En 1965 A.J. Lea recogió datos sobre la temperatura anual media en varias ciudades (de Gran Bretaña, Noruega y Suecia) y la tasa de mortalidad en un tipo de cáncer de pecho en mujeres. Los datos que obtuvo fueron los siguientes:

Temperatura anual media (grados Fa.)	Índice de mortalidad
51′3	102′5
49′9	104′5
50′0	100′4
49′2	95′9
48′5	87′0
47′8	95′0
47′3	88′6
45′1	89′2
46′3	78′9
42′1	84′6
44′2	81′7
43′5	72′2
42′3	65′1
40′2	68′1
31′8	67′3
34′0	52′5

Determinar la recta de mínimos cuadrados así como la precisión conseguida con el ajuste obtenido mediante dicho método.

Aunque los datos del enunciado constituyen una distribución bidimensional de frecuencias, en donde la frecuencia absoluta de cada par es igual a 1, el principal interés sobre ellos suele ser el de determinar la ecuación de una función, generalmente una recta, que permita explicar una de las variables —denominada dependiente— en función de la otra —denominada independiente—, con el habitual propósito de hacer predicciones sobre la variable dependiente en función de la independiente.

En este ejercicio, el estudio de campo realizado tendrá interés si puede demostrarse una relación entre las variables *temperatura medio-ambiental* e *índice de mortalidad*. Si esto fuera así, se podría predecir, mediante la función ajustada, el índice de mortalidad que cabría esperar bajo una determinada temperatura medio-ambiental.

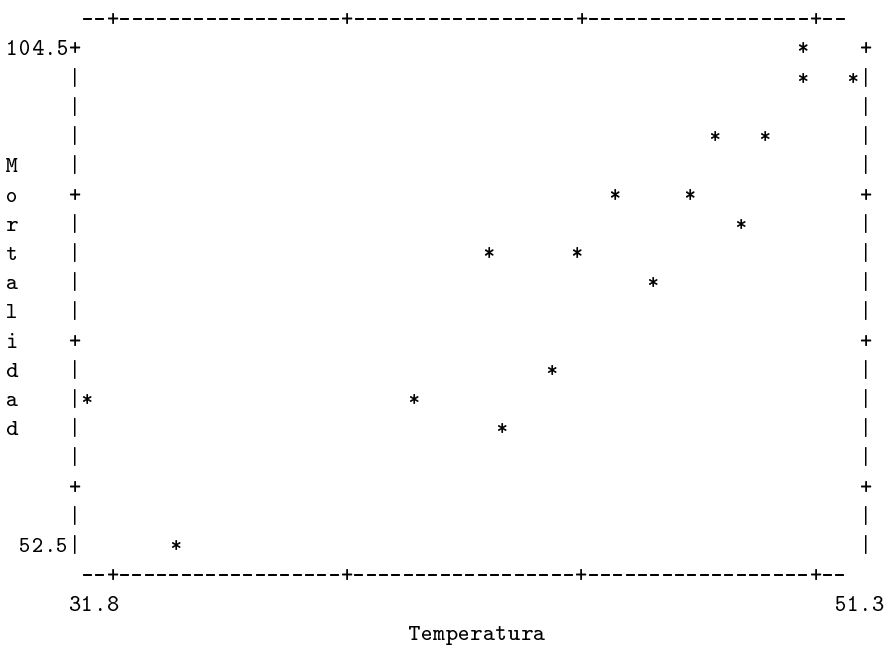
En este caso, por tanto, parece razonable considerar como variable independiente, X, la temperatura y como variable dependiente, Y, el índice de mortalidad.

No obstante todo lo que acabamos de decir, hacemos la observación de que, aunque con el coeficiente de determinación R^2, que calcularemos al final del problema, podemos calcular la bondad del ajuste que efectuemos, no será hasta que utilicemos las potentes técnicas de la Inferencia Estadística (en concreto de la Regresión Lineal) que podamos decidir si existe o no una relación lineal *significativa* entre ambas variables.

Aunque el ajuste por mínimos cuadrados (CB-sección 2.4.2) que se nos solicita es el de una recta, siempre es conveniente comenzar haciendo una representación gráfica de los pares de puntos dados, en lo que se denomina la *nube de puntos*, que no es más que la representación de los pares de puntos (x_i, y_i), $i = 1, ..., 16$, en unos ejes de coordenadas cartesianas, de forma que se pueda aventurar la bondad del ajuste que se va a realizar.

Es decir, si los datos aparecen alineados la recta de mínimos cuadrados explicará bien a la variable dependiente en función de la independiente, pero si los puntos muestran una gráfica en forma de parábola, es posible que un ajuste de tal función por mínimos cuadrados resulte más adecuado.

Para los datos de nuestro enunciado la nube de puntos es la siguiente



La disposición lineal de los datos, hace razonable el ajuste por mínimos cuadrados.

Como es sabido, la recta de mínimos cuadrados es la más próxima a la nube de puntos, la cual se determinó en CB que era la de ecuación

$$y = \widehat{\alpha} + \widehat{\beta}\, x$$

en donde $\widehat{\alpha}$ y $\widehat{\beta}$ eran los valores determinados por las ecuaciones

$$\widehat{\beta} = \frac{n \sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

y

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \widehat{\beta} \sum_{i=1}^{n} x_i}{n}.$$

Para calcularlos utilizaremos la siguiente tabla de cálculos

x_i	y_i	$x_i\, y_i$	x_i^2	y_i^2
$51'3$	$102'5$	$5258'25$	$2631'69$	$10506'25$
$49'9$	$104'5$	$5214'55$	$2490'01$	$10920'25$
$50'0$	$100'4$	5020	2500	$10080'16$
$49'2$	$95'9$	$4718'28$	$2420'64$	$9196'81$
$48'5$	$87'0$	$4219'5$	$2352'25$	7569
$47'8$	$95'0$	4541	$2284'84$	9025
$47'3$	$88'6$	$4190'78$	$2237'29$	$7849'96$
$45'1$	$89'2$	$4022'92$	$2034'01$	$7956'64$
$46'3$	$78'9$	$3653'07$	$2143'69$	$6225'21$
$42'1$	$84'6$	$3561'66$	$1772'41$	$7157'16$
$44'2$	$81'7$	$3611'14$	$1953'64$	$6674'89$
$43'5$	$72'2$	$3140'7$	$1892'25$	$5212'84$
$42'3$	$65'1$	$2753'73$	$1789'29$	$4238'01$
$40'2$	$68'1$	$2737'62$	$1616'04$	$4637'61$
$31'8$	$67'3$	$2140'14$	$1011'24$	$4529'29$
$34'0$	$52'5$	1785	1156	$2756'25$
$713'5$	$1333'5$	$60568'34$	$32285'29$	$114535'33$

De ella obtenemos que es

$$\widehat{\beta} = \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{16 \cdot 60568'34 - 713'5 \cdot 1333'5}{16 \cdot 32285'29 - 713'5^2} = 2'3577$$

y

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \widehat{\beta}\sum_{i=1}^{n} x_i}{n} = \frac{1333'5 - 2'3577 \cdot 713'5}{16} = -21'795$$

con lo que la recta de mínimos cuadrados será

$$y = -21'795 + 2'3577\, x.$$

Para analizar la bondad del ajuste de mínimos cuadrados (CB-sección 2.4.3) que acabamos de realizar, calcularemos el *coeficiente de determinación* R^2.

Al ser el ajuste de una recta, podemos calcular R^2 a través de la fórmula

$$\begin{aligned} R^2 = (r)^2 &= \frac{(\widehat{\beta})^2 \left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right)}{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n} \\ &= 2'3577^2 \cdot \frac{32285'29 - \dfrac{713'5^2}{16}}{114535'33 - \dfrac{1333'5^2}{16}} = 0'76537 \end{aligned}$$

o como cuadrado del *coeficiente de correlación de Pearson*

$$\begin{aligned} r &= \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \\ &= \frac{16 \cdot 60568'34 - 713'5 \cdot 1333'5}{\sqrt{16 \cdot 32285'29 - 713'5^2}\,\sqrt{16 \cdot 114535'33 - 1333'5^2}} = 0'87485 \end{aligned}$$

siendo

$$R^2 = r^2 = 0'87485^2 = 0'76536.$$

Aunque dicho valor puede calificarse de aceptable, no será hasta que utilicemos el contraste de la regresión lineal simple, cuando podamos decidir si éste se califica de bueno o no.