# Relaciones entre variables

Las técnicas de regresión permiten hacer predicciones sobre los valores de cierta variable $Y$ (dependiente), a partir de los de otra $X$ (independiente), entre las que se intuye que existe una relación. Para ilustrarlo retomemos los ejemplos mencionados al principio del tema anterior. Si sobre un grupo de personas observamos los valores que toman las variables

$$\begin{aligned} X &\equiv \text{Altura medida en cm} \\ Y &\equiv \text{Altura medida en metros} \end{aligned}$$

es trivial observar que la relación que hay entre ambas es: $Y = \frac{X}{100}$.

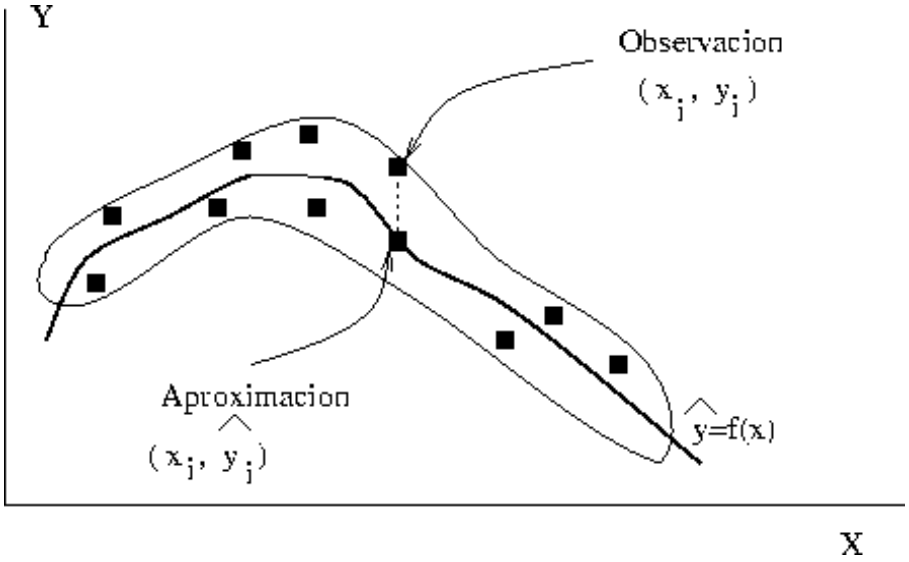
Obtener esta relación es menos evidente cuando lo que medimos sobre el mismo grupo de personas es, por ejemplo,

$$\begin{aligned} X &\equiv \text{Altura medida en cm} \\ Y &\equiv \text{Peso en kilos} \end{aligned}$$

La razón es que no es cierto que conocida la altura $x_i$ de un individuo, podamos determinar de modo exacto su peso $y_i$ (dos personas que miden 1,70 m pueden tener pesos de 60 y 65 kilos). Sin embargo, alguna relación entre ellas debe existir, pues parece mucho más probable que un individuo de $2m$ pese más que otro que mida $1.20m$. Es más, nos puede parecer más o menos aproximado una relación entre ambas variables como la siguiente

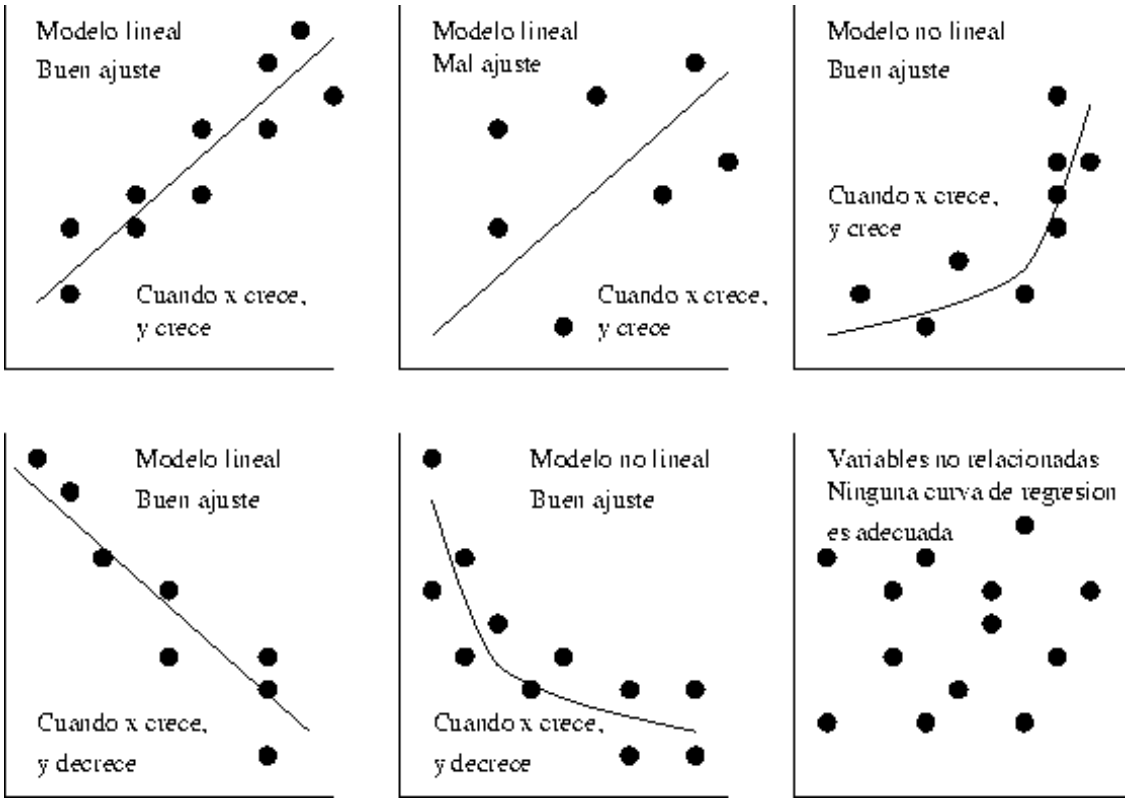
$$Y = X - 110 \pm (\text{error}).$$

A la deducción, a partir de una serie de datos, de este tipo de relaciones entre variables, es lo que denominamos **regresión**.




Mediante las técnicas de regresión de una variable $Y$ sobre una variable $X$, buscamos una función que sea una buena aproximación de una nube de puntos $(x_i, y_i)$, mediante una curva. Para ello hemos de asegurarnos de que la diferencia entre los valores $y_i$ e $\hat{y}_i$ sea tan pequeña como sea posible.

El término que hemos denominado *error* debe ser tan pequeño como sea posible (ver figura). El objetivo será buscar la función (también denominada modelo de regresión) $\hat{Y} = f(X)$ que minimice dicho error.




## Bondad de un ajuste

Consideremos un conjunto de observaciones sobre $n$ individuos de una población, en los que se miden ciertas variables $X$ e $Y$,

$$\begin{aligned} X &\hookrightarrow x_1, x_2, \ldots, x_n \\ Y &\hookrightarrow y_1, y_2, \ldots, y_n \end{aligned}$$

Estamos interesamos en hacer una regresión para determinar, de modo aproximado, los valores de $Y$ conocidos los de $X$. Así, debemos definir cierta variable $\hat{Y} = f(X)$, que debe tomar los valores

$$\begin{aligned} \hat{y}_1 &= f(x_1), \\ \hat{y}_2 &= f(x_2), \\ &\ldots \\ \hat{y}_n &= f(x_n), \end{aligned}$$

de modo que:

$$\begin{aligned} y_1 - \hat{y}_1 &\approx 0, \\ y_2 - \hat{y}_2 &\approx 0, \\ &\dots \\ y_n - \hat{y}_n &\approx 0, \end{aligned}$$

Ello se puede expresar definiendo una nueva variable $E = Y - \hat{Y}$ que mida las diferencias entre los auténticos valores de $Y$ y los teóricos suministrados por la regresión,

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1, \\ e_2 &= y_2 - \hat{y}_2, \\ &\dots \\ e_n &= y_n - \hat{y}_n. \end{aligned}$$

y calculando $\hat{Y}$ de modo que $E$ tome valores cercanos a 0. Dicho de otro modo, $E$ debe ser una variable cuya media debe ser 0 y cuya varianza $S_E^2$ debe ser pequeña (en comparación con la de $Y$).

Por ello, se define el coeficiente de **determinación** de la regresión de $Y$ sobre $X$, $R^2$, como

$$R^2 = \frac{S_Y^2 - S_E^2}{S_Y^2} = 1 - \frac{S_E^2}{S_Y^2}.$$

Si el ajuste de $Y$ mediante la curva de regresión $\hat{Y} = f(X)$ es bueno, cabe esperar que la cantidad $R^2$ tome un valor próximo a 1.

Análogamente, si nos interesa encontrar una curva de regresión para $X$ como función de $Y$, definiríamos $\hat{X} = f(Y)$ y se procedería del mismo modo en las definiciones.

El valor de $R^2$ sirve, entonces, para medir de qué modo las diferencias entre los verdaderos valores de una variable y los de su aproximación mediante una curva de regresión son pequeñas en relación con los de la variabilidad de la variable que intentamos aproximar. Por esta razón estas cantidades miden el grado de bondad del ajuste.

## Regresión lineal

La forma de la función $f$ en principio, podría ser arbitraria, y tal vez se tenga que la relación más exacta entre las variables peso y altura, definidas anteriormente, sea algo de forma muy complicada.

Por el momento no pretendemos encontrar relaciones complicadas entre variables, pues nos vamos a limitar al caso de la regresión lineal. Con este tipo de regresiones nos conformamos con encontrar relaciones funcionales de tipo lineal, es decir, buscamos cantidades $a$ y $b$ tales que se pueda escribir $\hat{Y} = a + bX$ con el menor error posible entre $\hat{Y}$ e $Y$.

**Observación**

Obsérvese que la relación anterior explica cosas como que si $X$ varía en 1 unidad, $Y$ varía la cantidad $b$. Por tanto:

1. Si $b > 0$, las dos variables aumentan o disminuyen a la vez;
2. Si $b < 0$, cuando una variable aumenta, la otra disminuye.

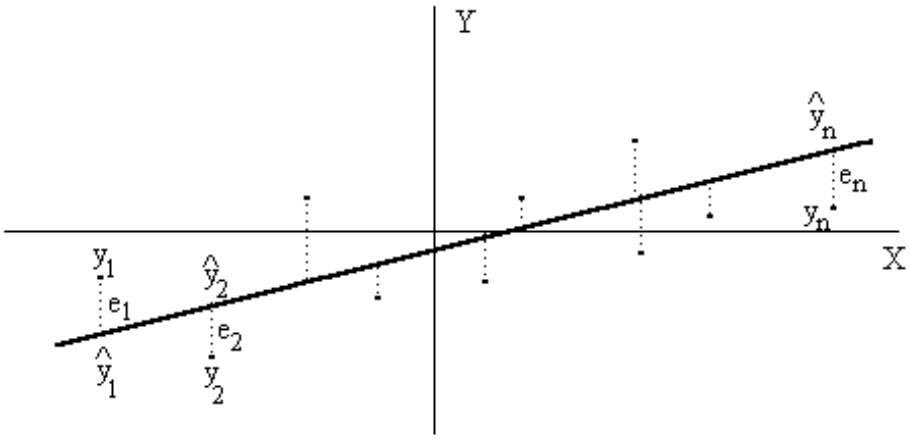
Por tanto, en el caso de las variables peso y altura lo lógico será encontrar que $b > 0$.

El problema que se plantea es, entonces, el de cómo calcular las cantidades $a$ y $b$ a partir de un conjunto de $n$ observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, de forma que se minimice el error. Las etapas en que se divide el proceso son de forma esquemática, las que siguen:

1. Dadas dos variables $X$, $Y$, sobre las que definimos $\hat{Y} = a + bX$ medimos el error que se comete al aproximar $Y$ mediante $\hat{Y}$ calculando la suma de las diferencias entre los valores reales y los aproximados al cuadrado (para que sean positivas y no se compensen los errores):
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$
2. Una aproximación $\hat{Y} = a + bX$ de $Y$, se define a partir de dos cantidades $a$ y $b$. Vamos a calcular aquellas que minimizan la función
$$Error\,(a, b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$
3. Posteriormente encontraremos fórmulas para el cálculo directo de $a$ y $b$ que sirvan para cualquier problema.

## Regresión de Y sobre X

Para calcular la recta de regresión de $Y$ sobre $X$ nos basamos en la siguiente figura




Una vez que tenemos definido el error de aproximación, los valores $a$ y $b$ que lo minimizan se calculan derivando con respecto a ambas e igualando a cero (procedimiento de los *mínimos cuadrados*):

$$\begin{aligned}
&\underset{a,b}{\text{Min}} \sum_i (y_i - a - bx_i)^2 = \\
&\underset{a,b}{\text{Min}} \sum_i \left(y_i^2 + a^2 + b^2 x_i^2 - 2ay_i - 2bx_i y_i + 2abx_i\right) = \\
&\underset{a,b}{\text{Min}} \left(\sum_i y_i^2 + na^2 + b^2 \sum_i x_i^2 - 2a \sum_i y_i - 2b \sum_i x_i y_i + 2ab \sum_i x_i\right).
\end{aligned}$$

Se deriva e iguala a 0:

$$\begin{aligned}
\frac{\partial}{\partial a} &= 2na - 2\sum_i y_i + 2b \sum_i x_i = 0 \\
\frac{\partial}{\partial b} &= 2b \sum_i x_i^2 - 2 \sum_i x_i y_i + 2a \sum_i x_i = 0
\end{aligned}$$

Despejando los valores de $a$ y $b$, se obtienen las relaciones buscadas:

$$\begin{aligned}
a &= \bar{y} - b\bar{x} \\
b &= \frac{S_{XY}}{S_X^2}
\end{aligned}$$

La cantidad $b$ se denomina *coeficiente de regresión* de $Y$ sobre $X$.

Las mismas conclusiones se sacan cuando intentamos hacer la regresión de $X$ sobre $Y$, pero, para calcular la recta de regresión de $X$ sobre $Y$ es totalmente incorrecto despejar de $\hat{Y} = a + bX$.

La regresión de $X$ sobre $Y$ se hace aproximando $X$ por $\hat{X}$ del modo $\hat{X} = a + bY$ donde

$$\begin{aligned} a &= \bar{x} - b\bar{y} \\ b &= \frac{S_{XY}}{S_Y^2} \end{aligned}$$

pues de este modo se minimiza, en el sentido de los mínimos cuadrados, los errores entre las cantidades $x_i$ y las $\hat{x}_i$

**Ejemplo**

En una muestra de 1.500 individuos se recogen datos sobre dos medidas antropométricas $X$ e $Y$. Los resultados se muestran resumidos en los siguientes estadísticos:

| | | |
|---|---|---|
| $\bar{x} = 14$ | $S_X = 2$ | $S_{XY} = 45$ |
| $\bar{y} = 100$ | $S_Y = 25$ | |

Obtener el modelo de regresión lineal que mejor aproxima $Y$ en función de $X$. Utilizando este modelo, calcular de modo aproximado la cantidad $Y$ esperada cuando $X = 15$.

**Solución**:

Lo que se busca es la recta, $\hat{Y} = a + bX$, que mejor aproxima los valores de $Y$ (según el criterio de los mínimos cuadrados) en la nube de puntos que resulta de representar en un plano $(X, Y)$ las 1.500 observaciones. Los coeficientes de esta recta son:

$$\begin{aligned} b &= \frac{S_{XY}}{S_X^2} = \frac{45}{4} = 11{,}25 \\ a &= \bar{y} - b\bar{x} = 100 - 11{,}25 \cdot 14 = -57{,}5 \end{aligned}$$

Así, el modelo lineal consiste en:

$$\hat{Y} = -57{,}5 + 11{,}25 \cdot X$$

Por tanto, si $x = 15$, el modelo lineal predice un valor de $Y$ de:

$$\hat{y} = -57{,}5 + 11{,}25 \cdot x = -57{,}5 + 11{,}25 \cdot 15 = 111{,}25$$

En este punto, hay que preguntarse si realmente esta predicción puede considerarse fiable. Para dar una respuesta, es necesario estudiar propiedades de la regresión lineal que están a continuación.

## Propiedades de la regresión lineal

Una vez que ya tenemos perfectamente definida $\hat{Y}$ nos preguntamos las relaciones que hay entre la media y la varianza de ésta y la de $Y$. La respuesta nos la ofrece la siguiente proposición:

### Proposición

En los ajustes lineales se conserva la media, es decir

$$\overline{\hat{y}} = \bar{y}$$

En cuanto a la varianza, no necesariamente es la misma para los verdaderos valores de $Y$ y su aproximación $\hat{Y}$, pues sólo se mantienen en un factor de $r^2$, es decir,

$$S^2_{\hat{Y}} = r^2 S^2_Y$$

**Demostración**: Se tiene que

$$\begin{aligned} \overline{\hat{y}} &= a + b\bar{x} = (\bar{y} - b\bar{x} + b\bar{x}) = \bar{y} \\ S^2_{\hat{Y}} &= b^2 S^2_X = \frac{S^2_{XY}}{S^2_X \cdot S^2_X} \cdot S^2_X = \\ &= \frac{S^2_{XY}}{S^2_X \cdot S^2_Y} \cdot S^2_Y = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2 \cdot S^2_Y = r^2 \cdot S^2_Y \end{aligned}$$

donde se ha utilizado la magnitud que denominamos coeficiente de correlación, $r$, y que ya definimos anteriormente como

$$r^2 = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2$$

### Observación

Como consecuencia de este resultado, podemos decir que la proporción de varianza explicada por la regresión lineal es del $r^2 \cdot 100\,\%$.

Nos gustaría obtener que $r = 1$, pues en ese caso ambas variables tendrían la misma varianza, pero esto no es cierto en general. Todo lo que se puede afirmar, como sabemos, es que $-1 \leq r^2 \leq 1$, y por tanto,

$$0 \leq S^2_{\hat{Y}} \leq S^2_Y$$

La cantidad que le falta a la varianza de la regresión, $S^2_{\hat{Y}}$, para llegar hasta la varianza total de $Y$, $S^2_Y$, es lo que se denomina *varianza residual*, que no es más que la varianza de $E = Y - \hat{Y}$, ya que

$$\begin{aligned} S^2_Y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}} + e_i\right)^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)^2 + \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)e_i = \\ &= S^2_{\hat{Y}} + S^2_E + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)e_i = S^2_{\hat{Y}} + S^2_E \end{aligned}$$

ya que el tercer sumando se anula según las ecuaciones *normales*:

$$\begin{aligned} \sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)e_i &= \sum_{i=1}^{n} e_i\left(a + bx_i - [a + b\bar{x}]\right) = \\ &= b\sum_{i=1}^{n} e_i\left(x_i - \bar{x}\right) = b\sum_{i=1}^{n} e_i x_i - b\bar{x}\sum_{i=1}^{n} e_i = \\ &= b\sum_{i=1}^{n} e_i x_i = 0 \end{aligned}$$

Por ello,

$$S^2_E = S^2_Y - S^2_{\hat{Y}}$$

Obsérvese que entonces la bondad del ajuste es

$$\begin{aligned} R^2 &= \frac{S^2_Y - S^2_E}{S^2_Y} = 1 - \frac{S^2_E}{S^2_Y} = 1 - \frac{S^2_Y - S^2_{\hat{Y}}}{S^2_Y} = \\ &= 1 - \frac{S^2_Y - r^2 \cdot S^2_Y}{S^2_Y} = 1 - \left(1 - r^2\right) = r^2 \end{aligned}$$

lo que resumimos en la siguiente proposición:

**Proposición**

Para los ajustes de tipo lineal se tiene que el coeficiente de determinación es igual a $r^2$, y por tanto representa la proporción de varianza explicada por la regresión lineal: $R^2 = r^2$.

Por ello:

- Si $|r| \approx 1$ el ajuste es bueno, es decir, $Y$ se puede calcular de modo bastante aproximado a partir de $X$ y viceversa.

- Si $|r| \approx 0$ las variables $X$ e $Y$ no están relacionadas (linealmente al menos), por tanto no tiene sentido hacer un ajuste lineal. Sin embargo no es seguro que las dos variables no posean ninguna relación en el caso $r = 0$, ya que si bien el ajuste lineal puede no ser procedente, tal vez otro tipo de ajuste de tipo cuadrático sí lo sea.

**Ejemplo**

De una muestra de ocho observaciones conjuntas de valores de dos variables $X$ e $Y$, se obtiene la siguiente información:

| | | |
|---|---|---|
| $\sum_i x_i = 24$ | $\sum_i x_i y_i = 64$ | $\sum_i y_i = 40$ |
| $S_Y^2 = 12$ | $S_X^2 = 6$ | |

Calcular:

1. La recta de regresión de $Y$ sobre $X$. Explicar el significado de los parámetros.
2. El coeficiente de determinación. Comentar el resultado e indique el tanto por ciento de la variación de $Y$ que no está explicada por el modelo lineal de regresión.
3. Si el modelo es adecuado, ¿cuál es la predicción $\hat{y}$ para $x = 4$?

**Solución**:

**1**. En primer lugar calculamos las medias y las covarianza entre ambas variables:

$$\begin{aligned} \bar{x} &= \frac{\sum_i x_i}{n} = \frac{24}{8} = 3 \\ \bar{y} &= \frac{\sum_i y_i}{n} = \frac{40}{8} = 5 \\ S_{XY} &= \frac{\sum_i x_i y_i}{n} - \bar{x} \cdot \bar{y} = \frac{64}{8} - 3 \cdot 5 = -7 \end{aligned}$$

Con estas cantidades podemos determinar los parámetros $a$ y $b$ de la recta. La pendiente de la misma es $b$, y mide la variación de $Y$ cuando $X$ aumenta en una unidad:

$$b = \frac{S_{XY}}{S_X^2} = \frac{-7}{6} = -1{,}167$$

Al ser esta cantidad negativa, tenemos que la pendiente de la recta es negativa, es decir, a medida que $X$ aumenta, la tendencia es a la disminución de $Y$. En cuanto al valor de la ordenada en el origen, $a$, tenemos

$$a = \bar{y} - b \cdot \bar{x} = 5 - \left(\frac{-7}{6}\right) \cdot 3 = 8{,}5$$

Así, la recta de regresión de $Y$ como función de $X$ es

$$\hat{Y} = 8{,}5 - 1{,}167 \cdot X$$

**2**. El grado de bondad del ajuste lo obtenemos a partir del coeficiente de determinación:

$$R^2 = r^2 = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2 = \frac{-7}{6 \cdot 12} = 0{,}68 \Longrightarrow 68\,\%$$

Es decir, el modelo de regresión lineal explica el 68 % de la variabilidad de $Y$ en función de la de $X$. Por tanto, queda un 32 % de variabilidad no explicada.

**3**. La predicción que realiza el modelo lineal de regresión para $x = 4$ es:

$$\hat{y} = 8{,}5 - 1{,}167 \cdot x = 8{,}5 - 1{,}167 \cdot 4 = 3{,}83$$

que hay que considerar con ciertas reservas pues, como hemos visto en el apartado anterior, hay una razonable cantidad de variabilidad que no es explicada por el modelo.

**Ejemplo**

En un grupo de 8 pacientes se miden las cantidades antropométricas peso y edad, obteniéndose los siguientes resultados:

Resultado de las mediciones

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $X \equiv$edad | 12 | 8 | 10 | 11 | 7 | 7 | 10 | 14 |
| $Y \equiv$peso | 58 | 42 | 51 | 54 | 40 | 39 | 49 | 56 |

¿Existe una relación lineal importante entre ambas variables? Calcular la recta de regresión de la edad en función del peso y la del peso en función de la edad. Calcular la bondad del ajuste ¿En qué medida, por término medio, varía el peso cada año? ¿En cuánto aumenta la edad por cada kilo de peso?

**Solución**:

Para saber si existe una relación lineal entre ambas variables se calcula el coeficiente de correlación lineal, que vale:

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{15{,}20}{2{,}32 \cdot 6{,}96} = 0{,}94$$

ya que

$$\begin{aligned}
\sum_i x_i &= 79 \Longrightarrow \bar{x} = \frac{79}{8} = 9{,}88 \\
\sum_i y_i &= 389 \Longrightarrow \bar{y} = \frac{389}{8} = 48{,}63 \\
\sum_i x_i^2 &= 823 \Longrightarrow S_X^2 = \frac{823}{8} - 9{,}88^2 = 5{,}36 \Longrightarrow S_X = 2{,}32 \\
\sum_i y_i^2 &= 19{,}30 \Longrightarrow S_Y^2 = \frac{19{,}30}{8} - 48{,}63^2 = 48{,}48 \Longrightarrow S_Y = 6{,}96 \\
\sum_i x_i y_i &= 3{,}96 \Longrightarrow S_{XY} = \frac{\sum_i x_i y_i}{n} - \bar{x} \cdot \bar{y} = \frac{3{,}96}{8} - 9{,}88 \cdot 48{,}63 = 15{,}20
\end{aligned}$$

Por tanto el ajuste lineal es muy bueno. Se puede decir que el ángulo entre el vector formado por las desviaciones del peso con respecto a su valor medio y el de la edad con respecto a su valor medio, $\theta$, es

$$r = \cos\theta \Longrightarrow \theta = \operatorname{arc}\cos(r) \approx 19 \text{ grados}$$

es decir, entre esos vectores hay un buen grado de paralelismo (sólo unos 19 grados de desviación).

La recta de regresión del peso en función de la edad es es

$$\begin{aligned}
\hat{Y} &= a_1 + b_1 X = 20{,}61 + 2{,}84 \cdot X \\
a_1 &= \bar{y} - b_1 \bar{x} = 20{,}61 \\
b_1 &= \frac{S_{XY}}{S_X^2} = 2{,}84
\end{aligned}$$

La recta de regresión de la edad como función del peso

$$\begin{aligned}
\hat{X} &= a_2 + b_2 Y = -5{,}37 + 0{,}31 \cdot Y \\
a_2 &= \bar{x} - b_2 \bar{y} = -5{,}37 \\
b_2 &= \frac{S_{XY}}{S_Y^2} = 0{,}31
\end{aligned}$$

que, como se puede comprobar, no resulta de despejar en la recta de regresión de $Y$ sobre $X$.

La bondad del ajuste es $R^2 = r^2 = 0{,}889$, por tanto podemos decir que el 88,9 % de la variabilidad del peso en función de la edad es explicada mediante la recta de regresión correspondiente. Lo mismo podemos decir en cuanto a la variabilidad de la edad en función

del peso. Del mismo modo, puede decirse que hay un $100 - 88,94 = 11,06\,\%$ de varianza que no es explicada por las rectas de regresión. Por tanto, la varianza residual de la regresión del peso en función de la edad es

$$S_E^2 = \left(1 - r^2\right) S_Y^2 = 0{,}11 \cdot 48{,}48 = 5{,}33$$

y la de la edad en función del peso:

$$S_E^2 = \left(1 - r^2\right) S_X^2 = 0{,}11 \cdot 5{,}36 = 0{,}59$$

Por último, la cantidad en que varía el peso de un paciente cada año es, según la recta de regresión del peso en función de la edad, la pendiente de esta recta es $b_1 = 2{,}84$ $Kg/año$. Cuando dos personas difieren en peso, en promedio la diferencia de edad entre ambas se rige por la cantidad $b_2 = 0,3136$ $años/Kg$ de diferencia.