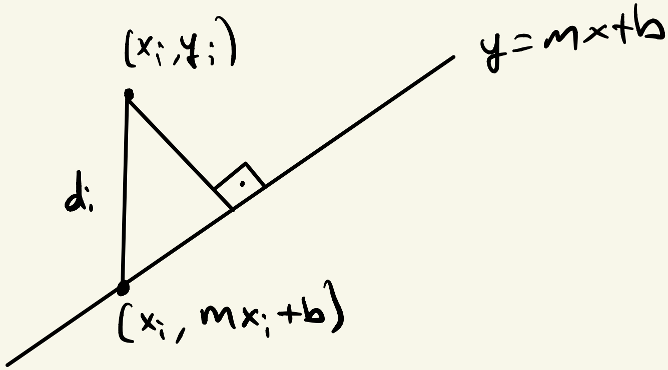
Linear Regression

Two points define a line, but how do you pass a line through more than two points? By minimizing the sum of squares of distances from the points to the line.



Instead of taking the regular distance from a point (x_i, y_i) to the line $y = mx + b$, we will consider the vertical distance d_i.

$$d_i = y_i - (mx_i + b)$$

For N given points, the sum of squares of vertical distances will be

$$\text{error} = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} \left[y_i - (mx_i + b) \right]^2$$

We are trying to find a line that will minimize this error, i.e., we are trying to minimize this expression with respect to m and b.

Call

$$f(m,b) = \text{error} = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} \left[y_i - (mx_i + b) \right]^2 .$$

If $\frac{\partial f}{\partial b} = 0$, then

$$0 = \frac{\partial f}{\partial b} = \sum_{i=1}^{N} 2 \left[y_i - (mx_i + b) \right]$$

$$\Rightarrow \sum_{i=1}^{N} (y_i - mx_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i - m \sum_{i=1}^{N} x_i - N \cdot b = 0$$

$$\Rightarrow N \cdot \bar{y} - m \cdot N \cdot \bar{x} - N \cdot b = 0$$

$$\Rightarrow \bar{y} - m\bar{x} - b = 0$$

$$\Rightarrow b = \bar{y} - m\bar{x} .$$

If $\frac{\partial f}{\partial m} = 0$, then

$$0 = \frac{\partial f}{\partial m} = \sum_{i=1}^{N} 2\{y_i - (mx_i+b)\}\cdot(-x_i)$$

$$\Rightarrow \sum_{i=1}^{N} \left(-x_iy_i + mx_i^2 + bx_i\right) = 0$$

$$\Rightarrow -\sum_{i=1}^{N} x_iy_i + m\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i = 0$$

Putting $b = \bar{y} - m\bar{x}$,

$$-\sum_{i=1}^{N} x_iy_i + m\sum_{i=1}^{N} x_i^2 + \bar{y}\sum_{i=1}^{N} x_i - m\bar{x}\sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow -\sum_{i=1}^{N} x_iy_i + m\sum_{i=1}^{N} x_i^2 + N\bar{x}\bar{y} - mN(\bar{x})^2 = 0$$

$$\Rightarrow m\left(\left(\sum_{i=1}^{N} x_i^2\right) - N(\bar{x})^2\right) = \left(\sum_{i=1}^{N} x_iy_i\right) - N\bar{x}\bar{y}$$

$$\Rightarrow m = \frac{\left(\sum_{i=1}^{N} x_iy_i\right) - N\bar{x}\bar{y}}{\left(\sum_{i=1}^{N} x_i^2\right) - N(\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{S_x^2}$$

$$(n-1)S_{xx} = \sum_{i=1}^{N}(x_i-\bar{x})^2 = \left(\sum_{i=1}^{N} x_i^2\right) - 2\bar{x}\underbrace{\left(\sum_{i=1}^{N} x_i\right)}_{N\cdot\bar{x}} + N(\bar{x})^2$$

$$= \left(\sum_{i=1}^{N} x_i^2\right) - N(\bar{x})^2$$

$$(n-1)S_{xy} = \sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \sum_{i=1}^{N}\left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}\right)$$

$$= \sum_{i=1}^{N} x_i y_i - \bar{y}\underbrace{\sum_{i=1}^{N} x_i}_{N\cdot\bar{x}} - \bar{x}\underbrace{\sum_{i=1}^{N} y_i}_{N\cdot\bar{y}} + N\bar{x}\bar{y}$$

$$= \left(\sum_{i=1}^{N} x_i y_i\right) - N\bar{x}\bar{y}$$

$$= \left(\sum_{i=1}^{N} x_i y_i\right) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)$$

$$= \left(\sum_{i=1}^{N} x_i y_i\right) - N\bar{x}\bar{y}$$

To sum up: given the points

$(x_1, y_1), \ldots, (x_N, y_N)$

the regression line is $y = mx + b$

where $b = \bar{y} - m\bar{x}$

and $m = \dfrac{S_{xy}}{S_x^2}$.

<u>Ex.:</u> Let's find the correlation coefficient and regression line of

$$\{(-3,1), (-2,0), (-1,0), (0,-1), (1,-1), (2,-3)\}.$$

$\{x_i\} = \{-3,-2,-1,0,1,2\} \to \bar{x} = -\frac{1}{2}$

$\{y_i\} = \{1,0,0,-1,-1,-3\} \to \bar{y} = -\frac{2}{3}$

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$= \frac{\left(-\frac{5}{2}\right)^2 + \left(-\frac{3}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{5}{2}\right)^2}{5}$$

$$= \frac{25+9+1+1+9+25}{4\cdot 5} = \frac{70}{20} = \frac{7}{2} = 3.5$$

$$S_x = \sqrt{\frac{7}{2}} \approx 1.87$$

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

$$= \frac{\left(\frac{5}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{7}{3}\right)^2}{5}$$

$$= \frac{25+4+4+1+1+49}{9\cdot 5} = \frac{84}{45} = \frac{28}{15}$$

$$S_y = \sqrt{\frac{28}{15}} \approx 1.37$$

$$S_{xy} = \frac{1}{5}\left(\left(-\frac{5}{2}\right)\cdot\left(\frac{5}{3}\right) + \left(-\frac{3}{2}\right)\cdot\left(\frac{2}{3}\right) + \left(-\frac{1}{2}\right)\cdot\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\cdot\left(-\frac{1}{3}\right) + \left(\frac{3}{2}\right)\cdot\left(-\frac{1}{3}\right) + \left(\frac{5}{2}\right)\cdot\left(-\frac{7}{3}\right) \right)$$

$$= -\frac{72}{5\cdot 6} = -\frac{72}{30} = -2.4$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-2.4}{\sqrt{\frac{7}{2} \cdot \frac{28}{15}}} \approx -0.94.$$

$$m = \frac{s_{xy}}{s_x^2} = \frac{-2.4}{3.5} \approx -0.6857$$

and $b = \bar{y} - m\bar{x} = -\frac{2}{3} - \left(-\frac{24}{35}\right) \cdot \left(-\frac{1}{2}\right)$

$$= -\frac{2}{3} - \frac{12}{35} = -\frac{106}{105} \approx -1.0095$$

$\Rightarrow$ regression line is

$$y = -0.6857x - 1.0095$$

<u>Observation:</u> Let $m, n \in \mathbb{R}$, with $m \neq 0$.

If $y_i = m x_i + n$ for all i, then

$$r = \begin{cases} 1 & \text{if } m > 0 \\ -1 & \text{if } m < 0 \end{cases}$$

<u>Proof:</u>

$$r = \frac{1}{N-1} \frac{1}{s_x} \frac{1}{s_y} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{N-1} \frac{1}{s_x} \frac{1}{|m| s_x} \sum_{i=1}^{N} (x_i - \bar{x}) \big((m x_i + n) - (m\bar{x} + n) \big)$$

$$= \frac{1}{N-1} \frac{1}{s_x} \frac{1}{|m| s_x} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot m \cdot (x_i - \bar{x})$$

$$= \frac{1}{N-1} \frac{1}{s_x^2} \frac{1}{|m|} m \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{1}{N-1} \frac{1}{s_x^2} \frac{m}{|m|} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{m}{|m|}$$

$$= \begin{cases} 1 & \text{if } m > 0 \\ -1 & \text{if } m < 0 \end{cases}$$

$\square$