Regresión

Se han visto algunos ejemplos donde parece que haya una relación aproximadamente lineal entre dos variables. Supongamos que queremos estimar la relación entre las dos variables. ¿Cómo ajustamos una recta a los datos?

Un modelo para representar una relación aproximadamente lineal es

$$y = \alpha + \beta x + \epsilon$$

donde ϵ es un error de predicción.

En esta formulación, y es la **variable dependente** cuya valor depende del valor de la **variable independiente** x.

Mínimos Cuadrados

Dada una muestra de datos $(x_1, y_1), \ldots, (x_n, y_n)$ queremos utilizar la recta que se ajusta mejor.

Si ajustamos una recta $y = a + bx$ a los datos de la muestra, entonces los **residuos** o errores de predicción son

$$r_i = y_i - (a + bx_i)$$

para $i = 1, \ldots, n$.

De alguna manera, la recta que se ajusta mejor es la que minimiza el error total. Pero ?cómo definimos el error total?

Usamos la suma de errores cuadrados $E(a, b) = \sum_{i=1}^{n} r_i^2$.

Teorema 6 *Para una muestra de datos bivariantes* $(x_1, y_1), \ldots, (x_n, y_n)$*, la recta de forma* $y = a + bx$ *que minimiza la suma de errores cuadrados* $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ *tiene*

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Demostración (sólo para matemáticos)

Supongamos que ajustamos la recta $y = a+bx$. Queremos minimizar el valor de $E(a, b)$. Recordamos que en el mínimo se tiene

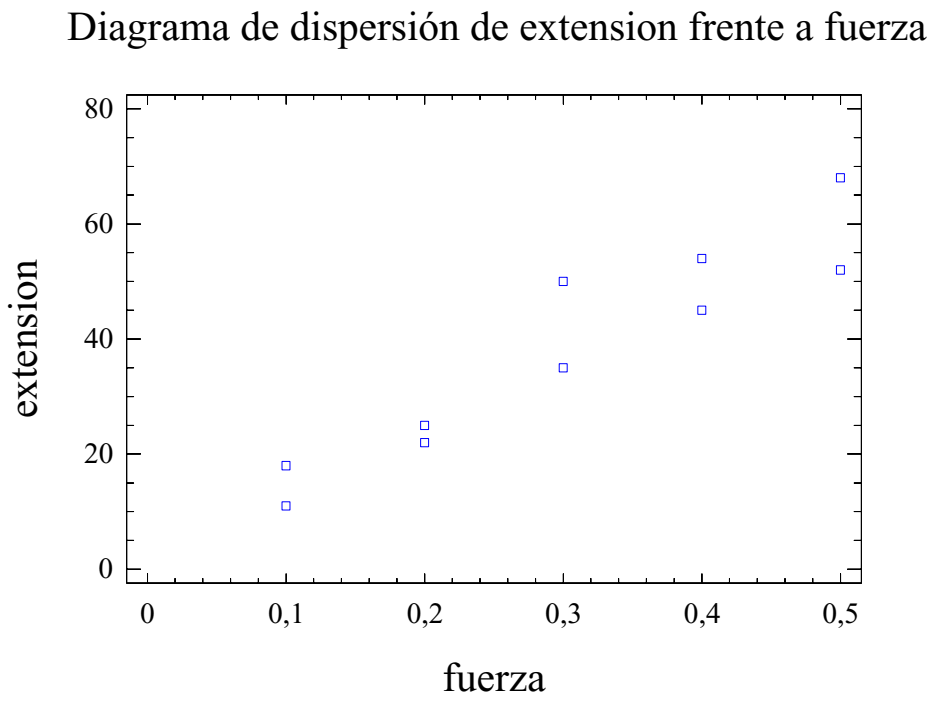
$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$$

Recordamos que $E = \sum_{i=1}^{n}(y_i - a - bx_i)^2$. Luego

$$
\begin{aligned}
\frac{\partial E}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) \quad \text{y al mínimo se tiene} \\
0 &= -2\sum_{i=1}^{n}(y_i - a - bx_i) \\
&= -2\left(n\bar{y} - na - nb\bar{x}\right) \\
a &= \bar{y} - b\bar{x} \\
\frac{\partial E}{\partial b} &= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) \quad \text{y al mínimo,} \\
0 &= -2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i(a + bx_i)\right) \\
\sum_{i=1}^{n} x_i y_i &= \sum_{i=1}^{n} x_i(a + bx_i) \\
&= \sum_{i=1}^{n} x_i(\bar{y} - b\bar{x} + bx_i) \quad \text{sustituyendo por } a \\
&= n\bar{x}\bar{y} + b\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\
b &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{n s_{xy}}{n s_x^2} = \frac{s_{xy}}{s_x^2} \qquad \diamond
\end{aligned}
$$

Ejemplo 72 *Se quiere probar la elasticidad de un muelle. Con este objetivo, se sometió el muelle a varios niveles de fuerza (x Newtons) y se midió la extensión total del muelle (y mm) en cada caso.*

fuerza	0,1	0,1	0,2	0,2	0,3	0,3	0,4	0,4	0,5	0,5
extensión	18	11	25	22	35	50	54	45	52	68



El diagrama de dispersión sugiere que existe una relación casi lineal entre fuerza y extensión. Para predecir la extensión del muelle en torno de la fuerza aplicada, aplicamos el model de regresión

$$y = \alpha + \beta x + \epsilon$$

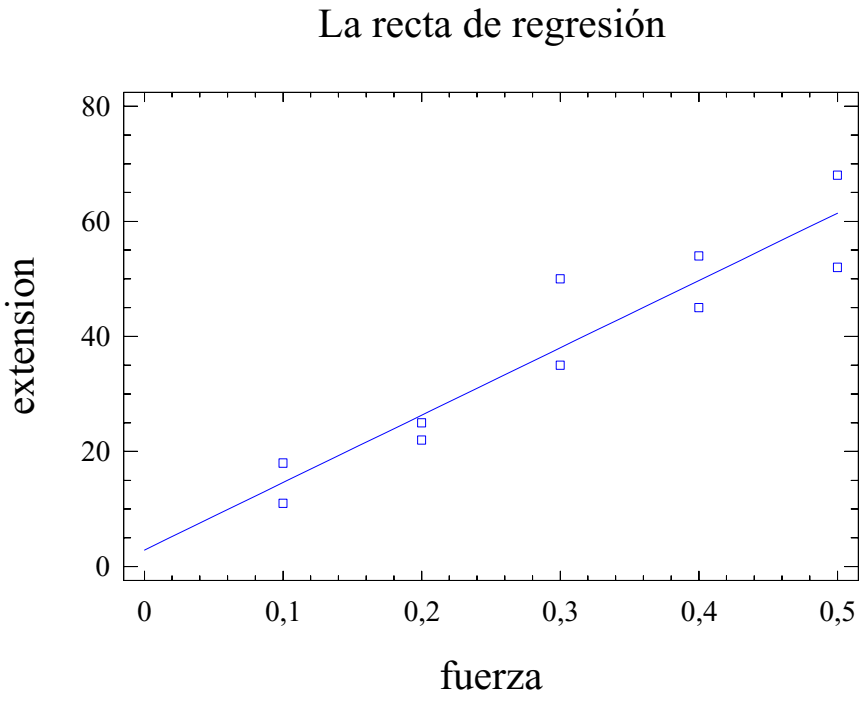
Dados los datos de la muestra, hallamos la recta estimada por mínimos cuadrados. Tenemos:

$$\begin{aligned} \bar{x} &= 0{,}3 \\ s_x^2 &= 0{,}02 \\ \bar{y} &= 38 \\ s_y^2 &= 310{,}8 \\ s_{xy} &= 2{,}34 \end{aligned}$$

Calculamos la recta de mínimos cuadrados.

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ &= \frac{2{,}34}{0{,}02} \\ &= 117 \\ a &= \bar{y} - b\bar{x} \\ &= 38 - 117 \times 0{,}3 \\ &= 2{,}9 \end{aligned}$$

La recta ajustada es $y = 2{,}9 + 117x$.



Ejemplo 73 *Calculamos la* **recta de regresión** *para los datos sobre diabéticos del Ejemplo 63.*

En el Ejemplo 64 demostramos que las medias de estos datos son $\bar{x} = 181{,}375$ *e* $\bar{y} = 18{,}125$ *y que la covarianza es* $s_{xy} = 361{,}64$*. En el Ejemplo 69, mostramos que* $s_x^2 = 1261{,}98$ *y* $s_y^2 = 211{,}23$*.*

Entonces, si queremos predecir los valores de y *(reducción de peso) en términos de* x *(peso original), la recta de regresión de mínimos cuadrados es*

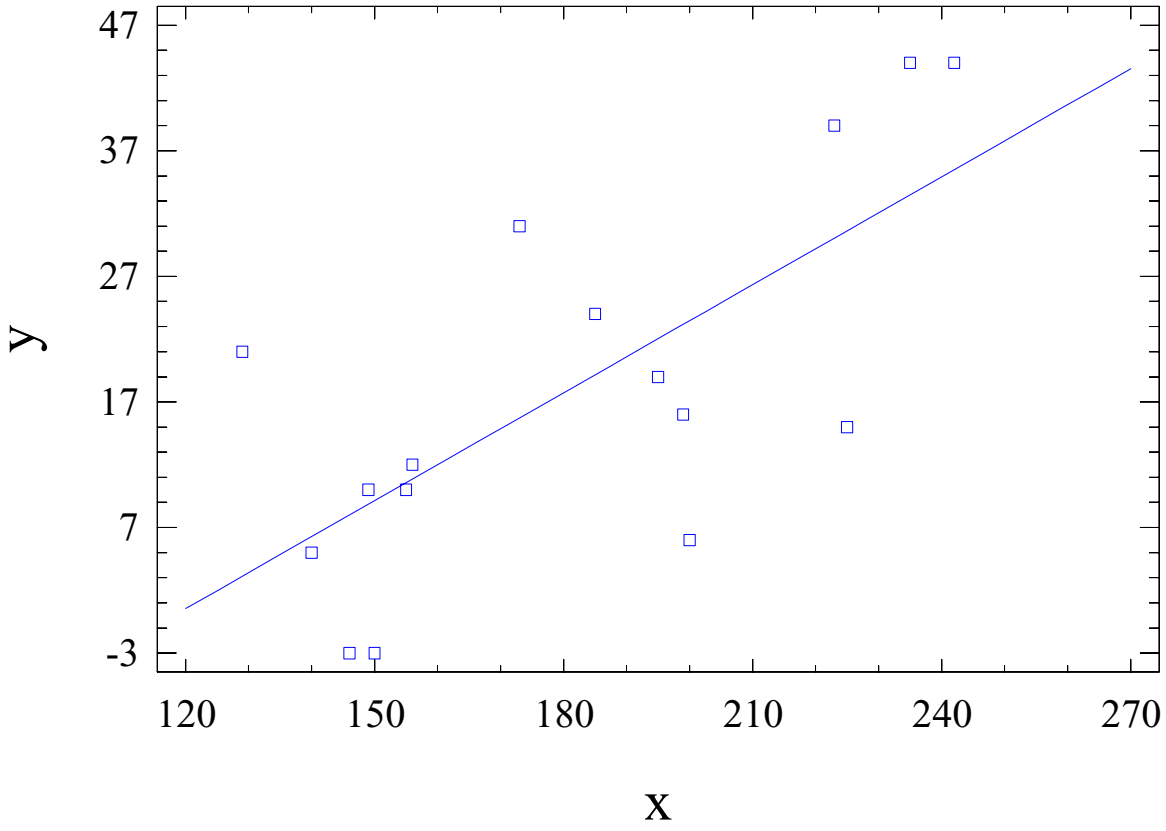
$$y = a + bx$$

donde

$$\begin{aligned} b &= \frac{361{,}64}{1261{,}98} \approx 0{,}287 \\ a &= 18{,}125 - 0{,}287 \times 181{,}375 \approx -33{,}85 \end{aligned}$$

Ajustamos la recta al diagrama de dispersión

Diagrama de dispersión con la recta de regresión añadida



Ejemplo 74 *Volvemos a los datos sobre el ácido úrico en la leche de vacas del Ejemplo 66.*

En el Ejemplo 66, demostramos que $\bar{x} = 29{,}56$, $\bar{y} = 167{,}43$ *y* $s_{xy} = -283{,}2$. *En el Ejemplo 68, sacamos que* $s_x^2 = 54{,}43$ *y* $s_y^2 = 1868{,}82$.

Luego, si queremos predecir la concentración de ácido úrico en la leche (y) *en términos de la cantidad de leche producida* (x), *la recta de mínimos cuadrados es*

$$y = a + bx$$

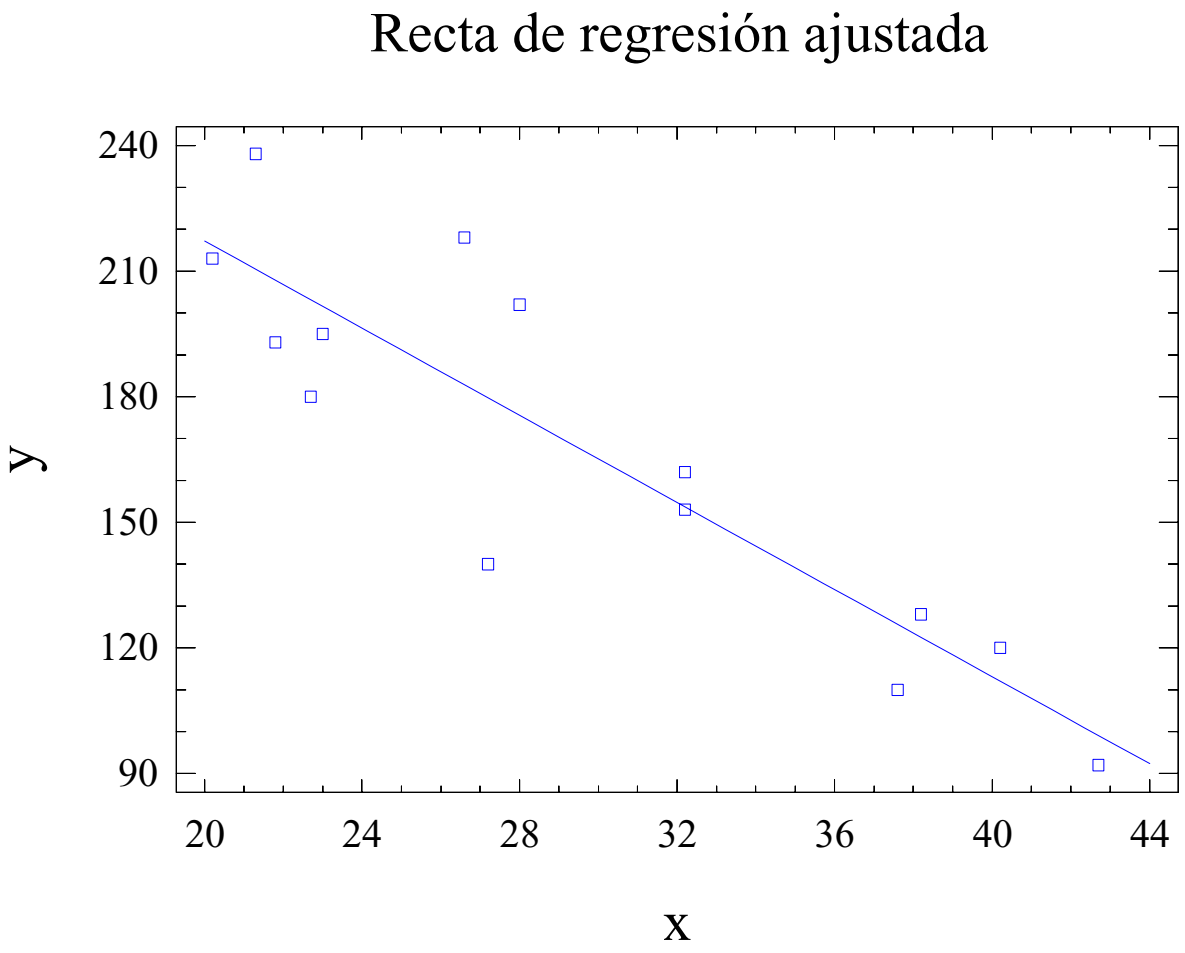
donde

$$\begin{aligned} b &= \frac{-283{,}2}{54{,}43} = -5{,}20 \\ a &= 167{,}43 - (-5{,}20) \times 29{,}56 \\ &= 321{,}24 \end{aligned}$$

Los resultados del análisis en Statgraphics son iguales

```
Regression Analysis - Linear model: Y = a + b*X
----------------------------------------------------------
Dependent variable: y Independent variable: x
----------------------------------------------------------

Parameter     Estimate
---------------------------
Intercept      321,241
Slope           -5,20265
---------------------------

Correlation Coefficient = -0,887889
R-squared = 78,8347 percent
Standard Error of Est. = 21,4817.

The equation of the fitted model is

   y = 321,241 - 5,20265*x
```

En la siguiente transparencia, se muestra la recta de regresión.



El ajuste de la recta de regresión parece bastante bueno.

Predicción

Habiendo ajustado una recta $y = a + bx$ a los datos, podremos usarla para predecir el valor de y teniendo el valor de x.

Ejemplo 75 *En el Ejemplo anterior, supongamos que una vaca produce* $x = 30$ *kilos de leche por día. ?Cuál estimamos es la concentración de ácido úrico en la leche de esta vaca?*

Estimamos con

$\widehat{y} = 321{,}24 - (-5{,}20) \times 30 \approx 165{,}15$ *μmol/litro*

Ejemplo 76 *Predecimos la pérdida de peso de un diabetico quien pesa unas* 220 *libras.*

La recta ajustada es $y = -33{,}85 + 0{,}287x$ *y entonces* $\widehat{y} = -33{,}85 + 0{,}287 \times 220 = 29{,}29$ *libras es la pérdida de peso previsto.*

Ejemplo 77 *En el Ejemplo 72, predecimos la extensión del muelle si se aplica una fuerza de* 0,4 *Newtons.*

Se tiene $\widehat{y} = 2{,}9 + 117 \times 0{,}4 = 49{,}7$ *mm es la extensión estimada.*

?Qué pasaría si ponemos una fuerza de 0*?*

La extensión prevista por la recta de regresión en este caso es de 2,9 *mm. No obstante el resultado no tiene sentido. Con fuerza* 0*, la extensión del muelle debe ser cero.*

Es muy peligroso hacer predicción usando valores de x *fuera del rango de los datos observados.*

La desviación típica residual

Los residuos o errores de la predicción son las diferencias $y_i - (a + bx_i)$. Es útil dar una idea de si el error típico es grande o pequeño. Por eso, se calcúla la d**esviación típica residual**.

Definición 20 *Dado que se ajusta la regresión por mínimos cuadrados con $y = a + bx$ con a, b definidos como anteriormente, se define la desviación típica residual como*

$$s_r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2}.$$

Ejemplo 78 *Calculamos los residuos en el Ejemplo 72*

	x	0,1	0,1	0,2	0,2	0,3	0,3	0,4	0,4	0,5	0,5
	y	18	11	25	22	35	50	54	45	52	68
$2{,}9 + 117x$		14,6	14,6	26,3	26,3	38,0	38,0	49,7	49,7	61,4	61,4
	r	3,4	−3,6	−1,3	−4,3	−3,0	12,0	4,3	−4,7	−9,4	6,6

Entonces

$$\begin{aligned} \bar{r} &= \frac{1}{10}(3{,}4 + \ldots + 6{,}6) \\ &= 0 \\ s_r^2 &= \frac{1}{10}\left(3{,}4^2 + \ldots + 6{,}6^2\right) \\ &= 37{,}2 \\ s_r &= \sqrt{37{,}02} \approx 6{,}08 \end{aligned}$$

Existe una manera más rápido de hacer el cálculo. En primer lugar observamos que $\bar{r} = 0$ siempre si ajustamos la recta de mínimos cuadrados.

Demostración

$$\begin{aligned}
\bar{r} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\bar{y} - b\bar{x} + bx_i)) \quad \text{por definición de } a \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y}) - b\sum_{i=1}^{n}(x_i - \bar{x})\right) \\
&= 0 \qquad \diamond
\end{aligned}$$

En segundo lugar, tenemos el siguiente resultado.

Teorema 7

$$s_r^2 = s_y^2\left(1 - r_{xy}^2\right)$$

donde r_{xy} es el coeficiente de correlación.

Demostración

$$
\begin{aligned}
s_r^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i-(a+bx_i)\right)^2 \\
&= \frac{1}{n}\left(y_i-(\bar{y}-b\bar{x}+bx_i)\right)^2 \quad \text{por definición de } a \\
&= \frac{1}{n}\left((y_i-\bar{y})-b(x_i-\bar{x})\right)\Big)^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2-2b\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+\right. \\
&\quad \left. b^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \\
&= s_y^2-2bs_{xy}+b^2s_x^2 \\
&= s_y^2-2\frac{s_{xy}}{s_x^2}s_{xy}+\left(\frac{s_{xy}}{s_x^2}\right)^2 s_x^2 \quad \text{por definición de } b \\
&= s_y^2-\frac{s_{xy}^2}{s_x^2} \\
&= s_y^2\left(1-\frac{s_{xy}^2}{s_x^2s_y^2}\right) \\
&= s_y^2\left(1-\left(\frac{s_{xy}}{s_xs_y}\right)^2\right)=s_y^2\left(1-r_{xy}^2\right) \qquad \diamond
\end{aligned}
$$

Ejemplo 79 *Volviendo al Ejemplo 72, recordamos que* $s_x^2 = 0{,}02$, $s_y^2 = 310{,}8$ *y* $s_{xy} = 2{,}34$. *Luego, la correlación es*

$$r_{xy} = \frac{2{,}34}{\sqrt{0{,}2 \times 310{,}8}} \approx 0{,}939.$$

Entonces $s_r^2 = 310{,}8\left(1 - 0{,}939^2\right) = 37{,}02$ *como calculamos anteriormente.*

Podemos interpretar el teorema de otra manera. Tenemos

$$\frac{s_r^2}{s_y^2} = 1 - r_{xy}^2.$$

Pensamos en s_y^2 como la varianza o error total en predecir los valores de la variable y sin saber los valores de x. s_r^2 es el error total si usamos la variable x para predecir y. El porcentaje de reducción de la varianza original debido a la regresión es $r_{xy}^2 \times 100\,\%$.

Ejemplo 80 *En el Ejemplo 79, se ve que la porcentaje de reducción en varianza debida al conocimiento de los valores de la fuerza es de* $0{,}939^2 \times 100 = 88{,}8\,\%$.

Ejemplo 81 *En el Ejemplo 74 se ve que el coeficiente de correlación es* $-0{,}88789$ *y que el valor de* `R-squared` *es de un* $78{,}8347\,\% = (-0{,}88789)^2 \times 100\,\%$.

Conociendo las cantidades de leche producido por las vacas, se reduce la varianza un $78{,}8347\,\%$.

En el libro de Peña y Romo (1997), se interpreta el resultado de la siguiente manera. Queremos predecir la variable y. Si sólo tenemos los datos $y_1, \ldots, y_n$ de la muestra, predecimos con $\bar{y}$ y el error promedio de predicción es (aproximadamente) s_y. No obstante, si conocemos el valor de la variable independiente x, predecimos con la recta de regresión y el error de predicción es s_r.

Entonces, $\frac{s_r}{s_y} = \sqrt{1 - r_{xy}^2}$ y la reducción en variabilidad debida a la regresión es $(1 - \sqrt{1 - r_{xy}^2}) \times 100\,\%$.

Ejemplo 82 *Si* $r_{xy} = 0{,}8$ *entonces* $1 - r_{xy}^2 = 0{,}36$ *y* $\sqrt{1 - r_{xy}^2} = 0{,}6$. *Por lo tanto, la desviación típica residual es sólo un* 60 % *de la desviación típica original y se ha reducido la variabilidad por un* 40 %.

Otra conexión entre correlación y regresión

Consideramos la fórmula para el pendiente de la recta de regresión. Tenemos:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ &= \frac{s_y s_{xy}}{s_y s_x s_x} = \frac{s_y}{s_x}\frac{s_{xy}}{s_x s_y} \\ &= \frac{s_y}{s_x} r_{xy} \end{aligned}$$

Luego, se ve que si la correlación entre las dos variables es cero, también lo es la pendiente de la recta. Además, el Teorema 7 nos demuestra que la reducción en la varianza de los datos y debida a la regresión en este caso es 0.