En un conjunto de personas a las que pesamos y tallamos se comprueba que, aproximadamente, a más estatura corresponde mayor peso. Pero no hay ninguna función matemática que ligue esas dos variables. Se trata de una **relación estadística**.

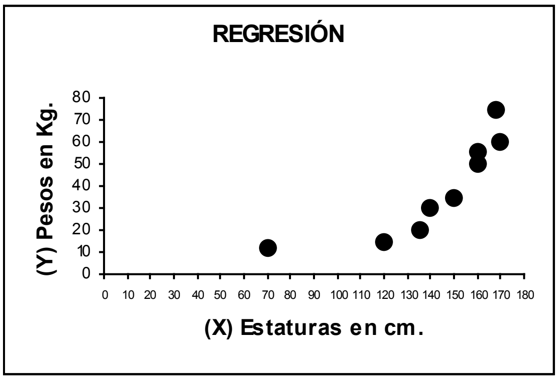
La **regresión** consiste en la búsqueda de una función que exprese lo mejor posible la relación existente entre dos (o más) variables estadísticas. Esto permitirá **predecir** el valor de una de las variables para un valor dado de la otra (u otras).

EJEMPLO

En la tabla siguiente se dan los pesos y estaturas de los nueve miembros de una familia:

Estatura (cm)	**X**	168	170	160	160	150	140	135	120	70
Peso (kg)	**Y**	75	60	55	50	35	30	20	15	12

La nube de puntos correspondiente a la variable bidimensional (X,Y) se ajusta bastante bien, aunque no exactamente, a una recta en las edades adultas. En cambio, el "ajuste" en los niños no es tan bueno.



La búsqueda de "buenas" rectas de ajuste es el objetivo de la **regresión lineal**.

Recta de regresión de Y sobre X

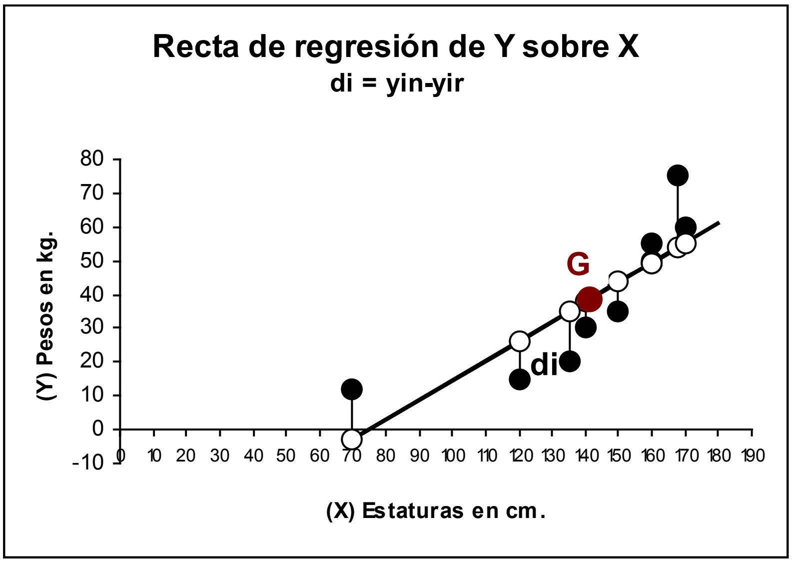
Se trata de hallar la ecuación de la recta **r** que mejor se ajuste a la nube de puntos en el siguiente sentido:

- **r** pasa por el punto $G(\overline{X},\overline{Y})$ (**centro de gravedad** de la distribución).

- La suma de los cuadrados de las diferencias de ordenadas entre los puntos-nube y los puntos-recta **r**, para cada valor x_i de X

$$S_{Y/X} = \Sigma d_i^{\,2} = \Sigma(y_{in} - y_{ir})^2$$

ha de ser mínima.



La **recta de regresión (mínimo-cuadrática) de Y sobre X** tiene por ecuación:

$$\mathbf{r_{Y/X}}: \quad y - \overline{Y} = \frac{\sigma_{XY}}{\sigma_X{}^2}(x - \overline{X})$$

Se emplea para predecir el valor de Y para un valor de X. Para un individuo que mida x_0 (cm), predice un peso de

$$\hat{y}_0 = \overline{Y} + \frac{\sigma_{XY}}{\sigma_X{}^2}(x_0 - \overline{X}) \quad \text{(kg)}$$

La *pendiente* de $\mathbf{r_{Y/X}}$, $\mathbf{m_{Y/X}} = \frac{\sigma_{XY}}{\sigma_X{}^2} = \frac{Cov(X,Y)}{Var(X)}$, se llama **coeficiente de regresión de Y sobre X**.

Recta de regresión de X sobre Y

Se trata de hallar la ecuación de la recta **r** que mejor se ajuste a la nube de puntos en el siguiente sentido:

- **r** pasa por el punto $G(\overline{X},\overline{Y})$ (**centro de gravedad** de la distribución).
- La suma de los cuadrados de las diferencias de abscisas entre los puntos-nube y los puntos-recta **r**, para cada valor y_i de Y

$$S_{X/Y} = \sum D_i{}^2 = \sum (x_{in} - x_{ir})^2$$

ha de ser mínima.