso, $1/n$, mientras que en (1.2) se promedian los y_j con las ponderaciones dadas por los f_j .

Estas dos expresiones no coinciden en general. Aunque si la marca y_j de la clase C_j es justamente la media de los datos de la clase C_j ,

$$y_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i,$$

entonces

$$\sum_{j=1}^k f_j y_j = \sum_{j=1}^k \frac{n_j}{n} \frac{1}{n_j} \sum_{x_i \in C_j} x_i = \frac{1}{n} \sum_{j=1}^k \sum_{x_i \in C_j} x_i = \frac{1}{n} \sum_{i=1}^n x_i,$$

y las definiciones (1.1) y (1.2) coinciden.

Obsérvese que siempre

$$\min_i x_i \leq \bar{x} \leq \max_i x_i$$

Además, si transformamos unos datos x_1, \dots, x_n en unos datos z_1, \dots, z_n mediante una transformación (lineal) del tipo $z_i = a + bx_i$, para $i = 1, \dots, n$, donde a, b son contantes, entonces

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n (a + bx_i) = a + b\bar{x}.$$

En estas transformaciones lineales exigimos, como parece razonable, que $b \neq 0$.

B. Mediana

Ordenamos los datos:

$$x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_n},$$

(obsérvese que los datos pueden repetirse). La *voluntad* de la **mediana** es ser el valor que deje tantos datos a la izquierda como a la derecha. Denotamos con MED a la mediana.

Si n es impar, escribimos $n = 2r + 1$ y se toma

$$\text{MED} = x_{i_{r+1}};$$

a la izquierda y a la derecha quedan r datos de la muestra.

Si n es par, escribimos $n = 2r$ y se toma (habitualmente)

$$\text{MED} = \frac{x_{i_r} + x_{i_{r+1}}}{2};$$

a la izquierda y a la derecha quedan r datos de la muestra. En este caso, en general, MED no es un valor de la muestra.



Nota 1.2.2. Aunque no se trate propiamente de una medida de centralización, reseñamos aquí que la **moda** es el valor que aparece con más frecuencia en la muestra; en datos agrupados, la moda es la clase C_j con la frecuencia f_j más alta. Puede haber más de una moda, y podría situarse en los extremos de los valores y no en la zona central.

1.2.2. Medidas de dispersión (respecto de la media)

Estas medias responden a la pregunta: ¿cuán dispersos o cuán variables son los datos?

Digamos, por ejemplo, que medimos una cierta cantidad con dos instrumentos distintos. Con el instrumento A las medidas son $68 = a_1 \leq \dots \leq a_{100} = 72$, con $\bar{a} = 70$. Con el instrumento B , $5 = b_1 \leq \dots \leq b_{75} = 220$, con $\bar{b} = 70$. Las medidas con A están más concentradas que las de B en torno a la media (común) 70. Queremos cuantificar esta diferencia.

Digamos que la muestra es x_1, \dots, x_n . La cantidad $x_i - \bar{x}$ mide la dispersión del dato x_i respecto de \bar{x} . La dispersión total sería

$$\sum_{i=1}^n (x_i - \bar{x})$$

pero esto es

$$\sum_{i=1}^n (x_i - \bar{x}) = \left(\sum_{i=1}^n x_i \right) - n\bar{x} = n\bar{x} - n\bar{x} = 0,$$

así que no informa de nada: las dispersiones positivas y negativas se compensan exactamente.

Para evitar compensaciones, podríamos tomar como medida de dispersión total a

$$\sum_{i=1}^n |x_i - \bar{x}|,$$

que tiene las mismas unidades que los datos originales, y que a veces se usa. Pero para cuestiones analíticas, el valor absoluto $x \mapsto |x|$ es incómodo porque no es derivable en $x = 0$, y preferimos usar cuadrados, como se muestra a continuación.

A. Varianzas y desviaciones típicas

A1. Varianza y cuasivarianza. Dada una muestra x_1, \dots, x_n , se define su **varianza** como

$$(1.3) \quad V_x = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$$

El subíndice hace referencia a la muestra, que genéricamente denotamos por x .

Se trata de un promedio de las dispersiones (de los datos) con respecto a la media, medidas en términos cuadráticos.

Si se divide por $n - 1$ en lugar de por n , se dice **cuasivarianza de la muestra**, que se denota por s_x^2 , haciendo de nuevo referencia a la muestra:

$$(1.4) \quad s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

La razón por la que conviene dividir por $n - 1$ se aclarará más adelante (en el capítulo 4 dedicado a la formalización de la noción de muestreo aleatorio).

La varianza (o la cuasivarianza) mide la dispersión total: no hay cancelación entre los sumandos.

A igualdad de medias $\bar{x} = \bar{z} = m$, si $V_x < V_z$, entonces los datos x se concentran más alrededor de m que los datos z .



Nota 1.2.3. Para datos agrupados (y_j = marca de la clase C_j), definimos

$$V_y = \sum_{j=1}^k f_j (y_j - \bar{y})^2$$

(aquí, la media \bar{y} se habrá calculado según (1.2)).

Para calcular:

- Obsérvese que

$$(n-1) s_x^2 = nV_x$$

- Las expresiones alternativas recogidas en (1.5) para s_x^2 y para V_x resultarán útiles más adelante. Si desarrollamos el cuadrado en la suma de las expresiones anteriores de s_x^2 y V_x se obtiene que

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 &= \frac{1}{n} \sum_{j=1}^n (x_j^2 + \bar{x}^2 - 2x_j \bar{x}) = \underbrace{\frac{1}{n} \sum_{j=1}^n x_j^2}_{:= \bar{x}^2} + \bar{x}^2 - 2\bar{x} \underbrace{\frac{1}{n} \sum_{j=1}^n x_j}_{\bar{x}} \\ &= \bar{x}^2 + \bar{x}^2 - 2\bar{x}^2 = (\bar{x}^2 - \bar{x}^2), \end{aligned}$$

donde, por simplificar notación, llamamos \bar{x}^2 a la media de los datos al cuadrado. De manera que

$$(1.5) \quad s_x^2 = \frac{n}{n-1} (\bar{x}^2 - \bar{x}^2) \quad \text{y} \quad V_x = \bar{x}^2 - \bar{x}^2$$

Letanía al uso: la varianza de los datos es la media del cuadrado de los datos menos el cuadrado de la media de los datos.

Obsérvese que

$$s_x^2 = 0 \iff V_x = 0 \iff \text{todos los } x_i \text{ son iguales}$$

(y, de hecho, si y sólo si $x_i = \bar{x}$, para $1 \leq i \leq n$). Así que $s_x^2 = 0$ o $V_x = 0$ significa dispersión 0.

Veamos. Como $nV_x = \sum_{j=1}^n (x_j - \bar{x})^2$ es una suma de términos no negativos, si $V_x = 0$, entonces $(x_j - \bar{x})^2 = 0$ para cada $j = 1, \dots, n$, y por tanto $x_j = \bar{x}$, para $j = 1, \dots, n$. Recíprocamente, si los x_j son todos iguales, digamos $x_j = m$, para $j = 1, \dots, n$, entonces $\bar{x} = m$ y $(x_j - \bar{x})^2 = 0$, para $j = 1, \dots, n$, por lo que $V_x = 0$.

A2. Desviación típica y cuasidesviación típica. Las unidades de la cuasivarianza son el cuadrado de las de los datos (si datos son metros, cuasivarianza viene en metros cuadrados). Para regresar a las unidades originales, se definen:

$$(1.6) \quad \begin{array}{ll} \text{desviación típica} & \text{cuasidesviación típica} \\ \boxed{\sqrt{V_x} = \sqrt{\frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2}} & \boxed{s_x = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}} \end{array}$$

 **Nota 1.2.4.** Obsérvese la *asimetría notacional*: V_x y s_x^2 (en el lado de las varianzas), y $\sqrt{V_x}$ y s_x (en el lado de las desviaciones típicas).

A3. Cambios de escala. Tenemos una muestra x_1, \dots, x_n , a la que nos referimos abreviadamente como x . Definimos ahora la muestra $a + bx$, que es la que se obtiene a partir de la muestra x con el cambio lineal $x_i \mapsto a + bx_i$, para cada $i = 1, \dots, n$. (Aunque las fórmulas que siguen son válidas para cualesquiera números reales a y b , supondremos siempre que $b \neq 0$.)

Entonces se tiene que

$$\overline{a + bx} = a + b\bar{x}, \quad V_{a+bx} = b^2 V_x \quad \text{y} \quad s_{a+bx}^2 = b^2 s_x^2,$$

Veamos. La primera expresión para la media ya se ha visto antes. Con ella, tenemos que $(a + bx_j - \overline{a + bx}) = b(x_j - \bar{x})$ para $1 \leq j \leq n$, y por tanto,

$$V_{a+bx} = \frac{1}{n} \sum_{j=1}^n (a + bx_j - \overline{a + bx})^2 = b^2 \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = b^2 V_x.$$

A4. Tipificación. Un cambio lineal $x_i \mapsto a + bx_i$ no altera la naturaleza de los datos (es cambio de origen y de escala/unidades).

El cambio

$$x_i \mapsto z_i = \frac{x_i - \bar{x}}{\sqrt{V_x}}$$

transforma los datos x_i iniciales en datos z_i sin unidades (y por tanto, comparables con los de otras muestras). Esta tipificación requiere que $V_x \neq 0$.

Los datos z_i tienen media $\bar{z} = 0$ y desviación típica $\sqrt{V_z} = 1$ (aunque la cuasidesviación típica no sería 1, sino $s_z = \sqrt{n/(n-1)}$).

Los datos z_i se dicen **tipificados** (de los x_i).

A5. Media, varianza y minimización. Tenemos una muestra x_1, \dots, x_n . Buscamos un valor de a que *minimice* la cantidad

$$d(a) = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2.$$

La cantidad $d(a)$ es la dispersión cuadrática (media) de los valores x_i con respecto al punto a . Vamos a comprobar que ese “punto de equilibrio” se alcanza justamente en la media muestral, en $a = \bar{x}$.

Veamos. Reescribimos

$$d(a) = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 + a^2 - 2a \frac{1}{n} \sum_{i=1}^n x_i = a^2 - (2\bar{x})a + \bar{x}^2,$$

así que $d(a)$ es un polinomio cuadrático en a (que es aquí la variable, tanto \bar{x} como \bar{x}^2 son calculados a partir de la muestra).

Esta parábola tiene un (único) mínimo, que se calcula con el procedimiento habitual de cálculo de puntos críticos:

$$0 = d'(a) = 2a - 2\bar{x},$$

lo que nos dice que, como anunciábamos, el mínimo se alcanza en \bar{x} .

Por cierto, el *valor* mínimo³ de la función $d(a)$, es decir, el número $d(\bar{x})$, es justamente la varianza de la muestra V_x .

B. Cuartiles y rango intercuartílico

Se trata de dividir la muestra (una vez ordenada) en cuatro bloques, cada uno con una “cuarta” parte de los datos de la muestra, y marcar dónde se producen las transiciones entre cuartas partes sucesivas.

Pongamos que los datos de la muestra *ya ordenados* son

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

Los cortes entre los bloques son los **cuartiles**. Idealmente se querría que los cortes fueran elementos de la muestra, pero eso, en general, es imposible. Una notación estándar para la información cuartílica es

$$Q_0, Q_1, Q_2, Q_3, Q_4.$$

Para empezar, se toma $Q_0 = \min_i \{x_i\} = x_1$ y $Q_4 = \max_i \{x_i\} = x_n$; el mínimo y el máximo de la muestra. A la diferencia $Q_4 - Q_0$ se le conoce como **rango** de la muestra; es una medida de dispersión de la muestra.

El valor Q_2 es, naturalmente, la mediana (con su casuística en cuanto a su determinación, en función de si n es par o impar).

Querríamos que Q_1 dejara a su izquierda 1/4 de la muestra y a su derecha 3/4 de la muestra y que Q_3 deje 1/4 a su derecha y 3/4 a su izquierda. Una receta simple para determinar Q_1 y Q_3 que cumplan (aproximadamente) esta especificación consiste en:

³Si interpretamos $d(a)$ como una *energía*, la mínima energía se alcanza en \bar{x} , y el valor de ese mínima energía es V_x . Sugerente.

- si n es par $n = 2r$: se toma

$$Q_1 = \text{mediana de } x_1, \dots, x_r, \quad Q_3 = \text{mediana de } x_{r+1}, \dots, x_n = x_{2r};$$

en este caso, $Q_2 = (x_r + x_{r+1})/2$.

- Si n es impar, $n = 2r + 1$, se toma

$$Q_1 = \text{mediana de } x_1, \dots, x_r, \quad Q_3 = \text{mediana de } x_{r+2}, \dots, x_n = x_{2r+1};$$

en este caso, $Q_2 = x_{r+1}$.

Así, si $n = 100$, tendríamos que $Q_1 = (x_{25} + x_{26})/2$, $Q_2 = (x_{50} + x_{51})/2$ y $Q_3 = (x_{75} + x_{76})/2$.

Se llama **rango intercuartílico** a la diferencia

$$\boxed{\text{RIC} = Q_3 - Q_1}$$

Es una medida más de dispersión, y mide la longitud del rango donde se encuentra el 50 % central de la muestra.

Se suelen considerar valores **atípicamente grandes** aquellos valores de la muestra que superan $Q_3 + 1.5 \times \text{RIC}$, y valores **atípicamente pequeños** aquellos valores de la muestra por debajo de $Q_1 - 1.5 \times \text{RIC}$.

Si $Q_2 - Q_1 \ll Q_3 - Q_2$, los valores bajos están concentrados, y los valores altos están dispersos. La comparación de $Q_2 - Q_1$ con $Q_3 - Q_2$ es una medida de asimetría (véase el apartado siguiente).



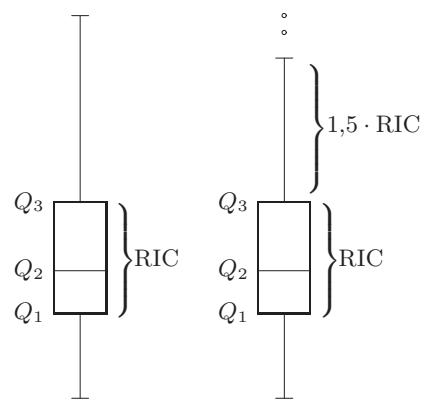
Nota 1.2.5. Una representación gráfica habitual: **diagramas de cajas (boxplots)**, que sirven fundamentalmente para comparar unas muestras con otras.

El primer paso consiste en marcar, en vertical, una caja que va desde altura Q_1 hasta altura Q_3 , marcando una raya horizontal a la altura de la mediana Q_2 . Luego, se trazan unas líneas verticales (conocidas a veces como *bijugates*⁴), una por arriba y otra por debajo, cuyos extremos se pueden marcar de diversas maneras. Dos posibilidades habituales son:

a) prolongar las líneas hasta alcanzar (por debajo) el mínimo de la muestra, y por arriba, el máximo (véase la primera figura de la derecha);

b) o bien

- prolongar la línea superior hasta el mayor dato no atípico (mayor dato inferior a $Q_3 + 1.5 \times \text{RIC}$),
- y la inferior hasta el menor dato no atípico (menor dato superior a $Q_1 - 1.5 \times \text{RIC}$),
- para después señalar los datos atípicos superiores e inferiores. Véase la segunda figura, con símbolos “o” para estos datos atípicos.



⁴No, no es broma.

1.2.3. Medidas de asimetría (respecto de la media)

Queremos medir si los datos de la muestra por encima de la media están más dispersos (respecto a la media) que los que están por debajo.

La dispersión individual sería $(x_i - \bar{x})$; aquí el signo sí es relevante, así que no tomamos valor absoluto ni cuadrado. Pero el promedio de estas cantidades, como hemos visto, es siempre 0. Por eso se suele considerar la cantidad

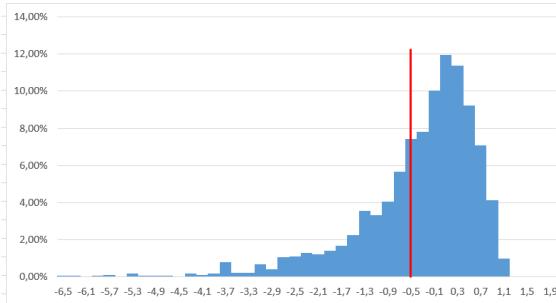
$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3,$$

que también preserva los signos (aunque la potencia 3 distorsiona las unidades, claro). Si esta cantidad es positiva, hay más dispersión a la derecha que a la izquierda de la media, y si es negativa, al revés, más a la izquierda que a la derecha..

Para tener una medida de la asimetría que no dependa de las unidades/dimensiones, se considera el **coeficiente de asimetría**⁵, asim_x , dado por

$$(1.7) \quad \text{asim}_x = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{V_x^{3/2}}$$

Los datos representados en la figura tienen $\bar{x} = -0.5$ y $\text{asim}_x = -1.9$.



- Obsérvese que, si a y b son dos constantes, con $b \neq 0$,

$$\text{asim}_{a+bx} = \text{signo}(b) \text{asim}_x,$$

donde $\text{signo}(b)$ es 1 si b es positivo, y -1 si b es negativo. Es decir, el coeficiente de asimetría no depende de traslaciones; y para los cambios de escala, sólo es relevante el signo (y no su magnitud).

- Para datos x_i tipificados, se tiene que $\text{asim}_x = \frac{1}{n} \sum_{i=1}^n x_i^3$.
- Hay quien define la asimetría dividiendo por s_x^3 .



Nota 1.2.6. Otras medidas posibles de asimetría, también adimensionales, son

$$\frac{\bar{x} - \text{MED}_x}{\sqrt{V_x}} \quad \text{y} \quad \frac{Q_3 + Q_1 - 2 \text{MED}_x}{Q_3 - Q_1}.$$

⁵En inglés, *skewness*.

1.3. Datos multidimensionales. Regresión lineal

Con frecuencia interesa analizar la relación de dos cantidades (variables X e Y) y se dispone de una muestra de datos que son pares (simultáneos) de valores de esas variables. Por ejemplo:

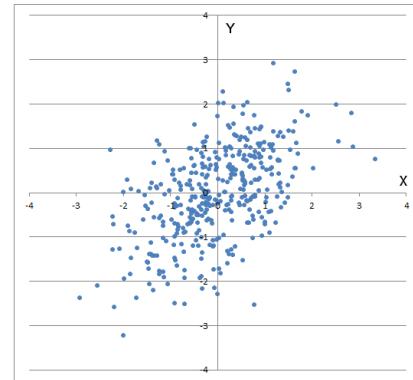
- peso y tensión arterial de población de individuos,
- número de habitantes y gasto por habitante en bibliotecas públicas de una lista de ciudades,
- variación diaria de la cotización de Telefónica y variación diaria del Ibex,
- variación de ventas de Zara y variación del PIB de una sucesión de meses.

Se busca habitualmente explicar una variable en función de la otra, e, incluso, intentar predecir el valor de una variable en función de la otra.

Ahora la muestra consiste de n datos, cada uno de los cuales es un par de números:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

De nuevo, n es el tamaño de la muestra del par (X, Y) de variables. A la muestra de (X, Y) (o a su representación gráfica) se le dice a veces **nube de puntos**. El centro (baricentro) de la nube es el punto (\bar{x}, \bar{y}) .



¡Atención!, antes de explorar los datos como pares, conviene describir (estadísticamente) las **muestras marginales**:

$$x_1, x_2, \dots, x_n, \quad \text{e} \quad y_1, y_2, \dots, y_n.$$

1.3.1. Medidas de dependencia (lineal)

A. Covarianza

La covarianza es la principal medida de dependencia. Se define como sigue:

$$(1.8) \quad \text{cov}_{x,y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Los subíndices hacen referencia a las dos series de datos en cuestión. Las unidades de la covarianza son (unidades de X) \times (unidades de Y).

A.1. Significado del signo de la covarianza. Si en general valores grandes (pequeños) de x_i –es decir, por encima (debajo) de su media \bar{x} – se corresponden con

valores grandes (pequeños) de y_i , por encima (debajo) de su media \bar{y} , entonces, en general,

$$(x_i - \bar{x}) > 0 \text{ se corresponde con } (y_i - \bar{y}) > 0,$$

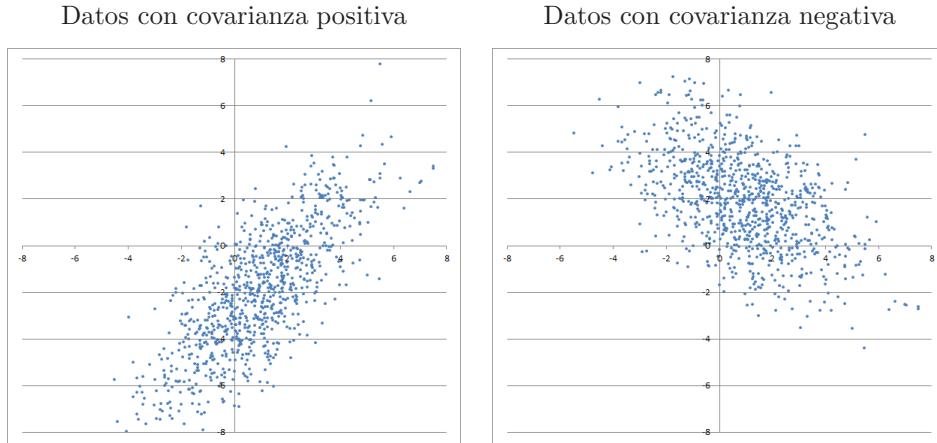
y, claro,

$$(x_i - \bar{x}) < 0 \text{ se corresponde con } (y_i - \bar{y}) < 0.$$

En este caso, se tendrá que los sumandos $(x_i - \bar{x})(y_i - \bar{y})$ son, en general, positivos y, por tanto, $\text{cov}_{x,y} > 0$.

Por el contrario si valores de x_i por encima de \bar{x} se corresponden con valores y_i por debajo de su media \bar{y} entonces se tendrá $\text{cov}_{x,y} < 0$

En términos generales, $\text{cov}_{x,y} > 0$ indica dependencia positiva (relación directa entre las variables X e Y) mientras que $\text{cov}_{x,y} < 0$ indica dependencia negativa (relación inversa entre las variables X e Y).



Abundando en esta interpretación, obsérvese que si para ciertas constantes a, b se tiene que

$$y_i = a + b x_i,$$

es decir, si los pares (x_i, y_i) yacen todos exactamente sobre la recta de ecuación $y = a + bx$, entonces $\bar{y} = a + b\bar{x}$ y

$$\text{cov}_{x,y} = b V_x.$$

A.2. Propiedades de la covarianza.

- $\text{cov}_{x,y} = \text{cov}_{y,x}$.
- $\text{cov}_{x,x} = V_x$, mientras que $\text{cov}_{x,-x} = -V_x$.
- Para constantes a, b, c, d se tiene que $\text{cov}_{a+bx,c+dy} = bd \text{cov}_{x,y}$. Es decir, la covarianza cambia bajo cambios de escala en los datos.

- Desarrollando el producto que conforma cada sumando en la definición (1.8), se obtiene la siguiente fórmula alternativa para la covarianza:

$$(1.9) \quad \text{cov}_{x,y} = \bar{xy} - \bar{x} \bar{y}$$

Aquí, \bar{xy} es la media de la muestra (unidimensional) formada por los datos (producto) $x_i y_i$, para $i = 1, \dots, n$.

Como se ha mencionado antes, la covarianza tiene unidades. Por eso, en muchas ocasiones se utiliza la versión adimensional conocida como:

B. Coeficiente de correlación

La desigualdad de Cauchy–Schwarz en \mathbb{R}^n , que dice que, dados dos vectores *no nulos* $(\alpha_1, \dots, \alpha_n)$ y $(\beta_1, \dots, \beta_n)$, se tiene que

$$\left| \sum_{i=1}^n \alpha_i \beta_i \right|^2 \leq \left(\sum_{i=1}^n \alpha_i^2 \right) \left(\sum_{i=1}^n \beta_i^2 \right),$$

con *igualdad* si y sólo si $\alpha_i = k \beta_i$ para cierto $k \in \mathbb{R}$ y todo $i = 1, \dots, n$ (es decir, si un vector es múltiplo del otro), nos da la acotación

$$(1.10) \quad |\text{cov}_{x,y}| \leq \sqrt{V_x} \sqrt{V_y}$$

pues

$$|\text{cov}_{x,y}|^2 = \left| \sum_{j=1}^n \frac{(x_j - \bar{x})}{\sqrt{n}} \frac{(y_j - \bar{y})}{\sqrt{n}} \right|^2 \leq \left(\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n} \right) \left(\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n} \right) = V_x \cdot V_y.$$

Además, se tiene igualdad si y sólo si el vector de coordenadas $(x_i - \bar{x})$ es un múltiplo del vector de coordenadas $(y_i - \bar{y})$; es decir, si y solo si se tiene que para ciertas constantes a, b se da que $y_i = a + b x_i$, para cada $1 \leq i \leq n$; esto es, si los datos (x_j, y_j) están todos sobre una misma recta.

Se define el **coeficiente de correlación** $\rho_{x,y}$ como

$$(1.11) \quad \rho_{x,y} = \frac{\text{cov}_{x,y}}{\sqrt{V_x} \sqrt{V_y}}$$

El coeficiente de correlación *no tiene unidades*.

Esta definición, claro, requiere $V_x \neq 0$ y $V_y \neq 0$. Es decir, que ni los datos x_i ni los y_i sean constantes.

B.1. Propiedades del coeficiente de correlación.

- $\rho_{x,y} = \rho_{y,x}$.
- $\rho_{x,x} = 1$ y $\rho_{x,-x} = -1$.

- $\rho_{a+bx,c+dy} = \rho_{x,y}$. Es decir, el coeficiente de correlación es invariante por traslaciones y cambios de escala.
- El signo de $\rho_{x,y}$ es el mismo que el de $\text{cov}_{x,y}$.
- Por la desigualdad de Cauchy–Schwarz (1.10) se tiene siempre que

$$(1.12) \quad -1 \leq \rho_{x,y} \leq +1.$$

De manera que el coeficiente de correlación es un número entre -1 y 1 . O en la terminología habitual, entre -100% y 100% .

- Por el caso de igualdad de Cauchy–Schwarz se tiene que $|\rho_{x,y}| = 1$ si y solo si $y_i = a + bx_i, 1 \leq i \leq n$, para ciertas constantes a y b . El signo de b es igual al valor de $\rho_{x,y}$ ($+1$ o -1).

1.3.2. Ajuste lineal: recta de regresión

Tenemos una muestra de tamaño $n \geq 2$ del par de variables (X, Y) :

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Planteamos la siguiente cuestión: de entre todas las rectas $y = a + bx$, ¿cuál es la que “mejor” aproxima/explica la muestra?

A. La recta de regresión

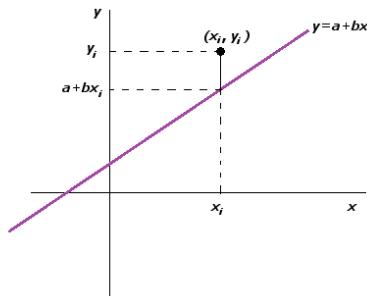
Suponemos de partida que $n \geq 2$, claro, y también que $V_x \neq 0$ y que $V_y \neq 0$. Si alguna de las varianzas es 0, los datos de la muestra correspondiente son todos iguales, y el ajuste es bien directo: por ejemplo, para el caso de $V_x = 0$, los datos x son constantes, de manera que la nube de puntos se ubica en una recta vertical (que claramente es la recta de mejor ajuste). En el otro caso, $V_y = 0$, la nube de puntos está sobre una recta horizontal.

Tenemos dos grados de libertad a y b . Definimos la función que a cada $(a, b) \in \mathbb{R}^2$ le asocia $E(a, b)$ dada por

$$(1.13) \quad E(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + b x_i))^2$$

La notación alude a que $E(a, b)$ es el **error cuadrático medio** que se comete al reemplazar y_i por su pretendida “explicación” $a + bx_i$. En términos geométricos, para cada i tomamos un error *vertical*. Intentamos explicar las muestras de Y en función de X ; los papeles de X e Y *no son simétricos*: Y es la variable que se pretende explicar, y X es la variable “explicativa”.

Usamos cuadrados por conveniencia analítica (de cálculo).





Nota 1.3.1. Tenemos, por un lado, el vector $(y_1, \dots, y_n) \in \mathbb{R}^n$ procedente de la muestra, y por otro el vector $(a + bx_1, \dots, a + bx_n) \in \mathbb{R}^n$, que sería el vector de imágenes de (x_1, \dots, x_n) si el modelo lineal se cumpliera. En la fórmula (1.13) estamos calculando la distancia euclídea (en \mathbb{R}^n) entre estos dos vectores. El objetivo es hallar a y b que minimicen esa distancia.

Podemos escribir (1.13) como

$$E(a, b) = a^2 + \bar{x}^2 b^2 - 2\bar{y}a + 2\bar{x}ab - 2\bar{x}\bar{y}b + \bar{y}^2,$$

para hacer patente que $E(a, b)$ es un polinomio cuadrático en a y b . Las expresiones que involucran medias de x e y (o de los cuadrados, o de xy) son los coeficientes, y las variables son a y b . La función $E(a, b)$ es no negativa, como resulta evidente a partir de su definición (1.13) (aunque, je, no tanto en la última fórmula).

En la siguiente reescritura

$$E(a, b) = \underbrace{(a, b) \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}}_{\text{forma cuadrática}} \underbrace{-2\bar{y}a - 2\bar{x}\bar{y}b + \bar{y}^2}_{\substack{\text{términos lineales} \\ \text{constante}}}$$

separamos las distintas contribuciones a $E(a, b)$. La matriz de la forma cuadrática tiene, como determinante, a la varianza V_x (¡ups!). En cualquier caso, es definida positiva, y por tanto $E(a, b)$ tiende a $+\infty$ cuando $(a, b) \rightarrow \infty_{\mathbb{R}^2}$ (es decir, cuando $\sqrt{a^2 + b^2} \rightarrow +\infty$). Este análisis preliminar garantiza que $E(a, b)$ tiene (al menos) un mínimo.

Como $E(a, b)$ es (infinitamente) diferenciable, esos mínimos deberán estar en los puntos críticos, cuya ubicación determinamos igualando a 0 las dos siguientes ecuaciones:

$$\begin{cases} \frac{\partial E(a, b)}{\partial a} = 2a + 2\bar{x}b - 2\bar{y}, \\ \frac{\partial E(a, b)}{\partial b} = 2\bar{x}^2b + 2\bar{x}a - 2\bar{x}\bar{y}, \end{cases}$$



Nota 1.3.2. En consonancia con la discusión de más arriba, observamos que el hessiano de E es la matriz $H_E = \begin{pmatrix} 2 & 2\bar{x} \\ 2\bar{x} & 2\bar{x}^2 \end{pmatrix}$, que es definida positiva. Obsérvese que $\det(H_E) = 4V_x > 0$.

Así que el único punto crítico es la solución del sistema

$$(1.14) \quad \begin{cases} \bar{y} = a + b\bar{x}, \\ \bar{x}\bar{y} = a\bar{x} + b\bar{x}^2 \end{cases}$$

Despejando (multiplicando por ejemplo la primera ecuación anterior por \bar{x} y restándole la segunda, etc.), deducimos que el mínimo de E se alcanza en el punto (\hat{a}, \hat{b}) , donde

$$(1.15) \quad \boxed{\begin{aligned} \hat{b} &= \frac{\text{cov}_{x,y}}{V_x} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} = \bar{y} - \left(\frac{\text{cov}_{x,y}}{V_x}\right)\bar{x} \end{aligned}}$$

La recta de ecuación

$$(1.16) \quad y = \hat{a} + \hat{b}x$$

que da el mínimo error cuadrático medio, es la **recta de regresión de Y sobre X** .



Nota 1.3.3. El nombre proviene del artículo *Regression towards mediocrity in hereditary stature*, publicado por Galton en 1886 en una revista de Antropología. La expresión “regression towards mediocrity” describe la tendencia a regresar o retornar a estados menos extremos, tras observar (Galton) que de padres altos salían hijos más bajos; y que de padres bajos, salían hijos más altos. En los términos de la ecuación (1.16), el papel de x se reserva para la altura de los padres (promedio de las alturas de los dos padres, tomaba Galton), y la y refleja altura de los hijos.

246

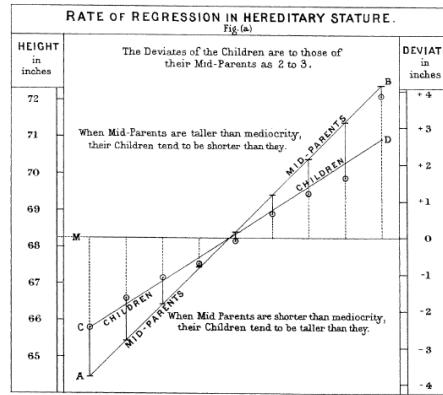
Anthropological Miscellanea.

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION towards MEDIOCITY in HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

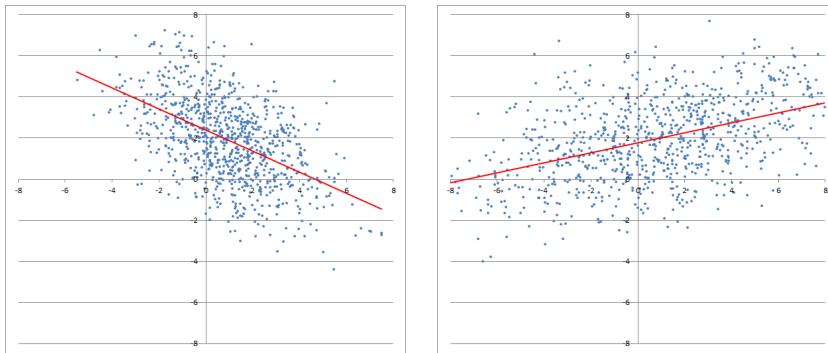
[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in “Nature,” September 24th. I reproduce here



El parámetro \hat{a} da el corte de la recta de regresión con el eje vertical OY . El parámetro \hat{b} es la pendiente, que tiene el mismo signo que $\text{cov}_{x,y}$.

Ambos parámetros \hat{a} y \hat{b} (y por tanto la recta de regresión) están unívocamente determinados por la muestra.



Usando que $\bar{y} = \hat{a} + \hat{b}\bar{x}$ (véase la primera ecuación de (1.14), o la propia solución (1.15)), observamos que la ecuación de la recta de regresión puede escribirse

como

$$(1.17) \quad y - \bar{y} = \hat{b} (x - \bar{x})$$

que nos dice que la recta de regresión pasa por el punto (\bar{x}, \bar{y}) , el centro de la nube de puntos.

Supongamos, por ejemplo, que $\hat{b} > 0$. Si $\hat{b} < 1$, entonces la pendiente de la recta queda por debajo de la de la bisectriz del primer cuadrante, y se produce el fenómeno de “regresión a la mediocridad” de que hablaba Galton. En caso contrario, si $\hat{b} > 1$, se produce una dispersión hacia los extremos.

Escribiendo la pendiente \hat{b} como

$$(1.18) \quad \hat{b} = \rho_{x,y} \frac{\sqrt{V_y}}{\sqrt{V_x}}$$

obtenemos la (sugerente) escritura alternativa para la recta de regresión:

$$(1.19) \quad \frac{y - \bar{y}}{\sqrt{V_y}} = \rho_{x,y} \frac{x - \bar{x}}{\sqrt{V_x}}$$

B. Algunas observaciones

- a) Para $n = 2$, la recta de regresión es la recta que pasa por los puntos (x_1, y_1) y (x_2, y_2) . En este caso, $\hat{b} = (y_2 - y_1)/(x_2 - x_1)$.
- b) Si $\text{cov}_{x,y} = 0$, la recta de regresión es la recta horizontal $y = \bar{y}$.
- c) Si los datos x e y están de partida tipificados (es decir, si $\bar{x} = \bar{y} = 0$ y $V_x = V_y = 1$), entonces la recta de regresión de y sobre x es, simplemente,

$$y = \rho_{x,y} x,$$

que pasa por el origen y cuya pendiente es la correlación ρ .

- d) La recta de regresión de x sobre y ($x = c + dy$) no coincide con la recta de regresión de y sobre x ($y = a + bx$), es decir, no es cierto, en general, que $b = 1/d$ o que $a = -c/d$. A pesar de que se trata de aproximar una recta a una misma nube de puntos, cuando se regresa a y sobre x el error es vertical, $(y_i - (a + bx_i)) = (y_i - a - bx_i)$, mientras que cuando x se regresa sobre y el error que se mide es horizontal, $(x_i - (c + dy_i)) = (x_i - c - dy_i)$.

Como ilustración, si los datos x e y estuvieran tipificados, la recta de regresión de y sobre x sería

$$y = \rho_{x,y} x,$$

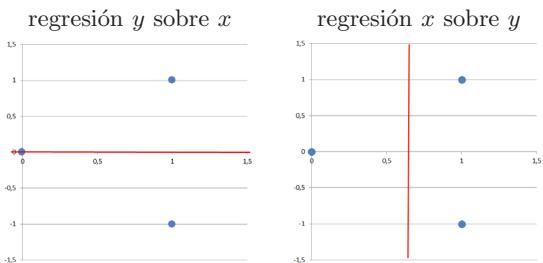
mientras que la de regresión de x sobre y sería

$$x = \rho_{x,y} y.$$

Como ejemplo adicional, para datos

| | | | |
|-----|---|---|----|
| X | 0 | 1 | 1 |
| Y | 0 | 1 | -1 |

tendríamos las rectas de regresión de las figuras de la derecha.



C. Bondad de ajuste

Supongamos que ya hemos calculado la recta de regresión. Podemos obtener el mínimo error cuadrático medio, es decir, el error cuadrático medio que se comete con la recta de regresión, sustituyendo los valores \hat{a} y \hat{b} en $E(\hat{a}, \hat{b})$. Las cantidades

$$E(\hat{a}, \hat{b}), \quad \text{o mejor,} \quad \sqrt{E(\hat{a}, \hat{b})},$$

miden ese error cometido: $E(\hat{a}, \hat{b})$ en términos cuadráticos, y la versión con la raíz, en unidades más convenientes.

Usando que la ecuación de la recta de regresión se escribe en la forma $y - \bar{y} = \hat{b}(x - \bar{x})$, y la fórmula (1.18) para \hat{b} podemos escribir

$$\begin{aligned} E(\hat{a}, \hat{b}) &= \frac{1}{n} \sum_{i=1}^n ((y_i - \bar{y}) - \hat{b}(x_i - \bar{x}))^2 \\ &= V_y - 2\hat{b} \text{cov}_{x,y} + \hat{b}^2 V_x = V_y (1 - \rho_{x,y}^2). \end{aligned}$$

Así que

$$(1.20) \quad \boxed{\text{error de ajuste} = \sqrt{E(\hat{a}, \hat{b})} = \sqrt{V_y} \sqrt{1 - \rho_{x,y}^2}}$$



Nota 1.3.4. Datos tipificados. Traslademos la nube de puntos para que la media sea $(0,0)$ y cambiemos unidades para que las varianzas sean unidad, es decir, hagamos

$$z_i = \frac{x_i - \bar{x}}{\sqrt{V_x}} \quad y \quad w_i = \frac{y_i - \bar{y}}{\sqrt{V_y}}$$

Nótese que $\rho_{x,y} = \rho_{z,w}$. Con esta transformación lineal $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ llevamos la muestra (x, y) en la muestra (z, w) . La recta de regresión de (z, w) pasa por el $(0,0)$ y tiene pendiente $\rho_{z,w} = \rho_{x,y}$. El error de ajuste es $\sqrt{1 - \rho_{x,y}^2}$.

Esta forma de medir el error tiene las unidades de y . Es más natural definir la siguiente cantidad, que no tiene unidades:

$$(1.21) \quad \boxed{\frac{\text{error de ajuste}}{\sqrt{V_y}} = \sqrt{1 - \rho_{x,y}^2}}$$

y que se conoce como la **bondad de ajuste por la recta de regresión**. Un valor de $\rho_{x,y}^2$ próximo a 1 significa buen ajuste.

La cantidad $\rho_{x,y}^2$ es conocida también como **coeficiente de determinación**⁶, y se escribe R^2 . Así que un valor de R^2 próximo a 1 significa buen ajuste. Hay quien sitúa el listón para adjudicar el adjetivo “bueno” al ajuste en 0.8 o 0.9.

(Nótese que la correlación es un número menor, en módulo, que 1, y por tanto elevar al cuadrado decrece su magnitud. Por ejemplo, un coeficiente de correlación (muy) alto como un 90 % se corresponde a $R^2 = 0.81$).

En la práctica, tras un análisis de regresión lineal, se dan

- los coeficientes \hat{a} y \hat{b} de la recta de regresión;
- y el valor de R^2 , que cuantifica la bondad del ajuste de la recta a la nube de puntos.



Nota 1.3.5. Si definimos una nueva variable \hat{Y} como $\hat{Y} = \hat{a} + \hat{b}X$, es decir, con los valores y dados por la recta de regresión, entonces

$$E(\hat{a}, \hat{b}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = V_{y-\hat{y}}.$$

La cantidad

$$V_{y-\hat{y}} = V_y(1 - \rho_{x,y}^2)$$

se conoce como **varianza residual**. Es la varianza de los datos de la muestra de Y cuando se les resta la explicación lineal \hat{Y} en términos de los datos de X , es decir, la varianza “no explicada”:

$$\text{error de ajuste} = \sqrt{\text{varianza residual}}, \quad \text{y} \quad \frac{\text{varianza residual de } y}{\text{varianza de } y} = \frac{V_y(1 - \rho^2)}{V_y} = 1 - \rho^2.$$

EJEMPLO 1.3.1. Muestra

| X | -3 | -1 | 0 | 0 | 1 | 5 |
|-----|----|----|----|---|---|---|
| Y | -2 | 0 | -1 | 1 | 1 | 3 |

Se tiene

$$\bar{x} = 1/3 \quad \bar{y} = 1/3$$

$$V_x = 5.88 \quad V_y = 2.55$$

$$s_x^2 = 7.06 \quad s_y^2 = 3.06$$

Además, $\text{cov}_{x,y} = 3.55$ y $\rho_{x,y} = 0.91$. La recta de regresión tiene $\hat{a} = 0.132$ y $\hat{b} = 0.604$. El coeficiente de determinación es $R^2 = 0.84$, relativamente cercano a 1. ♣

⁶Parece excesivo acuñar un nuevo nombre para una cantidad, el coeficiente de correlación (al cuadrado), ya conocida. Sin embargo, en más de dos dimensiones, el correspondiente coeficiente de determinación R^2 tiene una expresión más complicada. Véase el apartado 1.3.3, y en particular la fórmula (1.30).

D. Extrapolando con la recta de regresión

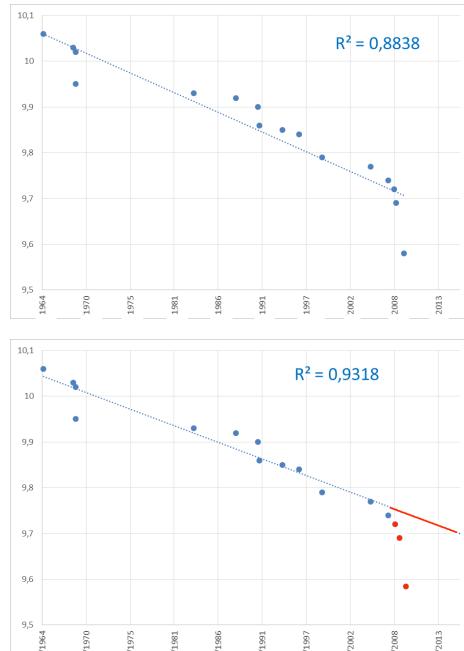
Si el ajuste lineal a los datos de la muestra que nos da la recta de regresión es “bueno”, entendemos que los datos de la muestra siguen la tendencia lineal que marca esa recta. Eso permite extrapolar. Si quisieramos el valor de y que debería corresponder a un valor x^* que no es uno de los x_i , tomariamos el dato dado por la recta de regresión

$$y^* = \hat{a} + \hat{b}x^*.$$

EJEMPLO 1.3.2. Los récords del mundo de los 100 metros lisos.

Los que siguen son los datos de la evolución del record del mundo de los 100 metros lisos. Los tres últimos corresponden a Usain Bolt.

| fecha | tiempo |
|------------|--------|
| 15/10/1964 | 10.06 |
| 20/06/1968 | 10.03 |
| 13/10/1968 | 10.02 |
| 14/10/1968 | 9.95 |
| 03/07/1983 | 9.93 |
| 24/09/1988 | 9.92 |
| 14/06/1991 | 9.90 |
| 25/08/1991 | 9.86 |
| 06/07/1994 | 9.85 |
| 27/07/1996 | 9.84 |
| 16/06/1999 | 9.79 |
| 14/06/2005 | 9.77 |
| 09/09/2007 | 9.74 |
| 31/05/2008 | 9.72 |
| 16/08/2008 | 9.69 |
| 16/08/2009 | 9.58 |



En la primera figura, que contiene todos los datos, se aprecia un ajuste razonablemente bueno; véase el valor de R^2 . Destacan, como valores un tanto atípicos: un dato al principio, que corresponde a los Juegos Olímpicos de México⁷; y los dos últimos puntos, que son los récords de Usain Bolt. Extraiga el lector conclusiones.

En la segunda gráfica hemos quitado los tres últimos récords (época pre-Bolt). La recta de regresión tiene una pendiente algo menos acusada (hacia abajo), y el ajuste es algo mejor (mayor R^2). La extrapolación de esa recta está marcada en rojo. Con esa extrapolación, allá por 2009, el record del mundo debería estar en torno a 9.73 (en lugar del increíble 9.58 del Mundial de Berlín). En ocasiones, extrapolar lleva a conclusiones precipitadas. ♣

⁷En los que, por circunstancias particulares (altitud, etc.), se consiguieron otros resultados extraordinarios, como el record de longitud de Bob Beamon.

E. Transformación de datos: ajustes logarítmico y exponencial

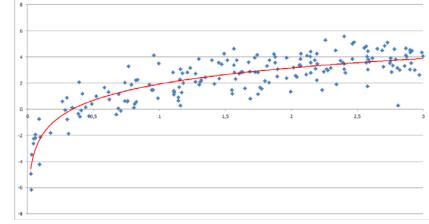
La recta de regresión es el mejor ajuste lineal a los datos de la muestra. Si la bondad de ajuste es buena, los datos de la muestra siguen (bastante aproximadamente) una relación lineal, y podremos usar la recta de regresión para interpolar, extrapolar, ...

Es habitual que la relación aparente entre x e y no sea lineal. Hay dos casos bien frecuentes de relaciones no lineales para los que la recta de regresión sigue siendo útil.

Caso 1. Los datos (x_i, y_i) parecen seguir una **relación logarítmica**, es decir, queremos ajustarles una curva de la familia

$$y = B \ln(x) + A.$$

En ese caso los x_i han de ser $x_i > 0$.



1. Introducimos una nueva variable $Z = \ln(X)$,
2. transformamos los datos de la muestra: definimos $z_i = \ln(x_i)$,
3. los pares (z_i, y_i) son muestra de (Z, Y) ,
4. ajustamos recta de regresión a los (z_i, y_i) ,
5. digamos $y = \hat{a} + \hat{b}z$,
6. el ajuste a los datos originales será

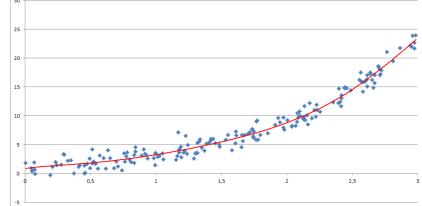
$$y = \hat{b} \ln(x) + \hat{a}$$

$$7. A = \hat{a} \text{ y } B = \hat{b} = \frac{\text{cov}_{\ln(x), y}}{V_{\ln(x)}}.$$

Caso 2. Los datos (x_i, y_i) parecen seguir una **relación exponencial**, es decir, queremos ajustarles una curva de la familia

$$y = Ce^{Dx}.$$

En ese caso los y_i han de ser $y_i > 0$.



1. Introducimos una nueva variable $W = \ln(Y)$; equivalentemente, $Y = e^W$,
2. transformamos los datos de la muestra: definimos $w_i = \ln(y_i)$,
3. los pares (x_i, w_i) son muestra de (X, W) ,
4. ajustamos recta de regresión a los (x_i, w_i) ,
5. digamos $w = \hat{a} + \hat{b}x$,
6. el ajuste a los datos originales será

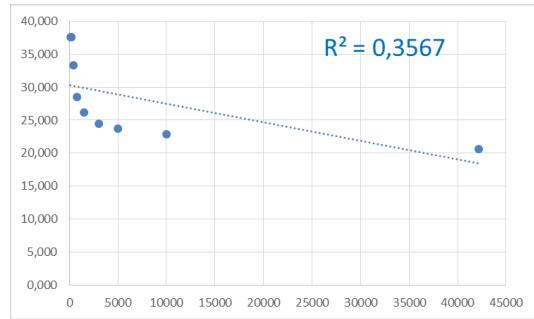
$$y = e^w = e^{\hat{a}} e^{\hat{b}x}.$$

$$7. C = e^{\hat{a}} \text{ y } D = \hat{b} = \frac{\text{cov}_{x, \ln(y)}}{V_x}.$$

EJEMPLO 1.3.3. *Más sobre récords del mundo.*

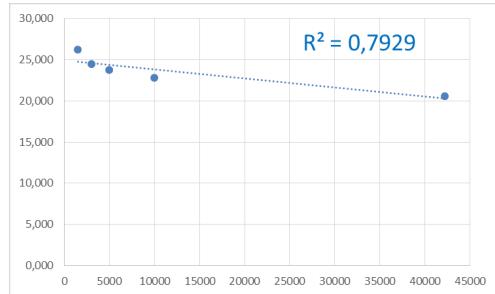
La siguiente tabla contiene las mejores marcas mundiales de las carreras incluidas en el programa olímpico. Se incluye una columna con los tiempos en segundos, y otra con la velocidad en kilómetros por hora.

| distancia | tiempo | velocidad |
|-----------|---------|-----------|
| 100 | 9.58 | 37.578 |
| 200 | 19.19 | 37.520 |
| 400 | 43.18 | 33.349 |
| 800 | 100.91 | 28.540 |
| 1500 | 206 | 26.214 |
| 3000 | 440.67 | 24.508 |
| 5000 | 757.35 | 23.767 |
| 10000 | 1577.53 | 22.820 |
| 42195 | 7377 | 20.591 |

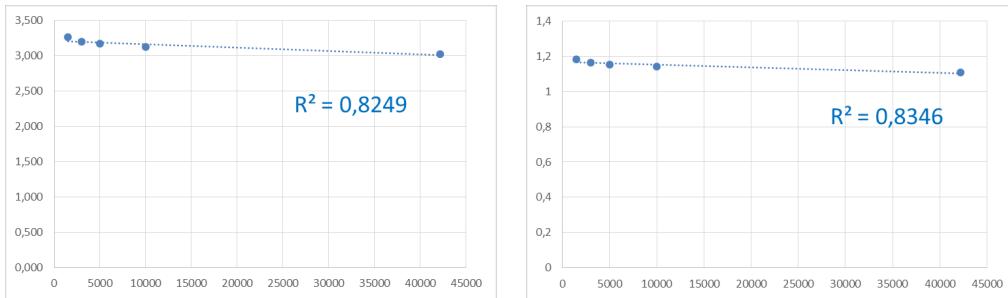


A la derecha representamos la nube de puntos (distancias-velocidades), recta de regresión incluida. Obsérvese, gráficamente o con el valor del R^2 , cómo el ajuste es bastante pobre.

Adelantamos ya que el objetivo en este ejemplo es extrapolar los datos para, por ejemplo, predecir a qué velocidad debería irse en una carrera muy larga, de 100 km o más. Con este planteamiento, conviene no considerar, en la lista anterior, los récords de carreras cortas (que en realidad es otro deporte), y limitarnos a los récords de fondo; digamos del 1500 en adelante. La figura de la derecha recoge el análisis correspondiente; obsérvese que el ajuste es bastante mejor. La extrapolación con esta recta nos daría, por ejemplo, que la velocidad para la prueba de 100 km debería ser 14.02 km/h, lo que se traduciría en un record de 7:08:02 (el record mundial está 6:13:33). Aunque para, por ejemplo, la carrera de 400 km, el modelo predeciría una velocidad ¡negativa!



Buscando mejor ajuste, y a la vista de la forma de la nube de puntos, podemos proceder como en el Caso 2 descrito antes, tomando logaritmos en los datos de velocidad. Sería el caso de la primera figura siguiente. Con este ajuste, que tiene mejor R^2 , se predice un record de los 100 km de 6:28:30, y una velocidad en los 400 km de 3.7 km/h (esto es, de paseo mirando escaparates). Si el lector queda aún insatisfecho, puede repetir el procedimiento, es decir, tomar un logaritmo de los logaritmos anteriores, obteniendo una recta como la de la segunda figura.



Este ajuste adicional predice 6:20:30 como record de los 100 km (lo que está bastante bien), y una velocidad en los 400 km de 5.7 km/h (que, bueno, sigue siendo andar, pero ya a ritmo ligero). ♣

1.3.3. Más de dos variables: regresiones múltiples

Para evitar excesos notacionales, vamos a ilustrar el análisis de datos multidimensionales con el caso de dimensión 3.

Supongamos que la muestra está formada por *tripletes* de números. Digamos que las variables son X , Y y Z , y que la muestra tiene tamaño n :

$$(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n).$$

Suponemos ya realizado el análisis de cada componente de la muestra, es decir, de las muestras marginales (medias, varianzas, etc.). Obsérvese que ahora, como medidas de dependencia lineal, tenemos tres covarianzas: $\text{cov}_{x,y}$, $\text{cov}_{x,z}$ y $\text{cov}_{y,z}$, y tres correspondientes coeficientes de correlación: $\rho_{x,y}$, $\rho_{x,z}$ y $\rho_{y,z}$.

A. Regresión de Z sobre X e Y

El objetivo es “explicar” una de las variables, por ejemplo Z , en términos de las otras dos. Más concretamente, pretendemos determinar (el plano de) la **regresión múltiple** de Z sobre X e Y . Para ello, escribimos la función

$$(1.22) \quad E(a, b, c) = \frac{1}{n} \sum_{j=1}^n (z_j - (a + b x_j + c y_j))^2$$

con el firme propósito de minimizarla. Las variables aquí son a , b y c .

Para los cálculos de esta sección, y por unificar notación, para las varianzas, en lugar de por ejemplo V_x , escribiremos $\text{cov}_{x,x}$.

Para minimizar la función E definida en (1.22), calculamos las tres derivadas

parciales de E e igualamos a 0 para obtener el sistema de ecuaciones siguiente

$$\begin{cases} 0 = \frac{\partial E}{\partial a} = 2(\bar{z} - a - b\bar{x} - c\bar{y}) \\ 0 = \frac{\partial E}{\partial b} = 2(\bar{zx} - a\bar{x} - b\bar{x^2} - c\bar{xy}) \\ 0 = \frac{\partial E}{\partial c} = 2(\bar{zy} - a\bar{y} - b\bar{xy} - c\bar{y^2}) \end{cases}$$

que, escrito en forma matricial, es

$$\begin{pmatrix} \bar{z} \\ \bar{zx} \\ \bar{zy} \end{pmatrix} = \begin{pmatrix} 1 & \bar{x} & \bar{y} \\ \bar{x} & \bar{x^2} & \bar{xy} \\ \bar{y} & \bar{xy} & \bar{y^2} \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Eliminando gaussianamente (restamos de la segunda fila la primera multiplicada por \bar{x} , y luego restamos de la tercera fila la primera multiplicada por \bar{y}), llegamos al sistema equivalente

$$(1.23) \quad \begin{pmatrix} \bar{z} \\ \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} 1 & \bar{x} & \bar{y} \\ 0 & \text{cov}_{x,x} & \text{cov}_{x,y} \\ 0 & \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

en el que, por conveniencia, separamos la primera ecuación de las otras dos: por un lado,

$$(1.24) \quad \bar{z} = a + b\bar{x} + c\bar{y}$$

y por otro

$$(1.25) \quad \begin{pmatrix} \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} \text{cov}_{x,x} & \text{cov}_{x,y} \\ \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} b \\ c \end{pmatrix}$$

Llamemos $(\hat{a}, \hat{b}, \hat{c})$ a la solución del sistema de ecuaciones (1.23). La ecuación matricial (1.25) permite hallar b y c para luego sustituir en (1.24).

El lector interesado puede hallar una expresión explícita para $(\hat{a}, \hat{b}, \hat{c})$, pero no seguimos ese camino. Aunque sí queremos apuntar ya que \hat{b} y \hat{c} no dependen de las medias de las muestras, sino solo de varianzas y covarianzas.

B. Coeficiente R^2 de determinación múltiple

En el punto de coordenadas $(\hat{a}, \hat{b}, \hat{c})$, la función E definida en (1.22) alcanza su mínimo. Llamemos

$$(1.26) \quad \text{EM} = E(\hat{a}, \hat{b}, \hat{c}),$$

donde EM significa “error (cuadrático) mínimo”. Recuérdese que EM sólo depende de los datos z_i , x_i e y_i .

Obsérvese que

$$E(\bar{z}, 0, 0) = \text{cov}_{z,z},$$

de manera que, como $E(\hat{a}, \hat{b}, \hat{c})$ es un mínimo,

$$(1.27) \quad 1 = \frac{E(\bar{z}, 0, 0)}{\text{cov}_{z,z}} \geq \frac{E(\hat{a}, \hat{b}, \hat{c})}{\text{cov}_{z,z}},$$

lo que permite escribir

$$(1.28) \quad \boxed{\frac{E(\hat{a}, \hat{b}, \hat{c})}{\text{cov}_{z,z}} = 1 - R^2}$$

Con la notación R^2 recogemos ya que R^2 es una cantidad positiva (por la condición (1.27)) que, además, es a lo sumo 1 (pues E es siempre positiva). Es decir,

$$0 \leq R^2 \leq 1.$$

A esta cantidad, R^2 , nos referiremos como el **coeficiente de determinación múltiple**. Como en el caso de dos dimensiones, R^2 cercano a 1 querrá decir “buen ajuste”, y R^2 cercano a 0 se corresponderá con un “mal ajuste”.

Por consistencia con el caso de dos dimensiones, la bondad del ajuste se mide con la cantidad $\sqrt{1 - R^2}$.

¿Cómo se calcula R^2 ? El procedimiento visto hasta aquí exige:

- partiendo de las muestras de Z , X e Y , hallar las medias, varianzas y covarianzas;
- luego resolver el sistema (1.23) para obtener los valores $(\hat{a}, \hat{b}, \hat{c})$;
- insertarlos en la expresión (1.22) para el error E para obtener el EM;
- que, dividido por la varianza de la muestra de Z (y restado de 1) nos da finalmente el R^2 .

Expresión explícita de R^2 . Desarrollamos seguidamente la expresión explícita de R^2 que se recoge más adelante en la fórmula (1.31).

Lector, si no tiene otra cosa que hacer, puede aplicarse a la tarea de obtener expresiones explícitas de $\hat{a}, \hat{b}, \hat{c}$, para luego llevarlos a (1.22), ponerse a simplificar.... Pero es mucho más instructivo argumentar como sigue.

Observamos primero que el error mínimo EM no depende de \hat{a} . Veamos. Como el triple $(\hat{a}, \hat{b}, \hat{c})$ es solución de (1.23), en particular se tiene, por la primera ecuación (1.24), que

$$\bar{z} = \hat{a} + \hat{b} \bar{x} + \hat{c} \bar{y}.$$

Llevando esta identidad en (1.22) observamos que

$$E(\hat{a}, \hat{b}, \hat{c}) = \frac{1}{n} \sum_{j=1}^n [(z_j - \bar{z}) - (\hat{b}(x_j - \bar{x}) + \hat{c}(y_j - \bar{y}))]^2,$$

donde ya no aparece \hat{a} .

Desarrollando el cuadrado,

$$\begin{aligned} E(\hat{a}, \hat{b}, \hat{c}) &= \text{cov}_{z,z} + \hat{b}^2 \text{cov}_{x,x} + \hat{c}^2 \text{cov}_{y,y} + 2\hat{b}\hat{c}\text{cov}_{x,y} - 2\hat{b}\hat{c}\text{cov}_{x,z} - 2\hat{c}\text{cov}_{y,z} \\ &= \text{cov}_{z,z} + (\hat{b}, \hat{c}) \begin{pmatrix} \text{cov}_{x,x} & \text{cov}_{x,y} \\ \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix} - 2(\hat{b}, \hat{c}) \begin{pmatrix} \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix} \end{aligned}$$

Ahora, usando las ecuaciones de (1.25), concluimos que

$$(1.29) \quad \text{EM} = E(\hat{a}, \hat{b}, \hat{c}) = \text{cov}_{z,z} - (\hat{b}, \hat{c}) \begin{pmatrix} \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix},$$

una expresión bastante más sencilla. Recuérdese que la cantidad de interés (para el cálculo de R^2) es $\text{EM}/\text{cov}_{z,z}$. Los cálculos se simplifican más si observamos que:

Lema 1.1 *La expresión $\text{EM}/\text{cov}_{z,z}$ es invariante bajo cambios de escala en las muestras de Z , X e Y .*

Y como ya hemos comprobado que $\text{EM}/\text{cov}_{z,z}$ tampoco depende de traslaciones de las muestras, podremos proseguir el argumento suponiendo que las series de partida están *tipificadas*. Pero antes:

DEMOSTRACIÓN DEL LEMA 1.1. Si partimos de las muestras de Z , X e Y , los valores \hat{b} y \hat{c} que minimizan la función de error verifican, por (1.25),

$$\begin{pmatrix} \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} \text{cov}_{x,x} & \text{cov}_{x,y} \\ \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix}$$

y el error mínimo es

$$\text{EM} = \text{cov}_{z,z} - (\hat{b}, \hat{c}) \begin{pmatrix} \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix}.$$

Digamos ahora que las muestras son Z , $X' = \lambda X$ e Y , donde $\lambda > 0$. Llamemos \hat{b}' y \hat{c}' a los valores que minimizan la función de error en este caso. Estos valores verifican, por (1.25),

$$\begin{aligned} \begin{pmatrix} \text{cov}_{x',z} \\ \text{cov}_{y,z} \end{pmatrix} &= \begin{pmatrix} \text{cov}_{x',x'} & \text{cov}_{x',y} \\ \text{cov}_{x',y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix} \\ \iff \begin{pmatrix} \lambda \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix} &= \begin{pmatrix} \lambda^2 \text{cov}_{x,x} & \lambda \text{cov}_{x,y} \\ \lambda \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix}, \end{aligned}$$

usando las propiedades de varianzas y covarianzas, lo que nos dice finalmente que podemos escribir \hat{b}' y \hat{c}' en términos de los \hat{b} y \hat{c} originales como

$$\begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix} = \begin{pmatrix} \hat{b}/\lambda \\ \hat{c} \end{pmatrix}$$

El error mínimo sería

$$\text{EM}' = \text{cov}_{z,z} - (\hat{b}', \hat{c}') \begin{pmatrix} \text{cov}_{x',z} \\ \text{cov}_{y,z} \end{pmatrix} = \text{cov}_{z,z} - (\hat{b}/\lambda, \hat{c}) \begin{pmatrix} \lambda \text{cov}_{x,z} \\ \text{cov}_{y,z} \end{pmatrix},$$

que coincide, tras cancelar un factor λ , con el EM original. Así que EM no varía bajo cambios de escala de X , y tampoco $\text{EM}/\text{cov}_{z,z}$, claro.

Exactamente el mismo argumento se aplica para cambios de escala en Y .

Veamos, finalmente, qué ocurre cuando cambiamos la escala de Z . Digamos que las muestras son de $Z' = \lambda Z$, X e Y . Volvemos a llamar \hat{b}' y \hat{c}' a los valores que minimizan la función de error en este caso, que verifican

$$\begin{aligned} \begin{pmatrix} \text{cov}_{x,z'} \\ \text{cov}_{y,z'} \end{pmatrix} &= \begin{pmatrix} \text{cov}_{x,x} & \text{cov}_{x,y} \\ \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix} \\ \iff \begin{pmatrix} \lambda \text{cov}_{x,z} \\ \lambda \text{cov}_{y,z} \end{pmatrix} &= \begin{pmatrix} \text{cov}_{x,x} & \text{cov}_{x,y} \\ \text{cov}_{x,y} & \text{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix}, \end{aligned}$$

Ahora tenemos que

$$\begin{pmatrix} \hat{b}' \\ \hat{c}' \end{pmatrix} = \begin{pmatrix} \lambda \hat{b} \\ \lambda \hat{c} \end{pmatrix},$$

y el error mínimo sería

$$\text{EM}' = \text{cov}_{z',z'} - (\hat{b}', \hat{c}') \begin{pmatrix} \text{cov}_{x,z'} \\ \text{cov}_{y,z'} \end{pmatrix} = \lambda^2 \text{cov}_{z,z} - (\lambda \hat{b}, \lambda \hat{c}) \begin{pmatrix} \lambda \text{cov}_{x,z} \\ \lambda \text{cov}_{y,z} \end{pmatrix} = \lambda^2 \text{EM}.$$

El error mínimo cambia en esta ocasión, pero

$$\frac{\text{EM}'}{\text{cov}_{z',z'}} = \frac{\lambda^2 \text{EM}}{\lambda^2 \text{cov}_{z,z}} = \frac{\text{EM}}{\text{cov}_{z,z}},$$

lo que concluye la demostración. ■

A la vista del lema anterior, y para obtener una fórmula cerrada para R^2 , podemos suponer que las muestras de Z , X e Y están ya tipificadas, de manera que las medias son 0, las varianzas valen 1, y las covarianzas son, simplemente, correlaciones. Para este caso, el valor del coeficiente de determinación es

$$R^2 = (\hat{b}, \hat{c}) \begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix}$$

(véanse (1.28) y (1.29)), donde \hat{b} y \hat{c} verifican, siguiendo (1.25),

$$\begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix},$$

o lo que es lo mismo, usando que la matriz anterior es simétrica (y trasponerla no causa efecto alguno),

$$(\hat{b}, \hat{c}) = (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}^{-1} = \frac{1}{1 - \rho_{x,y}^2} (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & -\rho_{x,y} \\ -\rho_{x,y} & 1 \end{pmatrix}.$$

Así que, finalmente,

$$(1.30) \quad R^2 = \frac{1}{1 - \rho_{x,y}^2} (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & -\rho_{x,y} \\ -\rho_{x,y} & 1 \end{pmatrix} \begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix}$$

que también se puede escribir como

$$R^2 = \frac{1}{1 - \rho_{x,y}^2} [\rho_{x,z}^2 - 2 \rho_{x,y} \rho_{y,z} \rho_{x,z} + \rho_{y,z}^2].$$

Aunque el lector haría bien en recordar la expresión (1.30), o mejor,

$$(1.31) \quad R^2 = (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix}$$

(con la inversa de la matriz de correlaciones entre X e Y) si es que quiere extender este análisis al caso de la regresión de Z sobre unas variables X_1, \dots, X_k , con $k \geq 2$.

1.4. Comandos de Excel

1.4.1. Una variable

Suponemos que los datos x_1, \dots, x_n están recogidos en un rango de celdas al que llamaremos **rango**.

- Tamaño de la muestra: `=contar(rango)`.
- Media muestral: `=promedio(rango)`.
- Mediana: `=mediana(rango)`.
- Moda: `=moda(rango)`.
- Cuasidesviación típica muestral: `=desvest(rango)`.
- Cuasivarianza muestral: `=var(rango)` (o elevar al cuadrado `desvest`).
- Varianza y desviación típica muestrales: `=varp(rango)` y `=desvestp(rango)`.
- Máximo y mínimo: `=max(rango)`, `=min(rango)`.
- Cuartiles: `=cuartil(rango;n)`, donde $n = 0, 1, 2, 3, 4, 5$. Si $n = 0$, da el mínimo; si $n = 5$, el máximo, y $n = 1, 2, 3$ da el primer, segundo (mediana) y tercer cuartiles.
- Coeficiente de asimetría: la función `=coeficiente.asimetria(rango)` calcula

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{s_x^3},$$

donde s_x es la cuasidesviación típica. Obsérvese que no coincide exactamente con la definición de (1.7).

Si los datos vienen agrupados en clases C_1, \dots, C_k , con marcas de clase y_1, \dots, y_k en **rangovalores** y frecuencias relativas f_1, \dots, f_k en **rangofrecuencias**, entonces la fórmula (1.2) para la media muestral se escribe

`=sumaproducto(rangovalores;rangofrecuencias)`.

Para la confección de histogramas, puede ser útil la instrucción `contar.si`:

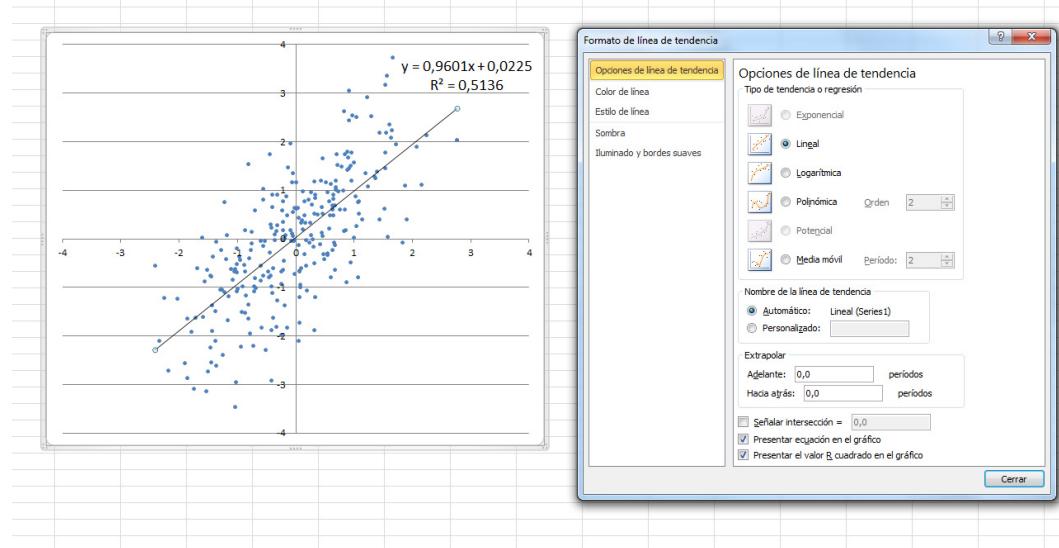
- con una instrucción del tipo `=contar.si(rango;F4)` contamos cuántas celdas del `rango` contienen el valor que aparece en la celda F4;
- con una instrucción del tipo `=contar.si(rango;"<="&F4)` contamos cuántas celdas del `rango` contienen valores menores o iguales que el valor que aparece en la celda F4.

1.4.2. Dos variables

Suponemos que los datos x_1, \dots, x_n están recogidos en el rango de celdas al que llamaremos `rango1`, y los datos y_1, \dots, y_n en `rango2`.

- Covarianza muestral: `=covar(rango1;rango2)`.
- Coeficiente de correlación muestral: `=coef.de.correl(rango1;rango2)`.

Recta de regresión. Dibujamos nubes de puntos con gráficos de dispersión. Luego, pinchando con el botón derecho del ratón sobre el gráfico, “agregar linea de tendencia” permite dibujar la recta de regresión, y da la opción de señalar la ecuación del recta y el valor de R^2 .



Por supuesto, estos valores se pueden calcular también con las fórmulas correspondientes (1.15), o utilizando las funciones

- `=interseccion.eje(rango2;rango1)` para \hat{a} ,
- `=pendiente(rango2;rango1)` para \hat{b} ,
- `=coeficiente.R2(rango2;rango1)` para el coeficiente de determinación R^2 .

Nótese el orden en que se escriben los rangos en las dos primeras (en la tercera es irrelevante).