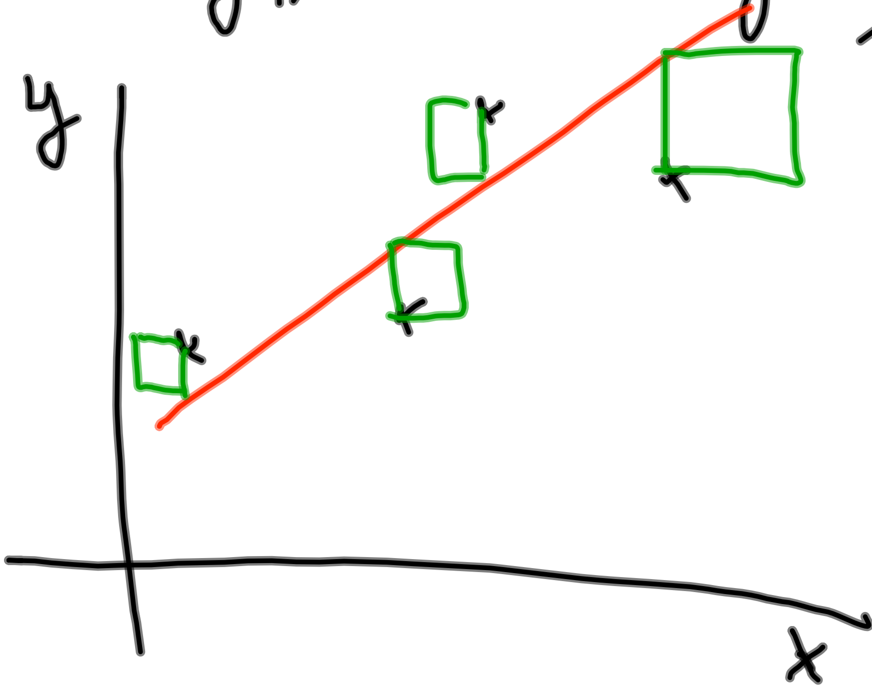
KORELACE A REGRESE

$(x_1, y_1), \dots, (x_n, y_n)$



$$y - \bar{y} = \frac{k}{s_x^2}(x - \bar{x}),$$

1. regresní přímka

$$k = \frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{k}{s_x \cdot s_y}$$ výběrový korelační koeficient

$$|r| = 1$$

Kovariance a korelační koeficient

$(\Omega, \mathcal{S}, P)$

nechť X, Y jsou náhodné prom.

Víme, že

$$E(a+bX) = a + bE(X), \quad a, b \in R$$

$$E(X+Y) = E(X) + E(Y)$$

$$D(a+bX) = b^2 \cdot D(X)$$

$$D(X+Y) = D(X) + D(Y) + 2\underbrace{E\left[(X-E(X))(Y-E(Y))\right]}_{cov(X,Y)}$$

Pom. $\boxed{cov(X,Y)} =$

$$= E(XY - YE(X) - XE(Y) + E(X)\cdot E(Y)) = E(XY) - E(X)\cdot E(Y)$$

Průměr $E(XY)$ je

a) pro diskrétní X, Y

$$E(XY)=\sum_i\sum_j x_i\, y_j\, p(x_i, y_j)$$

b) pro spojité X, Y

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, y\, f(x, y)\, dx\, dy$$

<u>Věta</u>. Jsou-li X, Y nezávislé náhodné veličiny, je
$$\mathrm{cov}(X,Y)=0.$$
(Říkáme, že X, Y jsou nekorelované.)

Pozn.: Nezávislost $X, Y \Rightarrow$
$$E(XY)=E(X)\cdot E(Y)$$
Obráceně: Je-li $\mathrm{cov}(X,Y)\neq 0$, pak jsou X, Y závislé.

$\mathrm{cov}(X,Y)$ jako míra závislosti X, Y?

Když zdvojnásobím $X \to 2X$, $Y \to 2Y$, tak

$$\mathrm{cov}(2X, 2Y) = 8\,\mathrm{cov}(X,Y)$$

Ještě ne; je třeba „normovat".

$$\mathrm{cov}\left(\frac{X-E(X)}{\sigma_X}, \frac{Y-E(Y)}{\sigma_Y}\right) =$$

$$= \frac{E\left[(X-E(X))\cdot(Y-E(Y))\right]}{\sigma_X \cdot \sigma_Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Def. $\dfrac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \rho(X,Y)$

je korelační koeficient X, Y -

Věta. 1. Pro nezávislé náh. prom. X, Y na $(\Omega, \mathcal{S}, P)$ je $\rho(X,Y) = 0$

2. $-1 \leq \rho(X,Y) \leq 1$

3. Je-li $Y = a + bX$, pak je

$$\rho(X,Y) = 1 \ (b>0)$$
$$= -1 \ (b<0)$$

Závěr: korelační koeficient je ukazatel lineární závislosti.

Platí: $$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=S_{XY}$$

je nevychýlený odhad $\operatorname{cov}(X,Y)$.

Odhad korelačního koeficientu je $$R=\frac{S_{XY}}{\sqrt{S_X^2\cdot S_Y^2}}$$ (výběrový korelační koef.)

Realizace korelačního koef. je

$$\boxed{r = \frac{\sum_{1}^{n} x_i \cdot y_i - n \cdot \bar{x}\bar{y}}{\sqrt{\sum (x_i^2 - n \cdot \bar{x}^2) \sum (y_i^2 - n\bar{y}^2)}}}$$

$$(s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 ;$$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2 \underbrace{\left(\sum x_i\right)}_{n \cdot \bar{x}} \bar{x} + \bar{x}^2 \cdot n$$

Co dál? Kdyby r bylo „malé", testuje se hypotéza $\rho(X,Y)=0$

Platí: $\dfrac{R}{\sqrt{1-R^2}}\cdot\sqrt{n-2}$ má Studentovo rozdělení s $n-2$ stupni volnosti

Lineární regrese

Z metody nejmenšího součtu čtverců odchylek máme rovnici regresní přímky

$$y - E(Y) = \frac{\operatorname{cov}(X,Y)}{D(X)}\left(x - E(X)\right)$$

spočítáme

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Postupy:

1.) Jiné regresní funkce $y = f(x)$

$$s_y^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$$

$$\hat{y}_i = f(x_i)$$

$$s_{\hat{y}}^2 = \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2$$

Vezmeme-li podíl

$$\frac{s^2_{\hat{y}}}{s^2_y} = l^2_{yx} \text{ (index determinace)}$$

l_{yx} – index korelace

2. Vícenásobná regrese a korelace

$$y = a + b_1 x_1 + b_2 x_2$$

$\ldots$