Capítulo 1

Estadística descriptiva básica

1.1. Representación de datos 18
1.2. Medidas descriptivas. Resúmenes cuantitativos 21
1.2.1. Medidas de centralización 21
1.2.2. Medidas de dispersión (respecto de la media) 23
1.2.3. Medidas de asimetría (respecto de la media) 28
1.3. Datos multidimensionales. Regresión lineal 29
1.3.1. Medidas de dependencia (lineal) 29
1.3.2. Ajuste lineal: recta de regresión 33
1.3.3. Más de dos variables: regresiones múltiples 42
1.4. Comandos de Excel 48
1.4.1. Una variable 48
1.4.2. Dos variables 49

La estadística descriptiva que en este capítulo se presenta es material básico, que el alumno, en gran medida, ya conoce al menos desde el bachillerato[1].

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 cálculo y uso de la *recta de regresión* para datos bidimensionales (o más generalmente, para datos multidimensionales). Finalmente, el apartado 1.4 contiene algunas instrucciones útiles de Excel.

[1] Como quiera que se denomine ahora.

1.1. Representación de datos

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.

☞ **Nota** 1.1.1. Por *población* entendemos el universo de objetos donde estudiamos la variable. Podrían ser las personas que viven en una determinada región, los insectos de cierta especie, todos los posibles resultados de una cierta estrategia de apuestas en la ruleta de un casino, o incluso una noción más abstracta como todos los posibles valores que puede tomar una variable aleatoria. *Población* (DRAE, quinta acepción): conjunto de los elementos sometidos a una evaluación estadística mediante muestreo.

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, ...

Usaremos *sistemáticamente* la siguiente notación:

$$x_1, x_2, \ldots, x_n$$

para denotar los datos de una muestra de la variable X. 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

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

para indicar que se ha utilizado una permutación de $\{1, \ldots, n\}$ 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 $x_1, \ldots, x_n$. 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 $C_1, \ldots, C_k$, donde k es el número de clases consideradas. Cada uno de los datos de la muestra ha 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,

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

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

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

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

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

- 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

$$y_1, y_2, \ldots, y_k$$

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

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

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 $C_1, \ldots, C_k$,
- las marcas de clase $y_1, \ldots, y_k$,
- y las frecuencias absolutas $n_1, \ldots, n_k$ (y por ende[2], el tamaño n de la muestra y las frecuencias relativas $f_1, \ldots, f_k$);

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

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).

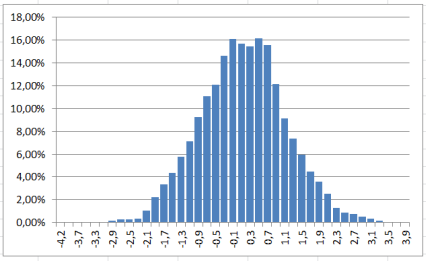
Si los datos no vienen agrupados, conviene agruparlos en clases, tomando por ejemplo un cierto ancho de clase común. Para que el histograma resultante sea informativo, el ancho no debe ser muy pequeño (habría muchas clases, pero probablemente contendrían muy pocos elementos) ni demasiado grande (quedarían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

	datos x
1	-1,15947
2	0,35253
3	0,45858
4	0,58237
5	-2,61563
6	0,41927
7	0,90431
8	-0,11321
9	-0,34396
10	-2,81868
11	0,46550
12	0,26764
13	-0,25731
14	-0,99053
15	-0,01991
16	2,40061
17	-0,13079
18	1,09168
19	-1,22362
20	0,04922
21	-0,54684
22	0,60299
23	-1,54120

número de datos 1334

Histograma

marcas	clases	frecuencias absolutas	frecuencias relativas
-4,2	< -4	0	0,00%
-3,9	-3,8	0	0,00%
-3,7	-3,6	0	0,00%
-3,5	-3,4	0	0,00%
-3,3	-3,2	0	0,00%
-3,1	-3,0	0	0,00%
-2,9	-2,8	2	0,15%
-2,7	-2,6	4	0,30%
-2,5	-2,4	4	0,30%
-2,3	-2,2	5	0,37%
-2,1	-2,0	14	1,05%
-1,9	-1,8	30	2,25%
-1,7	-1,6	45	3,37%
-1,5	-1,4	58	4,35%
-1,3	-1,2	77	5,77%
-1,1	-1,0	95	7,12%
-0,9	-0,8	123	9,22%
-0,7	-0,6	148	11,09%
-0,5	-0,4	161	12,07%
-0,3	-0,2	195	14,62%



☞ **Nota** 1.1.2. 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 $j = 1, \ldots, k$, se trata sólo de un cambio de escala en la figura.

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

☞ **Nota** 1.1.3. **Suavizado de histogramas**. Por razones varias (diluir papel del azar, o de posibles errores de medida), 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 $j \le j_1$ y para $j \ge j_2$ 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*

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

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 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 medidas cuantitativas más habituales que usaremos 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

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

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

$$\overline{y} = \sum_{j=1}^{k} f_j\, y_j\,. \tag{1.2}$$

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, para cada j, la marca y_j de la clase C_j es justamente la media de los datos de la clase C_j, es decir, si

$$y_j = \frac{1}{n_j}\sum_{x_i\in C_j} x_i, \qquad \text{para cada } j = 1, \ldots, k,$$

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 se cumple que

$$\min_i x_i \leq \overline{x} \leq \max_i x_i. \tag{1.3}$$

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

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\,\overline{x}. \tag{1.4}$$

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

B. Mediana

Ordenamos los datos de menor a mayor:

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

(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.

En general, no hay tal valor entre entre los datos de la muestra. Sí lo hay si por ejemplo la muestra es $(1, 2, 3, 4, 5)$ (la mediana sería el 3), pero no, por ejemplo, para el caso de la muestra $(1, 2, 3, 4)$. Si permitimos que la mediana no sea un valor de la muestra, entonces hay infinitos candidatos a ser mediana: en la muestra $(1, 2, 3, 4)$, valdría cualquier número estrictamente entre 2 y 3.

Veamos una manera (razonable y convencional) de calcular la mediana, que denotaremos por MED, de una muestra de tamaño n.

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)

Las medidas que presentamos a continuación responden, desde distintos puntos de vista, a la pregunta: ¿cuán dispersos o cuán variables son los datos?

Digamos, por ejemplo, que medimos una cierta característica con dos instrumentos distintos. Con el instrumento A las medidas son $68 = a_1 \leq \cdots \leq a_{100} = 72$, con media $\bar{a} = 70$; con el instrumento B, las medidas son $5 = b_1 \leq \cdots \leq b_{75} = 220$, con una media que es $\bar{b} = 70$ de nuevo. 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 $x_1, \ldots, x_n$. La cantidad $x_i - \overline{x}$ mide la dispersión del dato x_i respecto de $\overline{x}$. La dispersión total sería

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

pero esto es

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \Big(\sum_{i=1}^{n} x_i\Big) - n\overline{x} = n\overline{x} - n\overline{x} = 0\,,$$

que no es informativa: los términos positivos y negativos se compensan exactamente.

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

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

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 se prefiere usar cuadrados, como se muestra a continuación.

A. Varianzas y desviaciones típicas

A1. Varianza y cuasivarianza. Se define la **varianza** de una muestra $x_1, \ldots, x_n$ como

$$\boxed{V_x = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2} \tag{1.5}$$

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

Este valor V_x es un promedio de las dispersiones (de los datos) con respecto a la media, medidas en términos cuadráticos.

☞ **Nota** 1.2.3. Para datos agrupados en k clases (con y_j = la marca de la clase C_j), tras calcular la media $\overline{y}$ según (1.2), definiríamos

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

Si en lugar de dividir por n, dividimos por $n-1$, entonces se dice **cuasivarianza de la muestra**:

$$\boxed{s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2} \tag{1.6}$$

En la notación s_x^2 hacemos de nuevo referencia a la muestra x. 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 relación entre estas dos cantidades es directa:

$$(n-1)\, s_x^2 = nV_x.$$

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

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

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 $x_i = \overline{x}$, para $1 \le i \le n$). Así que $s_x^2 = 0$ o $V_x = 0$ significa dispersión 0.

☞ **Nota** 1.2.4. Detalle de la observación anterior: como $nV_x = \sum_{j=1}^{n}(x_j - \overline{x})^2$ es una suma de términos no negativos, si $V_x = 0$, entonces $(x_j - \overline{x})^2 = 0$ para cada $j = 1, \ldots, n$, y por tanto $x_j = \overline{x}$, para $j = 1, \ldots, n$. Recíprocamente, si los x_j son todos iguales, digamos $x_j = m$, para $j = 1, \ldots, n$, entonces $\overline{x} = m$ y $(x_j - \overline{x})^2 = 0$, para $j = 1, \ldots, n$, por lo que $V_x = 0$.

A2. Cálculo de varianzas y cuasivarianzas. Las expresiones de V_x y s_x^2 que recogemos en (1.7) suelen ser la forma más expeditiva de calcular estas cantidades. Si desarrollamos el cuadrado en la suma de la expresión (1.5), se obtiene que

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j^2 + \overline{x}^2 - 2x_j\,\overline{x}) = \underbrace{\frac{1}{n}\sum_{j=1}^{n} x_j^2}_{:=\overline{x^2}} + \overline{x}^2 - 2\,\overline{x}\,\frac{1}{n}\sum_{j=1}^{n} x_j$$

$$= \overline{x^2} + \overline{x}^2 - 2\,\overline{x}^2 = (\overline{x^2} - \overline{x}^2),$$

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

$$V_x = \overline{x^2} - \overline{x}^2 \qquad \text{y} \qquad s_x^2 = \frac{n}{n-1}\, V_x. \tag{1.7}$$

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.*

A3. Desviación típica y cuasidesviacion típica. Las unidades de la cuasivarianza son el cuadrado de las de los datos (si por ejemplo los datos se escriben en metros, la cuasivarianza tiene unidades de metros cuadrados). Para regresar a las unidades originales, se definen:

(1.8)

desviación típica	**cuasidesviación típica**
$\sqrt{V_x} = \sqrt{\dfrac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2}$	$s_x = \sqrt{\dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2}$

☞ **Nota** 1.2.5. 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).

A4. Cambios de escala. Tenemos una muestra $x_1, \ldots, x_n$, a la que nos referimos abreviadamente como x. Definimos ahora la muestra z, que es la que se obtiene a partir de la muestra x con el cambio lineal

$$x_i \mapsto z_i = a + bx_i, \qquad \text{para cada } i = 1, \ldots, n$$

(aunque las fórmulas que siguen con válidas para cualesquiera números reales a y b, supondremos siempre que $b \neq 0$). Entonces se tiene que

$$\overline{z} = a + b\,\overline{x}, \qquad V_z = b^2\,V_x \qquad \text{y} \qquad s_z^2 = b^2 s_x^2\,.$$

Veamos. La primera expresión para la media ya se ha visto, en (1.4). Para la segunda,

$$V_z = \frac{1}{n}\sum_{j=1}^{n}(z_i - \overline{z})^2 = \frac{1}{n}\sum_{j=1}^{n}\big(a + bx_j - (a + b\overline{x})\big)^2 = b^2\,\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = b^2\,V_x\,.$$

A5. Tipificación. Un cambio lineal $x_i \mapsto a + bx_i$ no altera la naturaleza de los datos (es un cambio de origen y de escala/unidades).

El cambio

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

transforma los datos $x_1, \ldots, x_n$ iniciales en datos $z_1, \ldots, z_n$ sin unidades (y por tanto, comparables con los de otras muestras).

El paso de los datos x_i a los datos z_i se conoce como **tipificación** de la muestra. Esta tipificación requiere que $V_x \neq 0$. Los datos z_i se dicen **tipificados** (de los x_i).

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

A6. Media, varianza y minimización. Tenemos una muestra $x_1, \ldots, x_n$. 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 $a = \overline{x}$.

Veamos. Reescribimos

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

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

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

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

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

Por cierto, el *valor* mínimo[3] de la función $d(a)$, es decir, el número $d(\overline{x})$, 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. Recuerde el lector que la mediana dividía la muestra en "mitades".

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 toman $Q_0 = \text{mín}\{x_1, \ldots, x_n\} = x_1$ y $Q_4 = \text{máx}\{x_1, \ldots, x_n\} = x_n$; el mínimo y el máximo de la muestra. A la diferencia $Q_4 - Q_0$ se le conoce como **rango** de la muestra; es también 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).

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

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, y al tiempo una convención estándar, para determinar Q_1 y Q_3 que cumplan (aproximadamente) esta especificación consiste en tomar:

- si n es par, $n = 2r$,

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

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

- Y si n es impar, $n = 2r + 1$,

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

  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

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

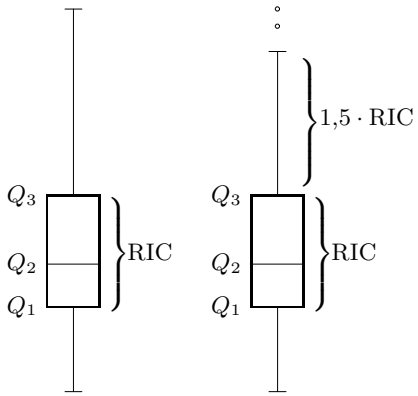
Es una medida más de dispersión, y mide la longitud del rango donde se encuentra el 50 % central de la muestra.

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 la nota 1.2.8).

☞ **Nota** 1.2.6. Una representación gráfica habitual, conocida como **diagramas de cajas (*boxplots*)**, sirve fundamentalmente para comparar unas muestras con otras.

Se marca en vertical, primero, una caja que va desde altura Q_1 hasta altura Q_3, incluyendo una raya horizontal a la altura de la mediana Q_2. Luego, se trazan líneas verticales (conocidas a veces como *bigotes*[4]), una por arriba y otra por debajo, cuyos extremos se pueden determinar 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 $Q_3 + 1.5\,\text{RIC}$),
- y la inferior hasta el menor dato no atípico (menor dato superior a $Q_1 - 1.5\,\text{RIC}$),
- para después señalar los datos atípicos superiores e inferiores. Véase la segunda figura, con símbolos "$\circ$"para estos datos atípicos.

[4] No, 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 $(x_i - \overline{x})$; 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

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

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

$$\boxed{\mathrm{asim}_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{V_x^{3/2}}} \tag{1.9}$$

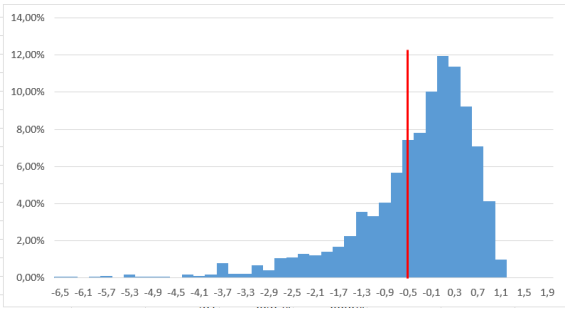
☞ **Nota** 1.2.7. La expresión anterior se puede escribir también como

$$\mathrm{asim}_x = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{x_i - \overline{x}}{\sqrt{V_x}}\Big)^3,$$

con la que se hace evidente que la asimetría depende únicamente de los datos *tipificados*.

Hay quien, por cierto, define el coeficiente de asimetría poniendo s_x^3 en el denominador de (1.9) en lugar de $V_x^{3/2}$.

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



Sean a y b dos constantes, con $b \neq 0$, y consideremos la habitual transformación lineal $x_i \mapsto z_i = a + bx_i$. La relación entre los coeficientes de asimetría de las muestras x y z viene dado por

$$\mathrm{asim}_z = \mathrm{signo}(b)\,\mathrm{asim}_x\,,$$

[5] En inglés, *skewness*.

donde signo(b) es 1 si b es positivo, y -1 si b es negativo. Es decir, el coeficiente de asimetría no depende de traslaciones (dadas por el valor de a); y para los cambios de escala (definidos por b), sólo es relevante el signo (y no su magnitud).

Si los datos x_i están tipificados, se tiene que $\text{asim}_x = \frac{1}{n}\sum_{i=1}^n x_i^3$.

☞ **Nota** 1.2.8. Hay quien define la asimetría dividiendo por s_x^3. Otras medidas posibles de asimetría, también adimensionales, son

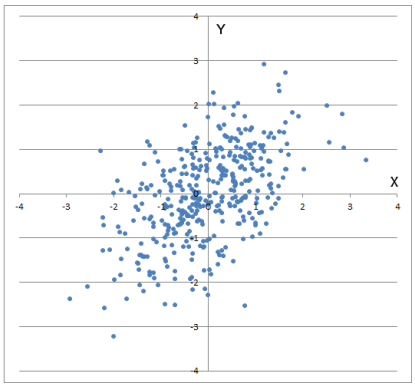
$$\frac{\bar{x} - \text{MED}_x}{\sqrt{V_x}} \qquad \text{y} \qquad \frac{Q_3 + Q_1 - 2\,\text{MED}_x}{Q_3 - Q_1}.$$

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 una 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 español 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:

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

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 $(\overline{x}, \overline{y})$.

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

$$x_1, x_2, \ldots, x_n \text{ por un lado,} \qquad \text{e} \qquad y_1, y_2, \ldots, y_n \text{ por otro.}$$

1.3.1. Medidas de dependencia (lineal)

Las medidas fundamentales de dependencia lineal son la *covarianza*, y su versión adimensional, el *coeficiente de correlación*.

A. Covarianza

La covarianza de la muestra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ se define como sigue:

$$\boxed{\text{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})} \tag{1.10}$$

Los subíndices en el símbolo $\text{cov}_{x,y}$ 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 $\overline{x}$– se corresponden con valores grandes (pequeños) de y_i, por encima (debajo) de su media $\overline{y}$, entonces, en general,

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

y, claro,

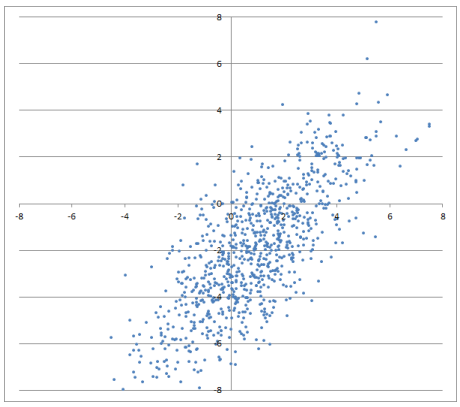
$$(x_i - \overline{x}) < 0 \text{ se corresponde con } (y_i - \overline{y}) < 0.$$

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

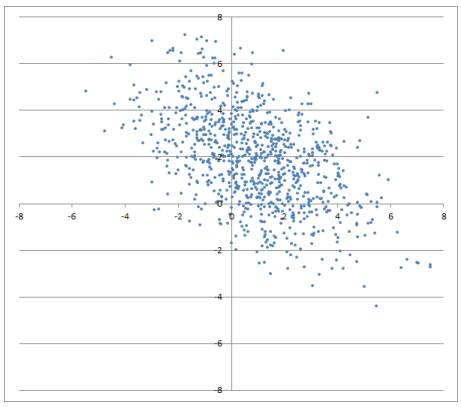
Por el contrario si valores de x_i por encima de $\overline{x}$ se corresponden con valores y_i por debajo de su media $\overline{y}$ entonces se tendrá $\text{cov}_{x,y} < 0$

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).

Datos con covarianza positiva



Datos con covarianza negativa



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 $\overline{y} = a + b\,\overline{x}$ y

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

A.2. Propiedades de la covarianza. Siguen a continuación algunas propiedades de la covarianza, cuya comprobación es casi inmediata a partir de la definición (1.10).

- (Simetría) $\mathrm{cov}_{x,y} = \mathrm{cov}_{y,x}$.
- $\mathrm{cov}_{x,x} = V_x$, mientras que $\mathrm{cov}_{x,-x} = -V_x$.
- (Traslaciones y cambios de escala) Para constantes a, b, c, d, si cambiamos de la muestra (x, y) original a la muestra $(a+bx, c+dy)$, se tiene que $\mathrm{cov}_{a+bx,c+dy} = bd\,\mathrm{cov}_{x,y}$. Es decir, la covarianza no cambia bajo traslaciones (dadas por a y c), pero sí bajo cambios de escala en los datos.
- (Fórmula alternativa para la covarianza) Desarrollando el producto que conforma cada sumando en la definición (1.10), se obtiene la siguiente fórmula alternativa para la covarianza:

  (1.11) $$\boxed{\mathrm{cov}_{x,y} = \overline{xy} - \overline{x}\;\overline{y}}$$

  Aquí, $\overline{xy}$ es la media de la muestra (unidimensional) formada por los datos (producto) $x_i y_i$, para $i = 1, \ldots, n$. Esto es, $\overline{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i$.

Señalamos, por último, que

(1.12) $$|\mathrm{cov}_{x,y}| \le \sqrt{V_x}\sqrt{V_y}.$$

Esta propiedad, extremadamente relevante, como veremos en breve, es una consecuencia directa del siguiente resultado:

Lema 1.1 (Desigualdad de Cauchy–Schwarz en $\mathbb{R}^n$) *Consideramos dos vectores $(\alpha_1, \ldots, \alpha_n)$ y $(\beta_1, \ldots, \beta_n)$. Entonces se tiene que*

$$\Big|\sum_{i=1}^n \alpha_i\,\beta_i\Big|^2 \le \Big(\sum_{i=1}^n \alpha_i^2\Big)\,\Big(\sum_{i=1}^n \beta_i^2\Big),$$

Si los dos vectores son no nulos, la igualdad se tiene si y sólo si $\alpha_i = \lambda\,\beta_i$ para cierto $\lambda \in \mathbb{R}$ y todo $i = 1, \ldots, n$ (es decir, si un vector es múltiplo del otro).

La comprobación de (1.12) usando la desigualdad de Cauchy–Schwarz es directa:

$$|\mathrm{cov}_{x,y}|^2 = \Big|\sum_{j=1}^n \frac{(x_i - \overline{x})}{\sqrt{n}}\,\frac{(y_i - \overline{y})}{\sqrt{n}}\Big|^2 \le \Big(\sum_{i=1}^n \frac{(x_i-\overline{x})^2}{n}\Big)\Big(\sum_{i=1}^n \frac{(y_i-\overline{y})^2}{n}\Big) = V_x \cdot V_y,$$

donde, en la notación del lema 1.1, tomamos $\alpha_i = (x_i - \overline{x})/\sqrt{n}$ y $\beta_i = (y_i - \overline{y})/\sqrt{n}$.

Supongamos que ni los datos x_i ni los datos y_i son constantes; es decir, que $V_x \neq 0$ y $V_y \neq 0$. Entonces, usando el caso de la igualdad del lema 1.1, concluimos que

$$|\mathrm{cov}_{x,y}| = \sqrt{V_x}\sqrt{V_y}$$

si y sólo si el vector de coordenadas $(x_i - \overline{x})$ es un múltiplo del vector de coordenadas $(y_i - \overline{y})$. 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 $1 \leq i \leq n$. En otras palabras, si los datos (x_j, y_j) están *todos sobre una misma recta*.

B. Coeficiente de correlación

Como se ha mencionado antes, la covarianza tiene unidades. Por esa razón, en muchas ocasiones se utiliza la versión adimensional conocida como el **coeficiente de correlación** $\rho_{x,y}$:

$$\boxed{\rho_{x,y} = \frac{\mathrm{cov}_{x,y}}{\sqrt{V_x}\sqrt{V_y}}} \tag{1.13}$$

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.

El coeficiente de correlación *no tiene unidades*. Y su escala es especialmente adecuada, como estará advirtiendo ya el lector, tras la inspección de la desigualdad (1.12).

En paralelo a las propiedades exhibidas antes de la covarianza, listamos a continuación las análogas para el coeficiente de correlación:

- (Simetría) $\rho_{x,y} = \rho_{y,x}$.
- $\rho_{x,x} = 1$ y $\rho_{x,-x} = -1$.
- (Traslaciones y cambios de escala) $\rho_{a+bx,c+dy} = \rho_{x,y}$. 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 $\mathrm{cov}_{x,y}$.

Por la desigualdad (1.12) (consecuencia a su vez de la de Cauchy–Schwarz), se tiene siempre que

$$-1 \leq \rho_{x,y} \leq +1\,. \tag{1.14}$$

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

Por último, obsérvese 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).

Es decir, $\rho_{x,y} = \pm 1$ significa que los datos (x_i, y_i) *están sobre una recta*.

1.3.2. Ajuste lineal: recta de regresión

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

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

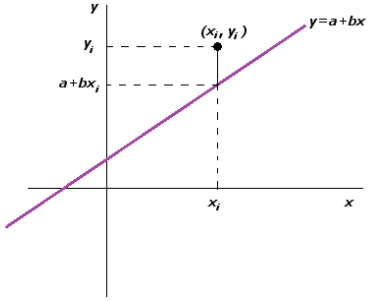
Planteamos la siguiente cuestión: de entre todas las rectas $y = a + bx$, *¿cuál es la que "mejor" aproxima/explica la muestra?*

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 para elegir la recta. Definimos la función que a cada $(a, b) \in \mathbb{R}^2$ le asocia $E(a, b)$ dada por

$$E(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + b\,x_i))^2 \tag{1.15}$$



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 $(y_1, \ldots, y_n) \in \mathbb{R}^n$ procedente de la muestra, y por otro el vector $(a + bx_1, \ldots, a + bx_n) \in \mathbb{R}^n$, que sería el vector de imágenes de $(x_1, \ldots, x_n)$ si el modelo lineal se cumpliera. En la fórmula (1.15) 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 (1.15) como

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

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 (1.15) (aunque no tanto en la última fórmula).

En la siguiente reescritura separamos las distintas contribuciones a $E(a,b)$:

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

La matriz de la forma cuadrática tiene, como determinante, a la varianza V_x (¡vaya!). En cualquier caso, es definida positiva[6], y por tanto $E(a,b)$ tiende a $+\infty$ cuando $(a,b) \to \infty_{\mathbb{R}^2}$ (es decir, cuando $\sqrt{a^2+b^2} \to +\infty$). Este análisis preliminar garantiza que $E(a,b)$ tiene (al menos) un mínimo.

Como $E(a,b)$ es (infinitamente) diferenciable, cualquiera de sus posibles mínimos deberá ser, también, un punto crítico. Para determinar la ubicación de los puntos críticos de $E(a,b)$, igualamos a 0 las dos siguientes ecuaciones:

$$\begin{cases} \dfrac{\partial E(a,b)}{\partial a} = 2a + 2\overline{x}b - 2\overline{y}\,, \\ \dfrac{\partial E(a,b)}{\partial b} = 2\overline{x^2}b + 2\overline{x}a - 2\overline{xy}\,. \end{cases}$$

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

Esto nos dice que la función $E(a,b)$ tiene un *único* punto crítico (que será el mínimo buscado), y que es la solución del sistema

$$\begin{cases} \overline{y} = a + b\,\overline{x}\,, \\ \overline{xy} = a\,\overline{x} + b\,\overline{x^2}\,. \end{cases} \tag{1.16}$$

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

$$\boxed{\begin{aligned} \widehat{b} &= \frac{\mathrm{COV}_{x,y}}{V_x} \\ \widehat{a} &= \overline{y} - \widehat{b}\,\overline{x} = \overline{y} - \Big(\frac{\mathrm{COV}_{x,y}}{V_x}\Big)\overline{x} \end{aligned}} \tag{1.17}$$

La recta de ecuación

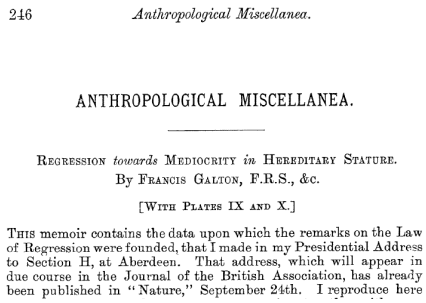
$$\boxed{y = \widehat{a} + \widehat{b}\,x} \tag{1.18}$$

donde $\widehat{a}$ y $\widehat{b}$ vienen dados en (1.17), y que da el mínimo error cuadrático medio, es la **recta de regresión de Y sobre X**.

[6] Sugerimos al lector que no esté cómodo con esta noción la consulta del apartado 3.1.

☞ **Nota** 1.3.3. El nombre de "regresión" 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. El uso moderno y técnico del término "regresión" ha olvidado ese origen.

En los términos de la ecuación (1.18), 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. En las figuras, la primera página del artículo y una gráfica del estudio.



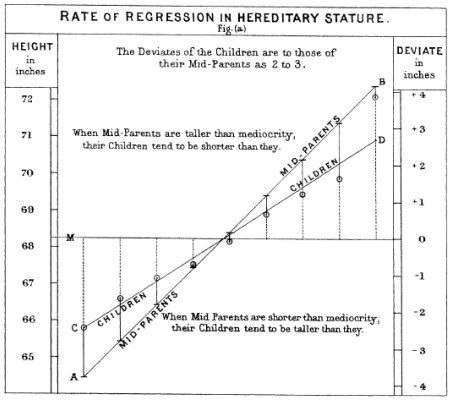
246 *Anthropological Miscellanea.*

ANTHROPOLOGICAL MISCELLANEA.

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

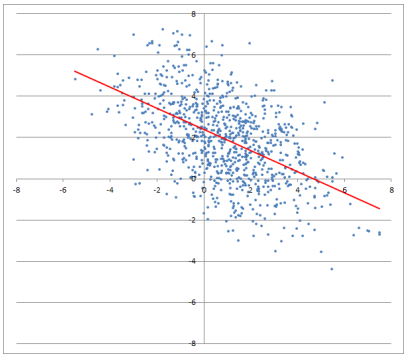
[WITH PLATES IX AND X.]

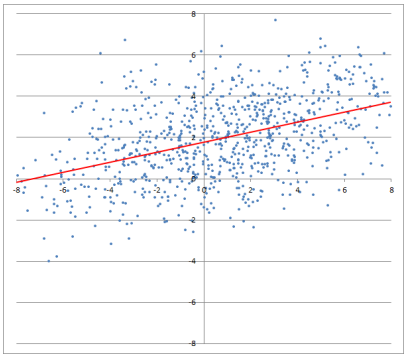
THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th.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



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

Ambos parámetros, $\widehat{a}$ y $\widehat{b}$, y por tanto la recta de regresión, están unívocamente determinados por la muestra. Véanse un par de ejemplos rectas de regresión en las figuras siguientes.





Usando que $\overline{y} = \widehat{a} + \widehat{b}\,\overline{x}$ (véase la primera ecuación de (1.16), o la propia solución (1.17)), observamos que la ecuación de la recta de regresión puede escribirse como

$$\boxed{y - \overline{y} = \widehat{b}\,(x - \overline{x})} \tag{1.19}$$

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

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

Escribiendo la fórmula en (1.17) para la pendiente $\widehat{b}$ como

$$\widehat{b} = \frac{\mathrm{cov} x, y}{V_x} = \rho_{x,y} \, \frac{\sqrt{V_y}}{\sqrt{V_x}}\,, \tag{1.20}$$

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

$$\boxed{\frac{y - \overline{y}}{\sqrt{V_y}} = \rho_{x,y} \, \frac{x - \overline{x}}{\sqrt{V_x}}} \tag{1.21}$$

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, $\widehat{b} = (y_2 - y_1)/(x_2 - x_1)$.

b) Si $\mathrm{cov}_{x,y} = 0$, la recta de regresión es la recta horizontal $y = \overline{y}$.

c) Si los datos x e y están de partida tipificados (es decir, si $\overline{x} = \overline{y} = 0$ 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 entre las series x e y.

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 en cuestión 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

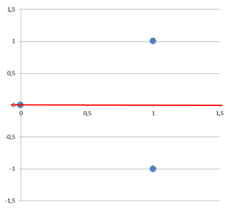
$$x = \rho_{x,y} \, y.$$

Como ejemplo adicional, para datos

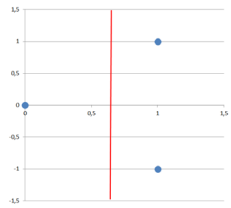
X	0	1	1
Y	0	1	-1

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

regresión Y sobre X



regresión X sobre Y



C. Bondad del ajuste a la recta de regresión

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 $\widehat{a}$ y $\widehat{b}$ en $E(a,b)$. Las cantidades

$$E(\widehat{a},\widehat{b}), \qquad \text{o mejor,} \qquad \sqrt{E(\widehat{a},\widehat{b})},$$

miden ese error cometido: $E(\widehat{a},\widehat{b})$ 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 $y-\overline{y}=\widehat{b}\,(x-\overline{x})$, y la fórmula (1.20) para $\widehat{b}$ podemos escribir

$$E(\widehat{a},\widehat{b})=\frac{1}{n}\sum_{i=1}^{n}\big((y_i-\overline{y})-\widehat{b}\,(x_i-\overline{x})\big)^2=V_y-2\,\widehat{b}\,\mathrm{cov}_{x,y}+\widehat{b}^{\,2}\,V_x=V_y\,(1-\rho^2_{x,y}).$$

Así que el error de ajuste a la recta de regresión se puede escribir como

$$\sqrt{E(\widehat{a},\widehat{b})}=\sqrt{V_y}\sqrt{1-\rho^2_{x,y}}. \tag{1.22}$$

☞ **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

$$z_i=\frac{x_i-\overline{x}}{\sqrt{V_x}} \qquad \text{y} \qquad w_i=\frac{y_i-\overline{y}}{\sqrt{V_y}}$$

Nótese que $\rho_{x,y}=\rho_{z,w}$. Con esta transformación lineal $\mathbb{R}^2\to\mathbb{R}^2$ 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 $\sqrt{1-\rho^2_{x,y}}$.

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

$$\frac{\sqrt{E(\widehat{a},\widehat{b})}}{\sqrt{V_y}}=\sqrt{1-\rho^2_{x,y}}, \tag{1.23}$$

y que mide cuán bueno es el ajuste de la nube de datos a la recta de regresión: si es pequeña (es decir, si $\rho^2_{x,y}$ es próximo a 1), significa buen ajuste.

La cantidad $\rho^2_{x,y}$ es conocida también como **coeficiente de determinación**[7], 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.

[7] En dos dimensiones, R^2 es simplemente el coeficiente de correlación al cuadrado. En más de dos dimensiones, el correspondiente coeficiente de determinación R^2 tiene una expresión no tan directa y que merece digna consideración. Véase el apartado 1.3.3, y en particular la fórmula (1.32).

(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, un análisis de regresión lineal consiste en calcular

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

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

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

La cantidad

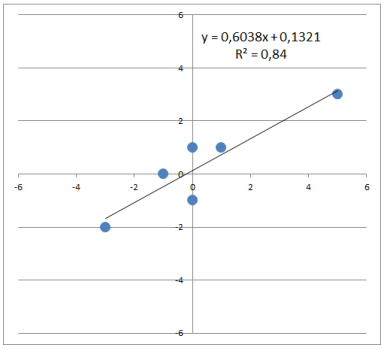
$$V_{y-\hat{y}} = V_y(1 - \rho^2_{x,y})$$

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

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

EJEMPLO 1.3.1. *Muestra*

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



Se tiene que

$$\overline{x} = 1/3, \qquad \overline{y} = 1/3,$$
$$V_x = 5.88, \qquad V_y = 2.55,$$
$$s_x^2 = 7.06, \qquad s_y^2 = 3.06.$$

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

D. Extrapolando con la recta de regresión

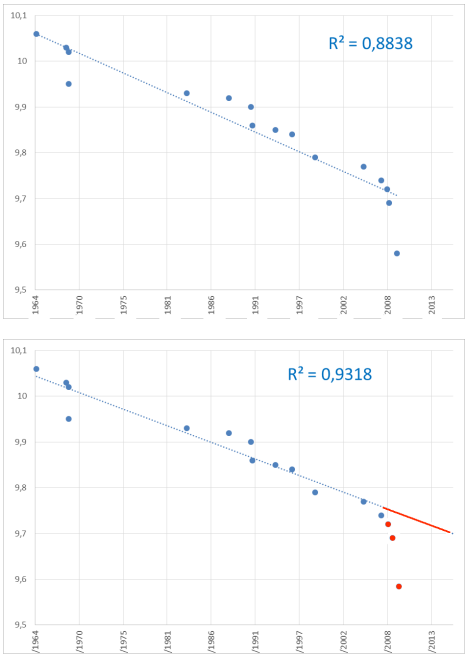
Si el ajuste lineal a los datos de la muestra que nos da la recta de regresión es "bueno", entendemos que los datos de la muestra siguen la tendencia lineal que marca esa recta. Eso permite extrapolar. Si quisiéramos el valor de y que *debería corresponder* a un valor x^* que no es uno de los x_i, tomaríamos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

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

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 México[8]; y los dos últimos puntos, que son los récords de Usain Bolt. Extraiga el lector conclusiones.

Si quitamos los tres últimos registros(época pre-Bolt), la recta de regresión tiene una pendiente menos acusada (hacia abajo), y el ajuste es algo mejor (mayor R^2). En la segunda gráfica prolongamos, en rojo, esta recta; esa extrapolación daría que el record del mundo debería haber estado, en 2009, en torno a 9.73, y no en el increíble 9.58 del Mundial de Berlín. Extrapolar puede llevar a conclusiones precipitadas. ♣

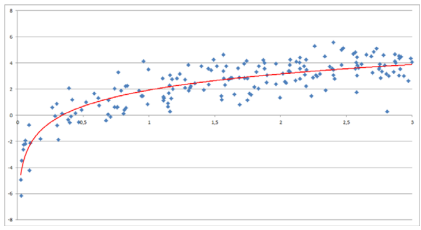
E. Transformación de datos: ajustes logarítmico y exponencial

La recta de regresión es el mejor ajuste lineal a los datos de la muestra. Si el error en el ajuste es pequeño, los datos de la muestra siguen (bastante aproximadamente) una relación lineal, y podremos por ejemplo usar la recta de regresión para 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.

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

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 $x_i > 0$.

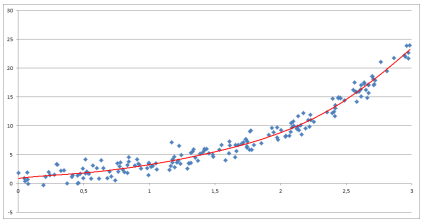
Podemos seguir el siguiente procedimiento:

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 $y = \widehat{a} + \widehat{b}z$,
6. el ajuste a los datos originales será

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

7. $A = \widehat{a}$ y $B = \widehat{b} = \dfrac{\mathrm{COV}_{\ln(x),y}}{V_{\ln(x)}}$.

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 $w = \widehat{a} + \widehat{b}x$,
6. el ajuste a los datos originales será

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

7. $C = e^{\widehat{a}}$ y $D = \widehat{b} = \dfrac{\mathrm{COV}_{x,\ln(y)}}{V_x}$.

En ambos casos, se considera que la bondad de ajuste viene dada por el coeficiente de determinación R^2 del ajuste lineal.

☞ **Nota** 1.3.6. Si planteamos el problema desde primeros principios, para el caso 1 podríamos considerar, en paralelo a (1.15), la función de error

$$E(A,B)=\frac{1}{n}\sum_{i=1}^{n}\big(y_i-(B\,\ln(x_i)+a)\big)^2,$$

para luego buscar A y B que minimizaran este error cuadrático medio. La solución coincide con la explicada en el texto.

Para el caso 2, sin embargo, la minimización de

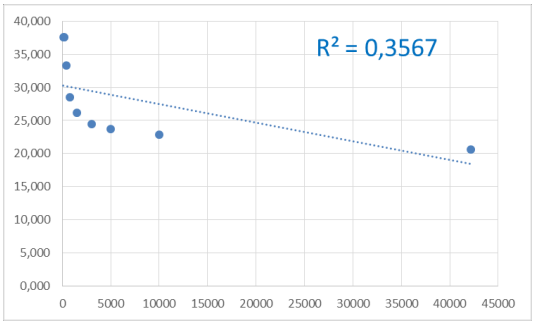
$$E(C,D)=\frac{1}{n}\sum_{i=1}^{n}\big(y_i-(C\,e^{Dx_i})\big)^2,$$

no tiene por qué coincidir con el resultado del procedimiento del texto, en el que en realidad minimizamos $E(C,D)=\frac{1}{n}\sum_{i=1}^{n}\big(\ln(y_i)-(Dx_i+\ln(C))\big)^2$.

Ejemplo 1.3.3. *Más sobre récords del mundo.*

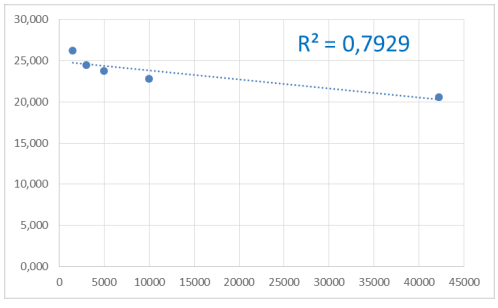
La siguiente tabla contiene las mejores marcas mundiales de las carreras incluidas en el programa olímpico. Se incluye una columna con los tiempos en segundos, y otra con la velocidad en kilómetros por hora.

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



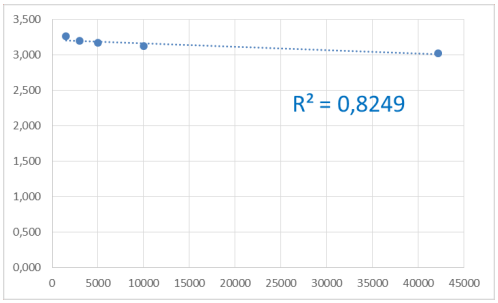
A la derecha representamos la nube de puntos (distancias-velocidades), recta de regresión incluida. Obsérvese, gráficamente o con el valor del R^2, cómo el ajuste es bastante pobre.

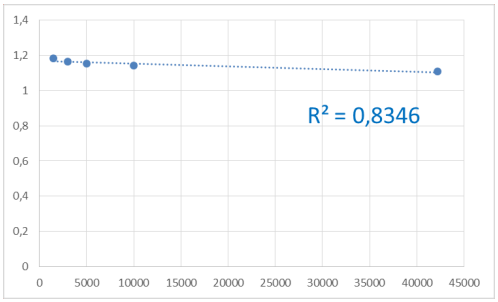
Adelantamos ya que el objetivo en este ejemplo es extrapolar los datos para, por ejemplo, predecir a qué velocidad debería irse en una carrera muy larga, de 100 km o más. Con este planteamiento, conviene no considerar, en la lista anterior, los récords de carreras cortas (que en realidad es otro deporte), 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 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 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:

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

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: $\mathrm{cov}_{x,y}$, $\mathrm{cov}_{x,z}$ y $\mathrm{cov}_{y,z}$, 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

$$E(a, b, c) = \frac{1}{n} \sum_{j=1}^{n} \big(z_j - (a + b\, x_j + c\, y_j)\big)^2 \tag{1.24}$$

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 $\mathrm{cov}_{x,x}$.

Para minimizar la función E definida en (1.24), calculamos las tres derivadas parciales de E e igualamos a 0 para obtener el sistema de ecuaciones siguiente:

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

que, escrito en forma matricial, es

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

☞ **Nota** 1.3.7. La matriz del sistema anterior es definida positiva, pues

$$(a,b,c) \begin{pmatrix} 1 & \overline{x} & \overline{y} \\ \overline{x} & \overline{x^2} & \overline{xy} \\ \overline{y} & \overline{xy} & \overline{y^2} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} (a + bx_i + cy_i)^2.$$

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

$$\begin{pmatrix} \overline{z} \\ \mathrm{cov}_{x,z} \\ \mathrm{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} 1 & \overline{x} & \overline{y} \\ 0 & \mathrm{cov}_{x,x} & \mathrm{cov}_{x,y} \\ 0 & \mathrm{cov}_{x,y} & \mathrm{cov}_{y,y} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{1.25}$$

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

$$\overline{z} = a + b\,\overline{x} + c\,\overline{y}, \tag{1.26}$$

y por otro

$$\begin{pmatrix} \mathrm{cov}_{x,z} \\ \mathrm{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} \mathrm{cov}_{x,x} & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & \mathrm{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} b \\ c \end{pmatrix}. \tag{1.27}$$

Llamemos $(\widehat{a}, \widehat{b}, \widehat{c})$ a la solución del sistema de ecuaciones (1.25). La ecuación matricial (1.27) permite hallar $\widehat{b}$ y $\widehat{c}$, para luego sustituir en (1.26) y deducir finalmente el valor de $\widehat{a}$. Apuntamos, por cierto, que como se aprecia en (1.27), los valores $\widehat{b}$ y $\widehat{c}$ no dependen de las medias de las muestras, sino solo de varianzas y covarianzas.

El lector interesado puede hallar una expresión explícita para $(\widehat{a}, \widehat{b}, \widehat{c})$ en términos de las tres covarianzas, las tres varianzas, y las tres medias muestrales.

B. Coeficiente R^2 de determinación múltiple

En el punto de coordenadas $(\widehat{a}, \widehat{b}, \widehat{c})$, la función E definida en (1.24) alcanza su mínimo. Llamemos

$$\text{EM} = E(\widehat{a}, \widehat{b}, \widehat{c}), \tag{1.28}$$

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

Usando la definición (1.24) de la función $E(a, b, c)$, observamos que

$$E(\overline{z}, 0, 0) = \frac{1}{n}\sum_{j=1}^{n}(z_j - \overline{z})^2 = \text{cov}_{z,z},$$

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

$$1 = \frac{E(\overline{z}, 0, 0)}{\text{cov}_{z,z}} \geq \frac{E(\widehat{a}, \widehat{b}, \widehat{c})}{\text{cov}_{z,z}}, \tag{1.29}$$

lo que permite escribir

$$\boxed{\frac{E(\widehat{a}, \widehat{b}, \widehat{c})}{\text{cov}_{z,z}} = 1 - R^2} \tag{1.30}$$

La cantidad R^2 es positiva (por la condición (1.29)), y 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 (1.25) para obtener los valores $(\widehat{a}, \widehat{b}, \widehat{c})$;
- insertarlos en la expresión (1.24) 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 (1.33).

Lector, si no tiene otra cosa que hacer, puede aplicarse a la tarea de obtener expresiones explícitas de $\widehat{a}, \widehat{b}, \widehat{c}$, para luego llevarlos a (1.24), ponerse a simplificar.... Pero es mucho más instructivo argumentar como sigue.

Observamos primero que el error mínimo EM no depende de $\widehat{a}$. Veamos. Como el triple $(\widehat{a}, \widehat{b}, \widehat{c})$ es solución de (1.25), en particular se tiene, por la primera ecuación (1.26), que

$$\overline{z} = \widehat{a} + \widehat{b}\,\overline{x} + \widehat{c}\,\overline{y}.$$

Llevando esta identidad en (1.24) observamos que

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

donde ya no aparece $\widehat{a}$.

Desarrollando el cuadrado,

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

Ahora, usando las ecuaciones de (1.27), concluimos que

$$\mathrm{EM} = E(\widehat{a}, \widehat{b}, \widehat{c}) = \mathrm{cov}_{z,z} - (\widehat{b}, \widehat{c}) \begin{pmatrix} \mathrm{cov}_{x,z} \\ \mathrm{cov}_{y,z} \end{pmatrix}. \tag{1.31}$$

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

Lema 1.2 *La expresión* $\mathrm{EM}/\mathrm{cov}_{z,z}$ *es invariante bajo cambios de escala en las muestras de* Z, X *e* Y.

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

Demostración del lema 1.2. Si partimos de las muestras de Z, X e Y, los valores $\widehat{b}$ y $\widehat{c}$ que minimizan la función de error verifican, por (1.27),

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

y el error mínimo es

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

Digamos ahora que las muestras son Z, $X' = \lambda X$ e Y, donde $\lambda > 0$. Llamemos $\widehat{b}'$ y $\widehat{c}'$ a los valores que minimizan la función de error en este caso. Estos valores verifican, por (1.27),

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

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

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

El error mínimo sería

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

que coincide, tras cancelar un factor λ, con el EM original. Así que EM no varía bajo cambios de escala de X, y tampoco $\mathrm{EM}/\mathrm{cov}_{z,z}$, 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 $\widehat{b}'$ y $\widehat{c}'$ a los valores que minimizan la función de error en este caso, que verifican

$$\begin{pmatrix} \mathrm{cov}_{x,z'} \\ \mathrm{cov}_{y,z'} \end{pmatrix} = \begin{pmatrix} \mathrm{cov}_{x,x} & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & \mathrm{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \widehat{b}' \\ \widehat{c}' \end{pmatrix}$$
$$\iff \begin{pmatrix} \lambda\,\mathrm{cov}_{x,z} \\ \lambda\,\mathrm{cov}_{y,z} \end{pmatrix} = \begin{pmatrix} \mathrm{cov}_{x,x} & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & \mathrm{cov}_{y,y} \end{pmatrix} \cdot \begin{pmatrix} \widehat{b}' \\ \widehat{c}' \end{pmatrix}.$$

Ahora tenemos que

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

y el error mínimo sería

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

El error mínimo cambia en esta ocasión, pero

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

lo que concluye la demostración. ■

Recordemos que buscamos una fórmula explícita para R^2. El lema anterior nos dice que 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

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

(véanse (1.30) y (1.31)), donde $\widehat{b}$ y $\widehat{c}$ verifican, siguiendo (1.27),

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

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

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

Así que, finalmente,

$$\boxed{R^2 = \frac{1}{1-\rho_{x,y}^2} (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & -\rho_{x,y} \\ -\rho_{x,y} & 1 \end{pmatrix} \begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix}} \tag{1.32}$$

que también se puede escribir como

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

Aunque el lector haría bien en recordar la expresión (1.32), o mejor,

$$\boxed{R^2 = (\rho_{x,z}, \rho_{y,z}) \begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{x,z} \\ \rho_{y,z} \end{pmatrix}} \tag{1.33}$$

(con la inversa de la matriz de correlaciones entre X e Y) si es que quisiera extender este análisis al caso de la regresión de una variable Z sobre unas variables $X_1, \ldots, X_k$, con $k \geq 3$.

1.4. Comandos de Excel

1.4.1. Una variable

Suponemos que los datos $x_1, \ldots, x_n$ 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)` o `=desvest.m(rango)`.
- Cuasivarianza muestral: `=var(rango)` o bien `=var.s(rango)` (o elevar al cuadrado las del punto anterior).
- Desviación típica muestral: `=desvestp(rango)` o `=desvest.p(rango)`.
- Varianza muestral: `=varp(rango)` o bien `=var.p(rango)`, o elevar al cuadrado las del punto anterior.
- 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

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

  donde s_x es la cuasidesviación típica. Obsérvese que no coincide exactamente con la definición de (1.9).

Si los datos vienen agrupados en clases $C_1, \ldots, C_k$, con marcas de clase $y_1, \ldots, y_k$ en `rangovalores` y frecuencias relativas $f_1, \ldots, f_k$ 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`:

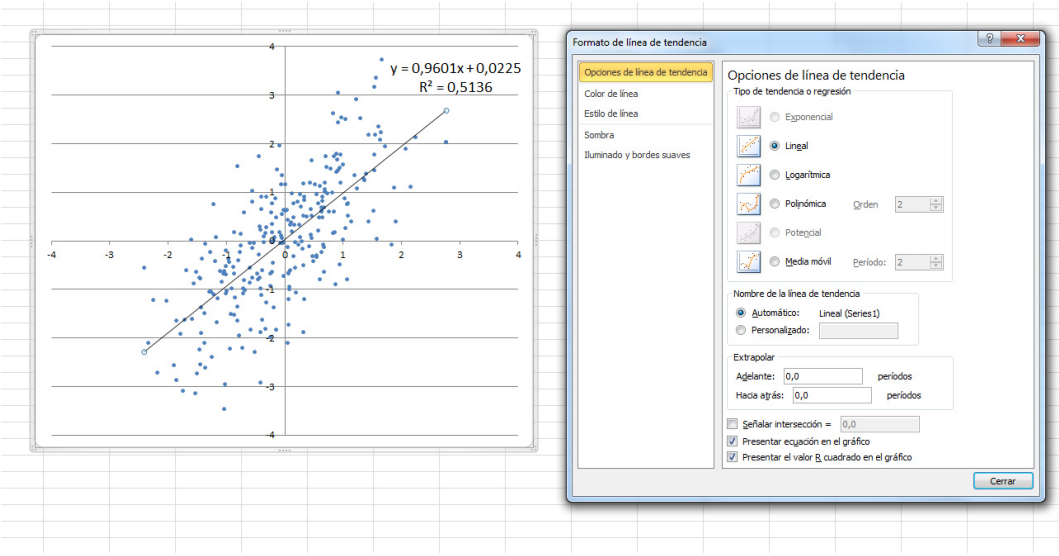
- con una instrucción del tipo `=contar.si(rango;F4)` contamos cuántas celdas del `rango` contienen el valor que aparece en la celda `F4`;
- con una instrucción del tipo `=contar.si(rango;"<="&F4)` contamos cuántas celdas del `rango` contienen valores menores o iguales que el valor que aparece en la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

1.4.2. Dos variables

Suponemos que los datos $x_1, \ldots, x_n$ están recogidos en el rango de celdas al que llamaremos `rangoX`, y los datos $y_1, \ldots, y_n$ en `rangoY`.

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

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 (marcando las casillas correspondientes al final de la ventana) de incluir en el gráfico la ecuación del recta y el valor de R^2.



Estos valores se pueden calcular también con las fórmulas correspondientes (1.17), o utilizando las funciones

- `=interseccion.eje(rangoY;rangoX)` para $\widehat{a}$,
- `=pendiente(rangoY;rangoX)` para $\widehat{b}$,
- `=coeficiente.R2(rangoY;rangoX)` 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).