

# Přednáška č. 4

## NÁHODNÉ VEKTORY A REGRESNÍ ANALÝZA

11. 3. 2021

P. Pecherková

# NÁHODNÝ VEKTOR

**Náhodný vektor** je spojení dvou (nebo více) náhodných veličin.

- V našem kurzu budeme uvažovat pouze dvě náhodné veličiny (ilustrační příklad)
- Náhodné veličiny mohou/nemusí mít mezi sebou vztah = závislost / nezávislost

**Nezávislost**

- Výhodné zejména, pokud chceme mít zajištěno, že naměřená data se neovlivňují navzájem

**Závislost**

- Lze hledat míru závislosti = korelace
- Lze zjistit vztah (předpis), jaký je mezi proměnnými = regresní analýza



$$f(x, y) = f(x|y)f(y)$$

$$f(x, y) = f(y|x)f(x)$$

$$f(x|y) = \frac{f(x, y)}{f(y)} \quad f(y|x) = \frac{f(x, y)}{f(x)}$$

$$f(y) = \frac{f(x, y)}{f(x|y)} \quad f(x) = \frac{f(x, y)}{f(y|x)}$$

Sledujeme chování lidí na přechodu – zda lidé v různé věkové kategorii (děti, dospělí, senioři) přecházejí na červenou či na zelenou.

### Sdružené rozdělení:

- Informace o současném chování veličiny  $x$  a  $y$
- S jakou pravděpodobností přechází dítě na červenou/zelenou, dospělý na červenou/zelenou...

### Podmíněné rozdělení:

- chování veličiny  $x$  za podmínky, že známe jaké bylo chování veličiny  $y$ .
  - S jakou pravděpodobností přejde dítě/dospělý/senior, když na semaforu svítí červená
- chování veličiny  $y$  za podmínky, že známe jaké bylo chování veličiny  $x$ .
  - S jakou pravděpodobností bude svítit červená/zelená, když přechází senior

### Marginální rozdělení:

- poskytuje informaci o veličině  $y$  bez vlivu veličiny  $x$ 
  - s jakou pravděpodobností je člověk dítě/dospělý, senior
- poskytuje informaci o veličině  $x$  bez vlivu veličiny  $y$ 
  - S jakou pravděpodobností svítí na semaforu červená/zelená.

Sledujeme chování lidí na přechodu – zda lidé v různé věkové kategorii (děti, dospělí, senioři) přecházejí na červenou či na zelenou.

| $f(x, y)$                                                                                 |  |  |  | $f(x)$  |
|-------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| zelená   | 0,10                                                                              | 0,30                                                                              | 0,20                                                                              | 0,6                                                                                        |
| červená  | 0,05                                                                              | 0,25                                                                              | 0,10                                                                              | 0,4                                                                                        |
| $f(y)$   | 0,15                                                                              | 0,55                                                                              | 0,3                                                                               |                                                                                            |

Víme, co svítilo na semaforu:

| $f(x y)$ |  |  |  |
|----------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| zelená   | $\frac{1}{6}$                                                                     | $\frac{3}{6}$                                                                     | $\frac{2}{6}$                                                                     |
| červená  | $\frac{1}{8}$                                                                     | $\frac{5}{8}$                                                                     | $\frac{2}{8}$                                                                     |



Víme, kdo přechází:

| $f(y x)$ |  |  |  |
|----------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| zelená   | $\frac{2}{3}$                                                                       | $\frac{6}{11}$                                                                      | $\frac{2}{3}$                                                                       |
| červená  | $\frac{1}{3}$                                                                       | $\frac{5}{11}$                                                                      | $\frac{1}{3}$                                                                       |



Sledujeme závislost rychlosti vozidla nad povolenou rychlost  $x$  na množství alkoholu v krvi řidiče  $y$ .

$$f_{XY}(x, y) = \frac{1}{16}xy \quad \text{pro } x \in (0; 40), y \in (0; 0,2)$$

Pravděpodobnost pod celou plochou musí být = 1

$$\int_0^{40} \int_0^{0,2} \frac{1}{16}xy \, dy \, dx = \int_0^{40} \frac{1}{16}x \left[ \frac{y^2}{2} \right]_0^{0,2} = \int_0^{40} \frac{1}{800}x \, dx = 1$$

Marginální hp

$$f(x) = \int_0^{0,2} \frac{1}{16}xy \, dy = \left[ \frac{1}{16}x \frac{y^2}{2} \right]_0^{0,2} = \frac{1}{800}x$$

$$f(y) = \int_0^{40} \frac{1}{16}xy \, dx = \left[ \frac{1}{16} \frac{x^2}{2} y \right]_0^{40} = 50y$$

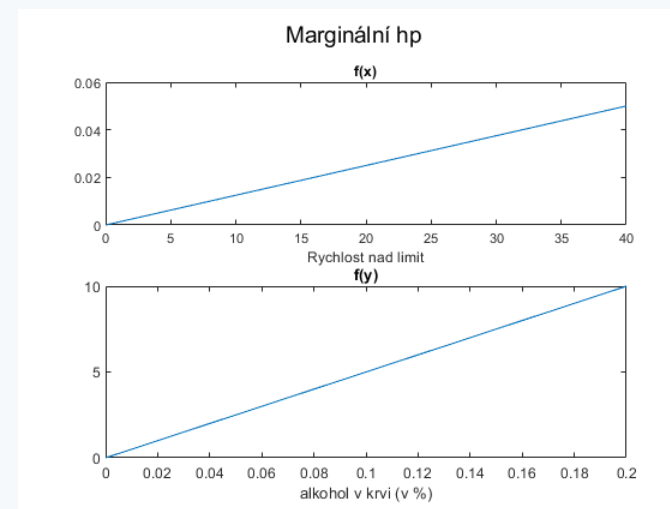
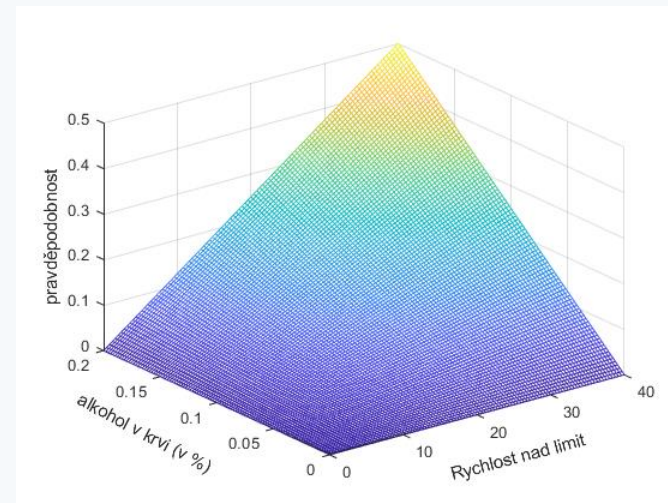
Podmíněná hp

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{\frac{1}{16}xy}{50y} = \frac{1}{800}x$$

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{\frac{1}{16}xy}{\frac{1}{800}x} = 50y$$

Jsou  $x$  a  $y$  nezávislé?

$$f(x)f(y) = \frac{1}{800}x \cdot 50y = \frac{1}{16}xy = f(x, y)$$



# KOVARIANCE

- Charakteristika dvou náhodných veličin  $x$  a  $y$ , která vypovídá o jejich vazbě / závislosti
  - Lze určit závislost veličin
  - Nelze určit míru závislosti veličin

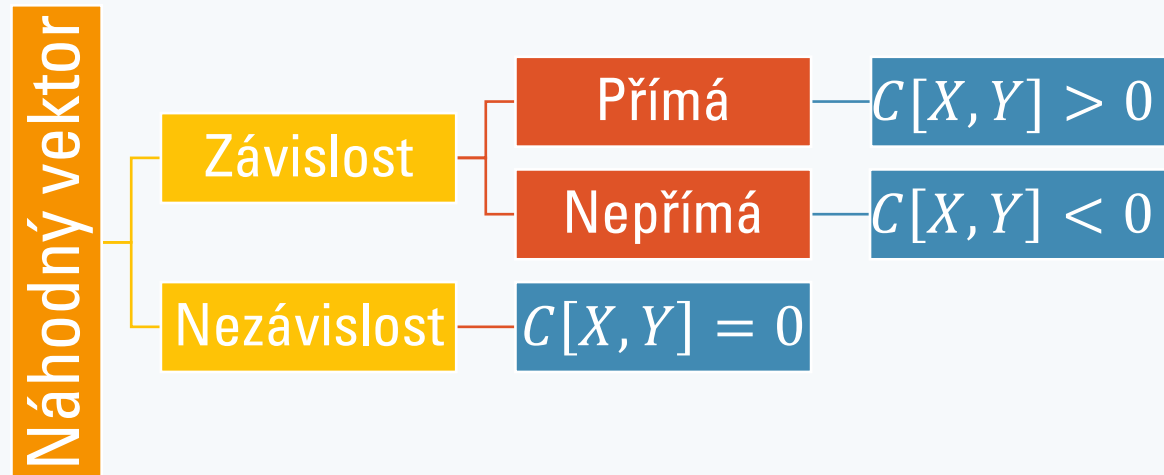
Kovarianční koeficient

$$C[X, Y] = \sum_i \sum_j (x_i - E[X])(y_j - E[Y])f(x_i, y_j)$$

$$C[X, Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y])f(x, y) dx dy$$

Kovarianční matice

$$\text{cov}[X, Y] = \begin{bmatrix} D[X] & C[X, Y] \\ C[X, Y] & D[Y] \end{bmatrix}$$

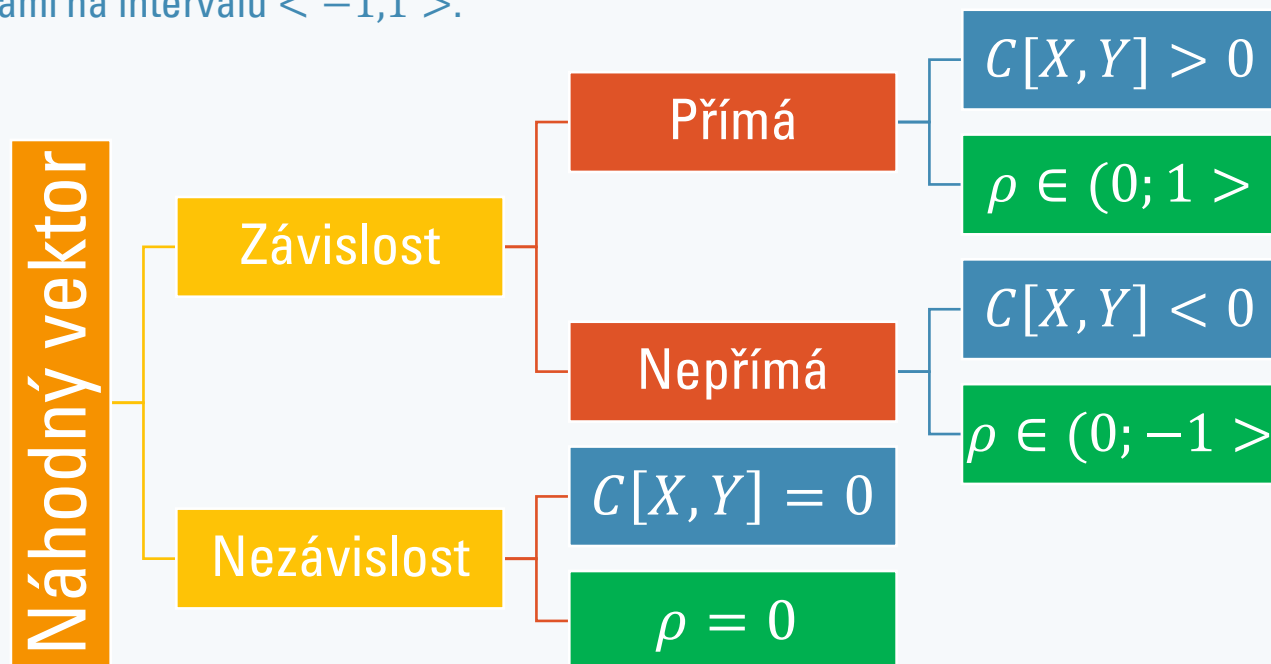


# KORELACE

Korelace vyjadřuje **míru** závislosti mezi veličinami na intervalu  $\langle -1, 1 \rangle$ .

Korelační koeficient:

- $\rho^2 = 1$  absolutní korelace
- $\rho^2 \in \langle 0,8; 1 \rangle$  velmi silná korelace
- $\rho^2 \in \langle 0,6; 0,8 \rangle$  silná korelace
- $\rho^2 \in \langle 0,4; 0,6 \rangle$  střední korelace
- $\rho^2 \in \langle 0,2; 0,4 \rangle$  slabá korelace
- $\rho^2 \in \langle 0; 0,2 \rangle$  velmi slabá korelace
- $\rho^2 = 0$  žádná korelace



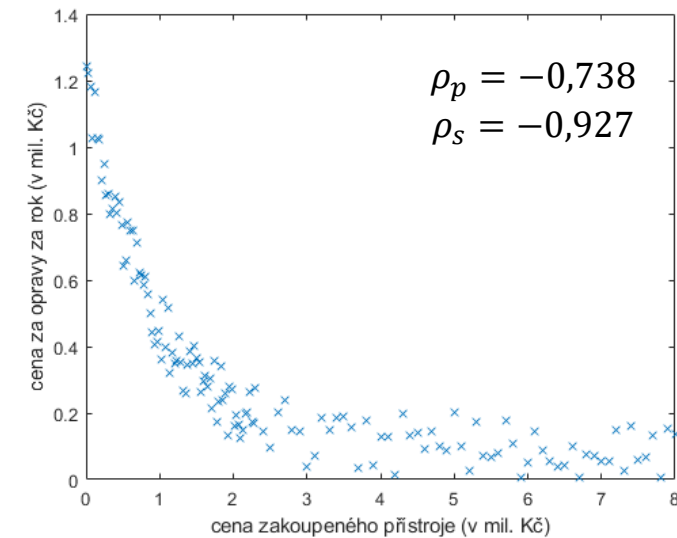
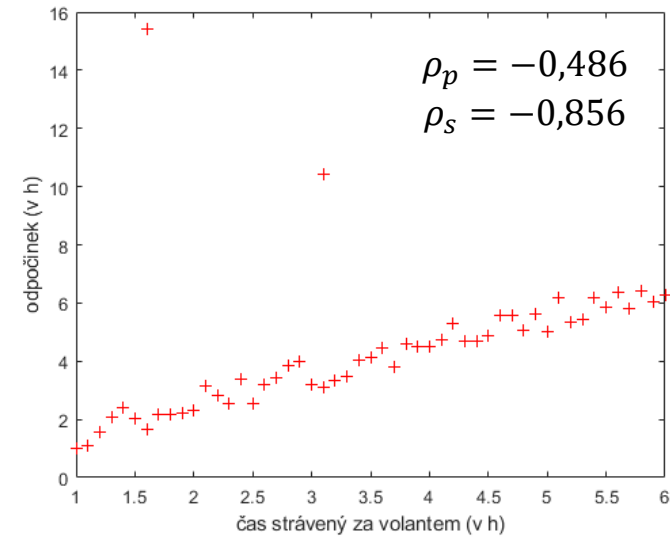
- Výpočet korelačního koeficientu:

- Pearsonův korelační koeficient

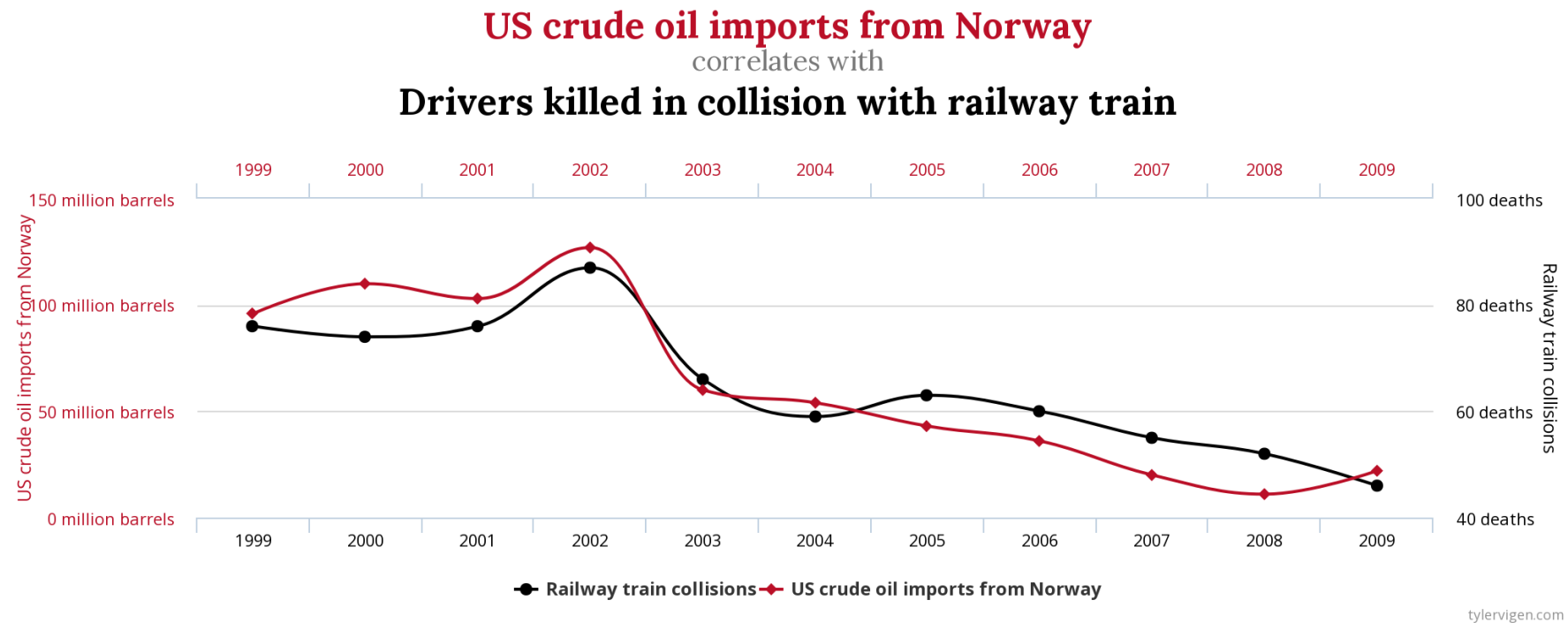
- $\rho_p = \frac{c[X,Y]}{\sigma_x \sigma_y}$
    - platí pouze pro lineární závislost
    - pracuje se skutečnou hodnotou

- Spearmanův korelační koeficient

- $\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$ 
      - kde  $d_i$  je rozdíl mezi dvěma pořadími stejného pozorování
    - platí pro lineární i nelineární závislost
    - nepracuje s hodnotami, ale s pořadím



# KAUZALITA

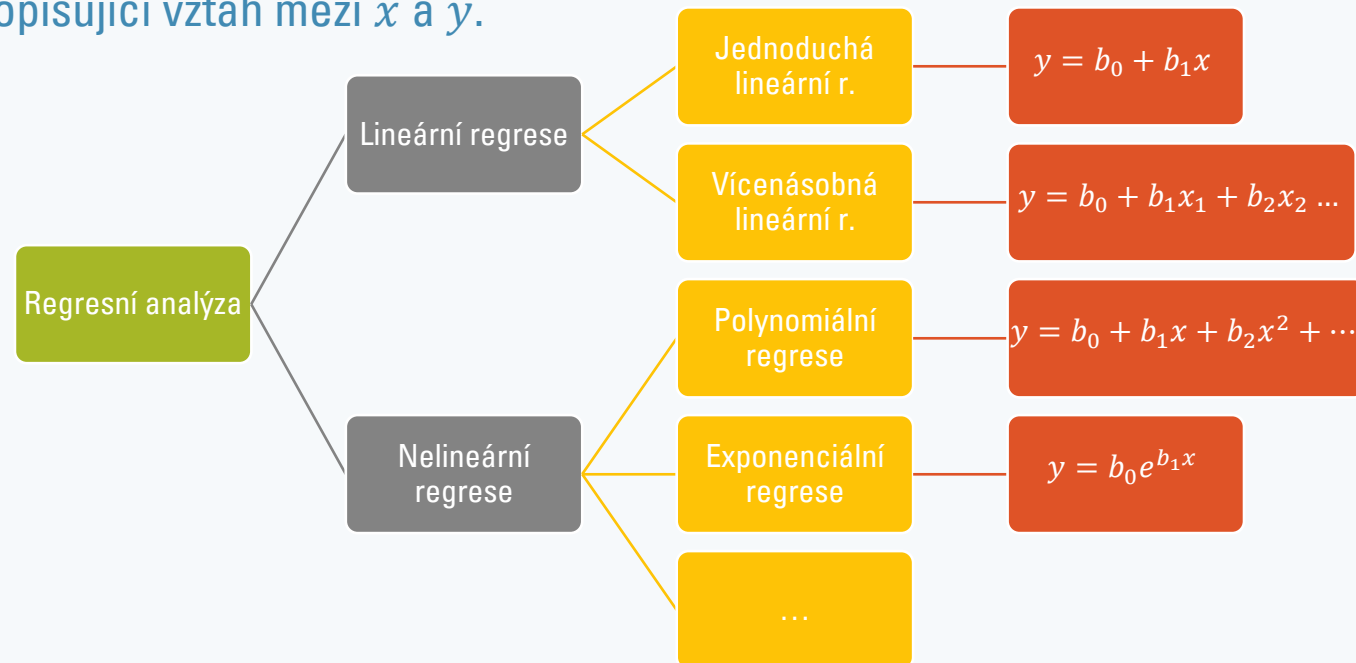


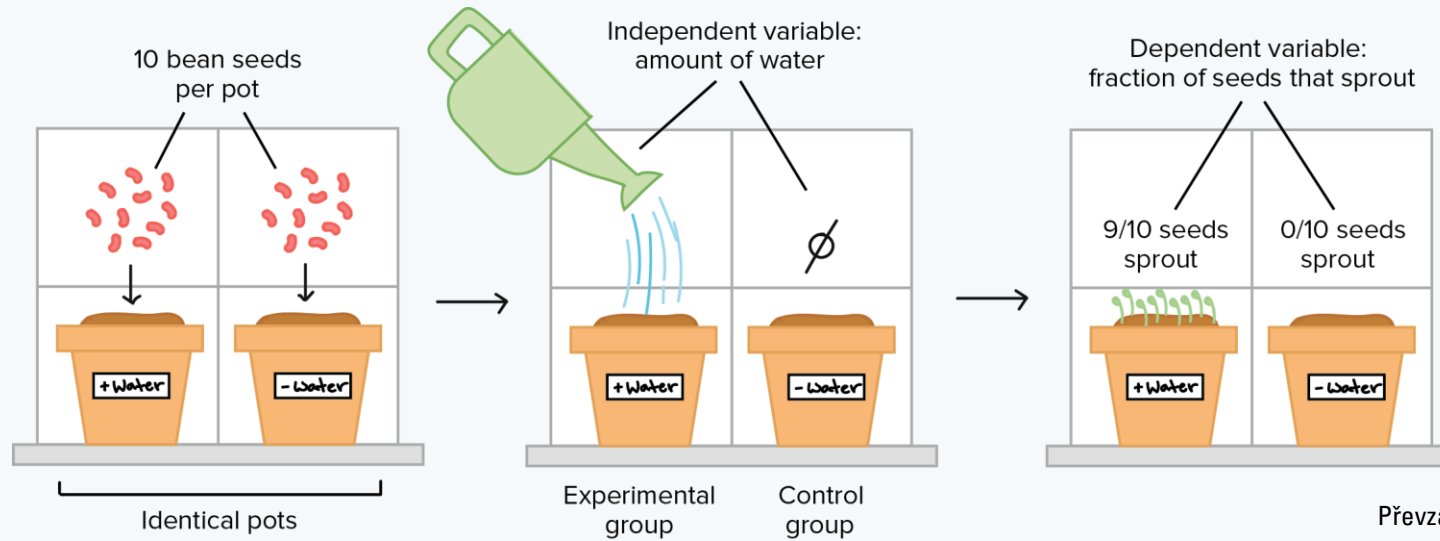
Korelace  $\neq$  kauzalita

# REGRESNÍ ANALÝZA

- Pokud máme data korelovaná **alespoň na úrovni střední korelace**, můžeme vytvořit tzv. regresní model, tedy vytvořit „rovnici“ popisující vztah mezi  $x$  a  $y$ .

Cílem je proložit naměřené body přímkou/křivkou danou předpisem tak, aby co nejlépe vystihovala trend dat.



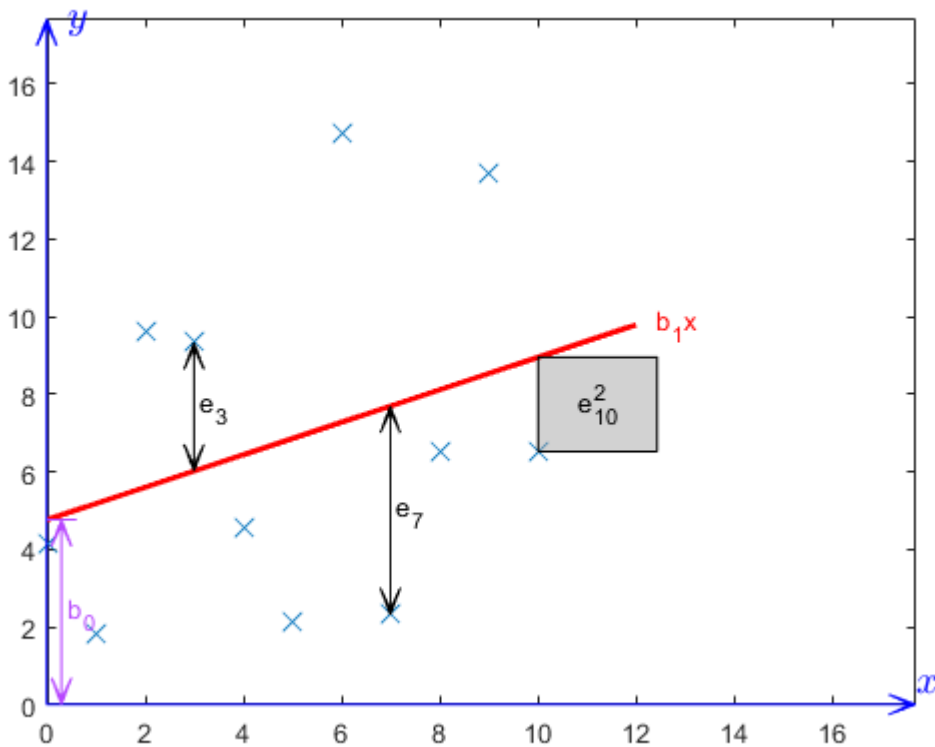


$x$  = nezávisle proměnná (independent variable)

$y$  = závisle proměnná (dependent variable)

Jak se určí, že zvolená přímka/ křivka je lepší než jiná? ... minimalizujeme součet reziduí, tedy:

$$\min \sum_i e_i^2$$



## • Jak najít tuto křivku?

### • Metoda nejmenších čtverců

$$\hat{b}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

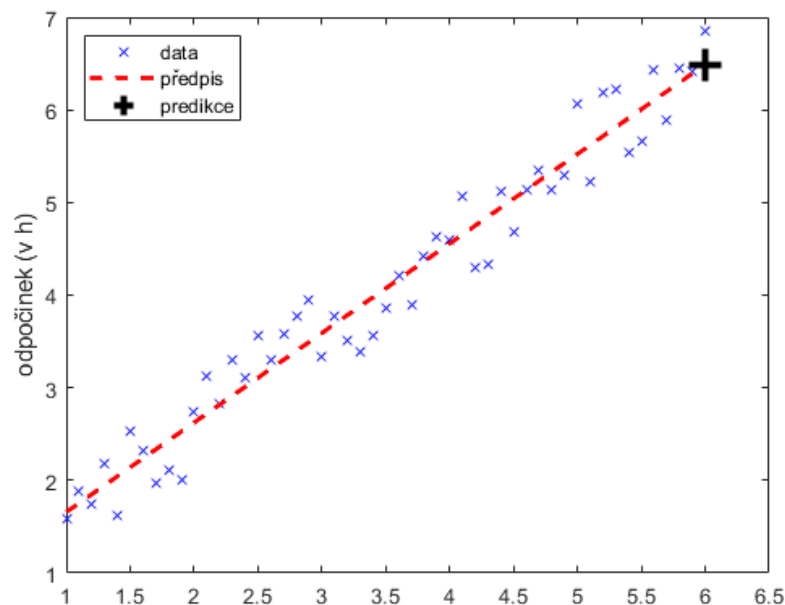
$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

### • Metoda nejmenších čtverců v maticovém tvaru

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_Y = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_X \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}_b + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_E$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}_b = (X'X)^{-1}X'Y$$

# Lineární regrese



- LR je metoda proložení naměřených hodnot přímkou

- Předpis:

$$y = b_0 + b_1x$$

- Mohou nastat 2 situace:

1. Hledáme regresní koeficienty –  $b_0$  a  $b_1$  (známe  $x$  a  $y$ )

- Použijeme metodu nejmenších čtverců

2. Hledáme očekávaný výstup (predikci) -  $y$  (známe  $b_0$ ,  $b_1$  a  $x$ )

- Příklad:

1. Hledáme předpis (vztah), kterým bychom popsali vliv doby strávené za volantem a dobou následného odpočinku
2. Známe předpis (vztah) pro vliv ceny benzínu a počet ujetých km a chceme zjistit, kolik bude odpočívat řidič, při 6 h jízdě

# Vícenásobná lineární regrese

```
Linear regression model:  
FINAL ~ 1 + EXAM1 + EXAM2 + EXAM3
```

| Estimated Coefficients: | Estimate | SE      |
|-------------------------|----------|---------|
| (Intercept)             | -4.3361  | 3.7642  |
| EXAM1                   | 0.35594  | 0.12139 |
| EXAM2                   | 0.54252  | 0.10085 |
| EXAM3                   | 1.1674   | 0.10301 |

- LR, kde je závisle proměnná  $y$  ovlivněna větším počtem nezávislých veličin  $x$ , které jsou **vzájemně nezávislé**.

- Předpis:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

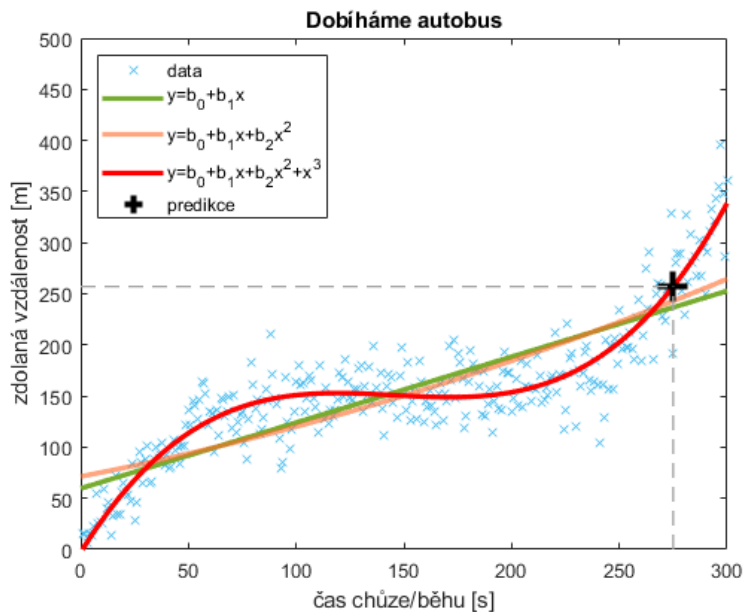
- Mohou nastat 2 situace:

1. Hledáme regresní koeficienty –  $b = [b_0, b_1, \dots, b_n]$  (známe  $x$  a  $y$ )
  - Použijeme metodu nejmenších čtverců v maticovém tvaru
2. Hledáme očekávaný výstup (predikci) -  $y$  (známe  $b$  a  $x$ )

- Příklad:

1. Hledáme předpis (vztah), kterým popíšeme předpokládaný výsledek závěrečného testu (0-200b) na základě výsledků 3 předchozích testů (0-100b)
2. Kolik bodů asi získáme z posledního testu, když jsme z předchozích získali
  - a)  $50 - 70 - 90$  b  $\rightarrow 156,5$  b
  - b)  $90 - 70 - 50$  b  $\rightarrow 124,0$  b

# Polynomiální regrese



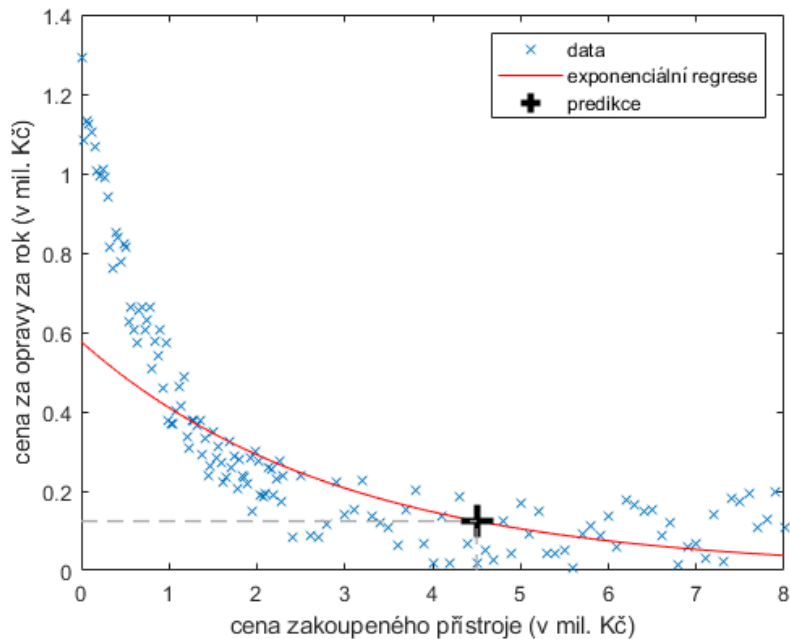
- Polynomiální regrese je regrese, která body prokládá křivkou určitého stupně
  - Čím vyšší stupeň (mocnina), tím složitější křivka – tím víc „zlomů“
- Předpis

$$y = b_0 + b_1x + b_2x^2 + \dots + b_nx^n$$

- Mohou nastat 2 situace:
  - Hledáme regresní koeficienty –  $b = [b_0, b_1, \dots, b_n]$  (známe  $x$  a  $y$ )
    - Použijeme metodu nejmenších čtverců v maticovém tvaru
  - Hledáme očekávaný výstup (predikci) -  $y$  (známe  $b$  a  $x$ )

- Příklad:
  - Hledáme předpis (vztah), kterým popíšeme náš běh na autobus, který je vzdálen 350 m
  - Známe předpis (vztah) pro běh a chceme vědět, zda stihneme za 275s doběhnout autobus, který je vzdálen 300 m

# Exponenciální regrese



- Exponenciální regrese patří k nelineárním metodám, ale lze linearizovat logaritmováním.

- Předpis:

$$y = b_0 e^{b_1 x}$$

- Předpis po zlogaritmování

$$\underbrace{\ln(y)}_Y = \underbrace{\ln(b_0)}_{B_0} + b_1 x$$

- Mohou nastat 2 situace:

1. Hledáme regresní koeficienty –  $B_0$  a  $b_1$  (známe  $x$  a  $Y$ )

- Použijeme metodu nejmenších čtverců

2. Hledáme očekávaný výstup (predikci) -  $Y$  (známe  $B_0$ ,  $b_1$  a  $x$ ). Po zjištění  $Y$  se skutečné  $y$  opět dopočte použitím exponenciální funkce

- Příklad:

1. Hledáme předpis (vztah), kterým popíšeme vliv ceny zakoupeného přístroje na cenu pozdějších oprav
2. Známe předpis (vztah) cen za zakoupení i oprav a chceme vědět, kolik budeme platit za opravy, pokud investujeme 4,5 mil. do nákupu.