Tutorial · #11

№ 22.5.1–5

Regression analysis:

We measure a random variable Y at fixed pts $X = [x_1, \dots, x_n]$.

Suppose we have just 1 sample:

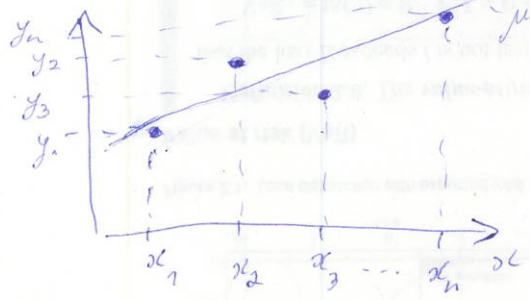
values $[y_1, \dots, y_n]$ measured at $[x_1, \dots, x_n]$ pts.

In linear regression, we assume mean value of Y

μ depends on x as $\mu(x) = a + bx$

– regression line – regression coefficient

For 1 sample, we have:



roughly speaking, we want to pass a line exactly "in between" all measured values $[y_1, \dots, y_n]$.

That is, we want to find a, b s.t. the deviation of $y_1, \dots, y_n$ from regression line $\mu(x)$ is minimal.

$$S(a,b) = \sum_{j=1}^{n} (y_j - a - b x_j)^2 \quad \text{should be minimized.}$$

$$\frac{\partial S}{\partial a} = 0 \Rightarrow -2 \cdot \sum_{j=1}^{n} (y_j - a - b x_j) = 0 \Rightarrow a = \underbrace{\frac{\sum_{j=1}^{n} y_j}{n}}_{=:\bar{y}} - b \cdot \underbrace{\frac{\sum_{j=1}^{n} x_i}{n}}_{=:\bar{x}}$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow -2 \cdot \sum_{j=1}^{n} x_j (y_j - a - b x_j) = 0 \Rightarrow b = \frac{\sum x_j y_j - a \sum x_j}{\sum x_j^2}$$

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum_{j=1}^{n} x_j y_j - a n \bar{x}}{\sum_{j=1}^{n} x_j^2} = \dfrac{\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}}{\sum_{j=1}^{n} x_j^2} + \dfrac{b n \bar{x}^2}{\sum_{j=1}^{n} x_j^2} \end{cases}$$

$$\Rightarrow \quad b = \frac{\sum_{j=1}^{n} x_j y_j - n \cdot \bar{x} \cdot \bar{y}}{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}, \qquad a = \bar{y} - b\bar{x}$$

It's conventional to rewrite this in terms of "variance / covariance" (even though x is not a random variable): $s_x^2 := \frac{1}{n-1} \cdot \sum_{j=1}^{n} (x_j - \bar{x})^2$

$$s_{xy} := \frac{1}{n-1} \cdot \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})$$

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{j=1}^{n} x_j^2 - \frac{2\bar{x}}{n-1} \cdot \underbrace{\sum_{j=1}^{n} x_j}_{= n\bar{x}} + \frac{n \cdot \bar{x}^2}{n-1} = \frac{1}{n-1} \cdot \sum_{j=1}^{n} x_j^2 - \frac{n\bar{x}^2}{n-1}$$

$$s_{xy} = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j y_j - \bar{x} \cdot \underbrace{\sum_{j=1}^{n} y_j}_{= n\bar{y}} - \bar{y} \cdot \underbrace{\sum_{j=1}^{n} x_j}_{= n\bar{x}} + n\bar{x}\bar{y} \right] =$$

$$= \frac{1}{n-1} \cdot \sum_{j=1}^{n} x_j y_j - n \cdot \frac{\bar{x}\bar{y}}{n-1}$$

$$\Rightarrow \begin{cases} \sum_{j=1}^{n} x_j^2 - n\bar{x}^2 = (n-1) \cdot s_x^2 \\ \sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y} = (n-1) \cdot s_{xy} \end{cases} \Rightarrow \quad b = \frac{s_{xy}}{s_x^2}$$

Also, introduce $S_y^2 := \frac{1}{n-1} \cdot \sum_{j=1}^{n} (y_j - \bar{y})^2$.

{ Similarly as for S_x^2, we have

$$\sum_{j=1}^{n} y_j^2 - n\bar{y}^2 = (n-1) S_y^2$$

} – don't need, in fact

Then, note that we can write

$$S(a,b) = \sum_{j=1}^{n} (y_j - a - b x_j)^2 = \sum_{j=1}^{n} \left[y_j - \bar{y} - b(x_j - \bar{x}) \right]^2 =$$

$$= \sum_{j=1}^{n} (y_j - \bar{y})^2 + b^2 \cdot \sum_{j=1}^{n} (x_j - \bar{x})^2 - 2b \cdot \sum_{j=1}^{n} (y_j - \bar{y}) \cdot (x_j - \bar{x}) =$$

$$= (n-1) \cdot \left[S_y^2 + b^2 S_x^2 - 2b S_{xy} \right] = (n-1) \cdot \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right) \geq 0$$

($b^2 S_x^2 = \frac{S_{xy}^2}{S_x^2}$, $2b S_{xy} = \frac{S_{xy}^2}{S_x^2}$)

$$\Leftrightarrow \quad S_y^2 \geq \frac{S_{xy}^2}{S_x^2} \quad \Leftrightarrow \quad \frac{|S_{xy}|}{|S_x| |S_y|} \leq 1 .$$

Moreover, $S_y^2 = \frac{S_{xy}^2}{S_x^2}$ $\Leftrightarrow$ $S(a,b) = 0$,

i.e. all the sample points lie on one line

$\Uparrow$

$$|S_{xy}| = |S_x| \cdot |S_y|$$

Correlation:

Now, let both X and Y be random variables.

Similarly to b, define correlation coefficient of the sample:

$$r = \frac{S_{xy}}{S_x S_y} \qquad (S_x > 0, \; S_y > 0)$$

This gives an estimate for correlation function of a population:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

σ_{xy}, covariance $= E[(X-\mu_x)(Y-\mu_y)]$

σ_x, variances $= E[(X-\mu_x)^2]$

$\sigma_y = E[(Y-\mu_y)^2]$

(this measures linear(!) dependence of X and Y

Note that S_{xy}, S_x, S_y introduced above are exactly unbiased estimates of σ_{xy}, σ_x, σ_y, respectively.

$0 \le |\rho| \le 1 \quad \Leftrightarrow \quad 0 \le |r| \le 1$

$|\rho| = 1$: exact linear dependence: deviation is zero

$\rho = 0$: no linear dependence: $(X-\mu_x)(Y-\mu_y)$ has positive & negative value equally likely

Compare:

In case of regression, $S_{xy} = 0 \Leftrightarrow b = 0$ implies Y being constant (i.e. independent) in X.