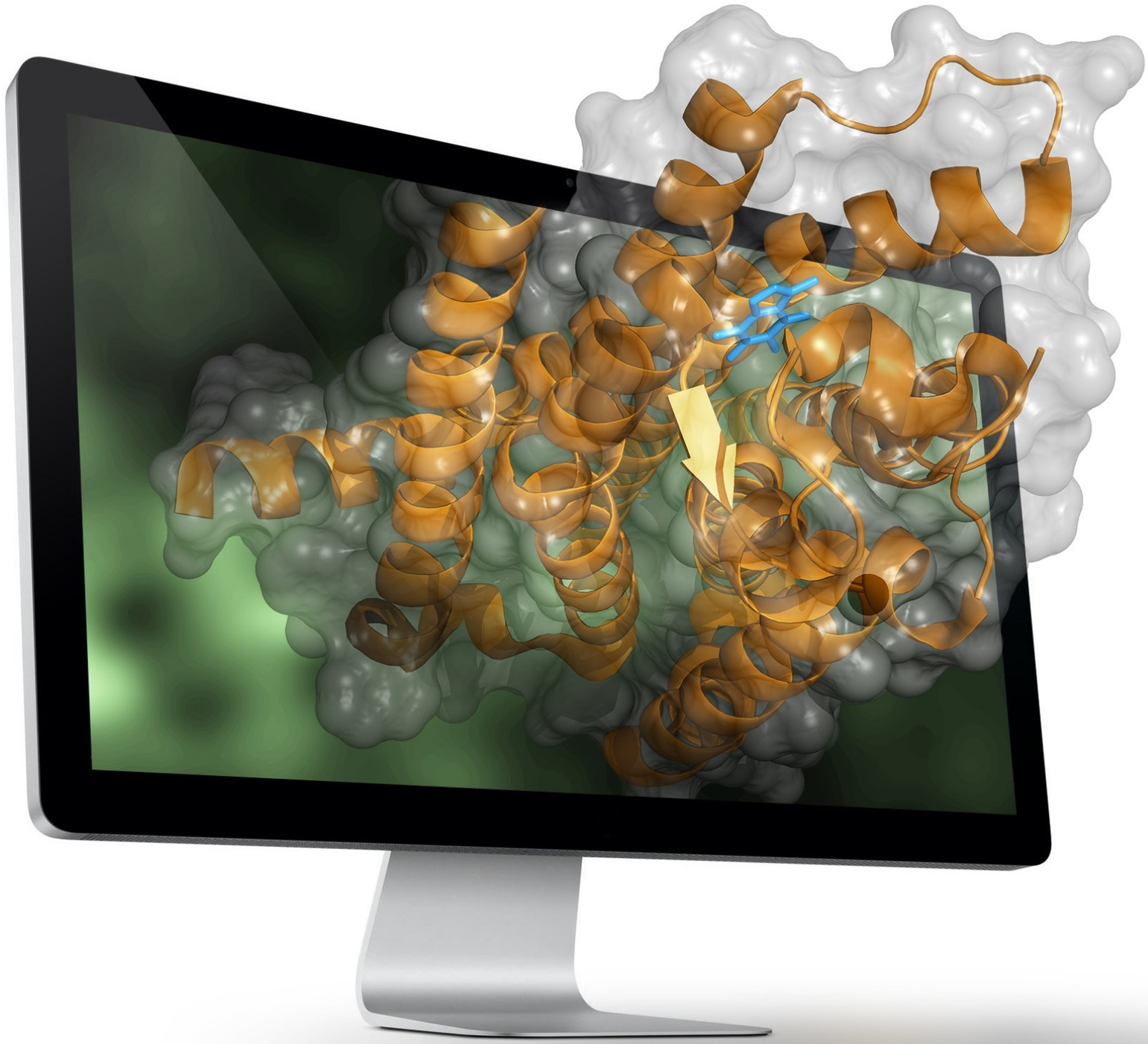


Pokročilá chemoinformatika

Lineární modely
únor 2017

První QSAR modely

Historie

- První QSAR modely využité pro popis biologické aktivity:

$$\log(1/C) = k_1 \log P + k_2\sigma + k_3$$

$$\log(1/C) = k_1\pi + k_2\sigma + k_3; \quad \pi = \log P_X - \log P_H$$

$$\log(1/C) = -k_1 (\log P)^2 + k_2 (\log P) + k_3\sigma + k_4$$

$$\log(1/C) = -0.44 (\log P)^2 + 1.58 (\log P) + 1.93$$

- Kde:
  - C je koncentrace nutná pro vyvolání reakce
  - log P je rozdělovací koeficient
  - σ je Hammetův parametr
- Publikováno pány Hanschem a Fujitou v letech 1964 - 1969

Metody tvorby QSAR/QSPR modelů

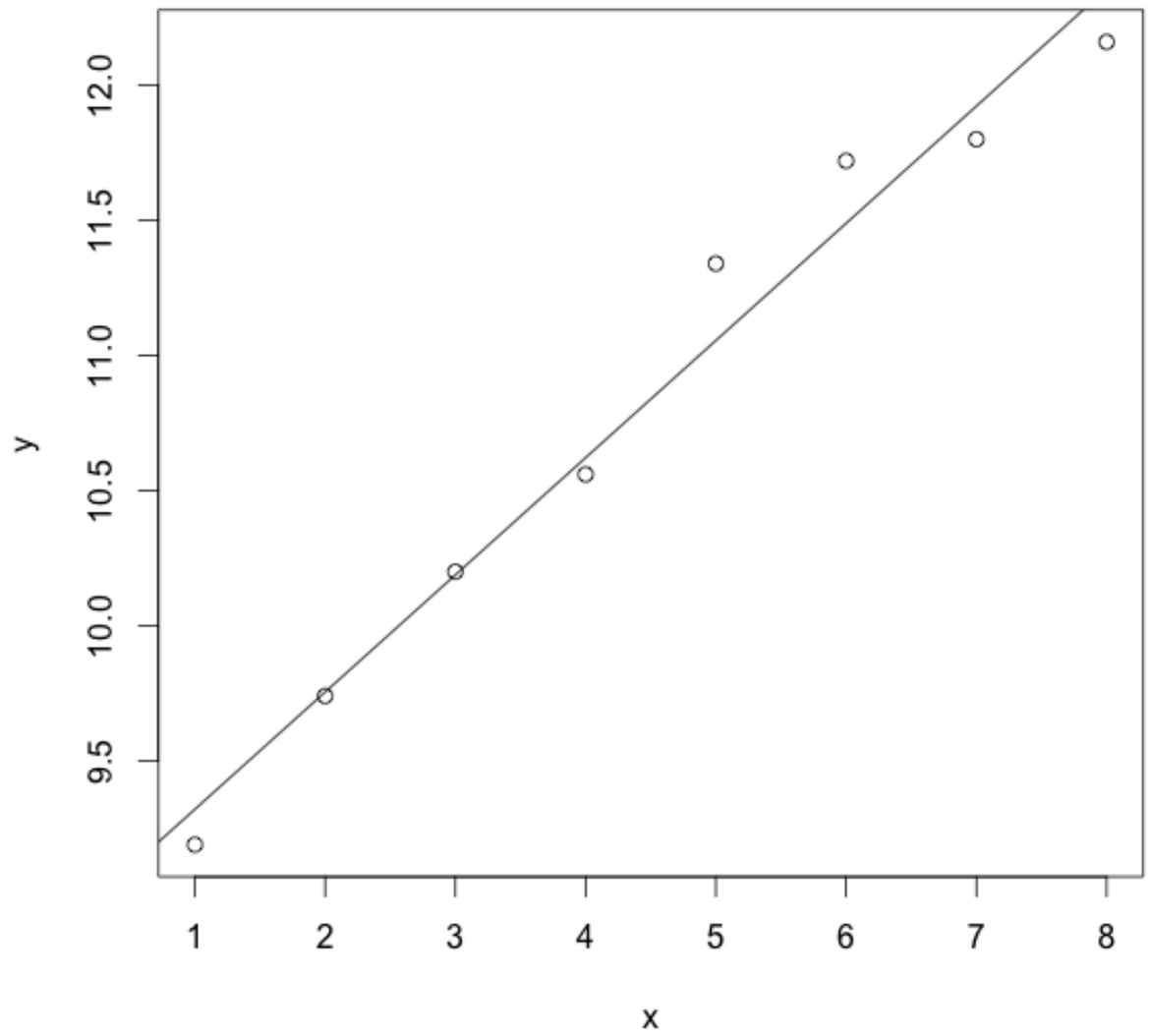
Metody tvorby modelů

- **lineární, logistická, zobecněná regrese**
- **MLR (Multiple Linear Regression)**
- KNN (K-Nearest Neighbors)
- Decision Tree a Random Forest
- ASNN (ASsociative Neural Networks)
- Naive Bayes
- Support Vector Machine

Regrese

Lineární regrese

- proměnné **x** a **y**, hledáme takové *a* a *b*, které nejlépe popiše vzájemný lineární vztah
- $y = ax + b$
- Používáme metodu nejmenších čtverců pro minimalizaci výsledné sumy vzdálenosti bodů od přímky
- Požadavek na normální rozložení dat!



Lineární regrese – výpočet parametrů

- Při hledání parametrů *a* a *b* potřebujeme znát sadu hodnot *x* a *y* (hodnoty deskriptorů a vlastností:

$$y = ax + b$$

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Lineární regrese – Excel a Calc

- V Excelu a Calcu (LibreOffice a OpenOffice.org) lze koeficient *a* zjistit funkcí **SLOPE(Y; X)** a konstantu *b* funkcí **INTERCEPT(Y; X)**.

- Případně lze oba koeficienty zjistit maticově zadanou funkcí **{=LINEST(Y;X)}**. V českém Excelu se tato funkce nazývá **LINREGRESE**.

SLOPE function

This article describes the formula syntax and usage of the **SLOPE** function in Microsoft Excel.

Description

Returns the slope of the linear regression line through data points in known_y's and known_x's. The slope is the vertical distance divided by the horizontal distance between any two points on the line, which is the rate of change along the regression line.

Syntax

SLOPE(known_y's, known_x's)

The SLOPE function syntax has the following arguments:

- **Known_y's** Required. An array or cell range of numeric dependent data points.
- **Known_x's** Required. The set of independent data points.

Lineární regrese v R

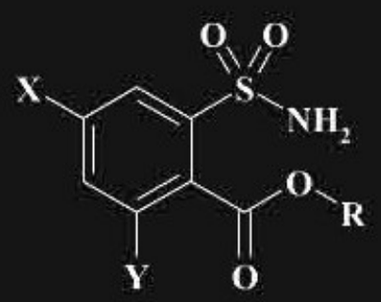
- dataset = read.csv(“filename.csv”, sep = ”;”)
  model = lm(Y~x, data = dataset)
  model
  plot(model)

```
Call:
lm(formula = pKa ~ qH, data = data)

Coefficients:
(Intercept)           qH
      29.62      -110.06
```

Proč více proměnných (regresorů, deskriptorů) v modelu?

Více deskriptorů v modelu - důvod

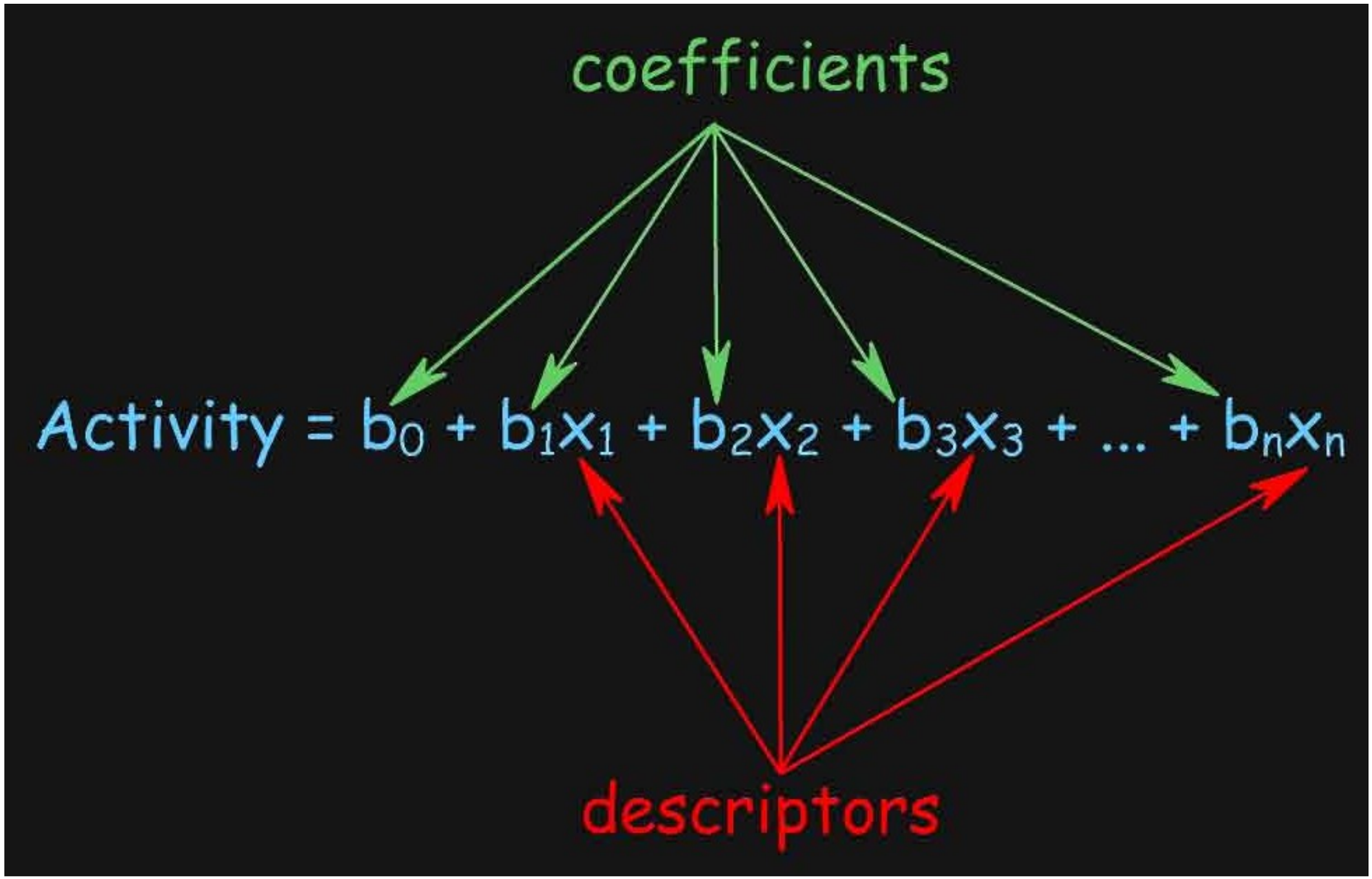


bad model

good model

model	r
log 1/C = 0.009 Es + 3.411	0.03
log 1/C = -0.626 σ + 3.314	0.27
log 1/C = -0.078 logP + 3.432	0.38
log 1/C = -0.210 logP - 2.214 σ + 3.154	0.80
log 1/C = 0.21 Es - 0.238 logP - 3.81 σ + 3.046	0.95

Více rozměrné modely



Vícerozměrné modely

- QSAR/QSPR modely jsou popsány rovnicí:

- $A = p_1.d_1 + p_2.d_2 + p_3.d_3 + \ldots + p_n.d_n + p_{n+1}$

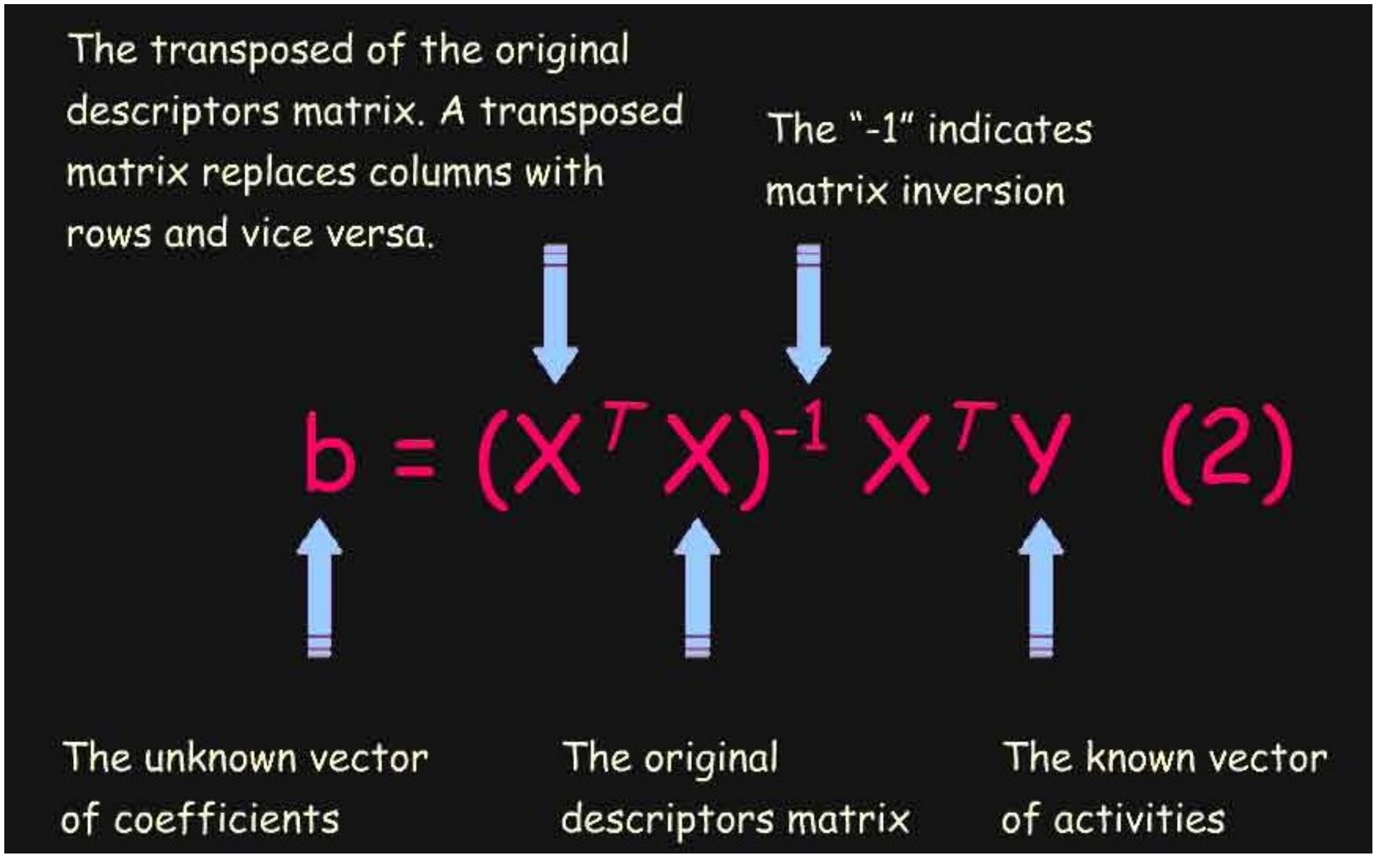
- Kde:
- $p_1, p_2, \ldots, p_{n+1}$ jsou parametry modelu
- p_{n+1} je intercept
- $d_1, d_2, \ldots, d_n$ jsou deskriptory (nezávislé proměnné, regresory)
- A je predikovaná aktivita případně vlastnost

Vícerozměrné modely – maticový zapis

$$y = x\,b + e$$

$$\underset{n}{\underset{1}{y}} = \underset{n}{\underset{m}{x}}\ \underset{m}{\underset{1}{b}} + \underset{n}{\underset{1}{e}}$$

Vícerozměrné modely – výpočet koeficientů



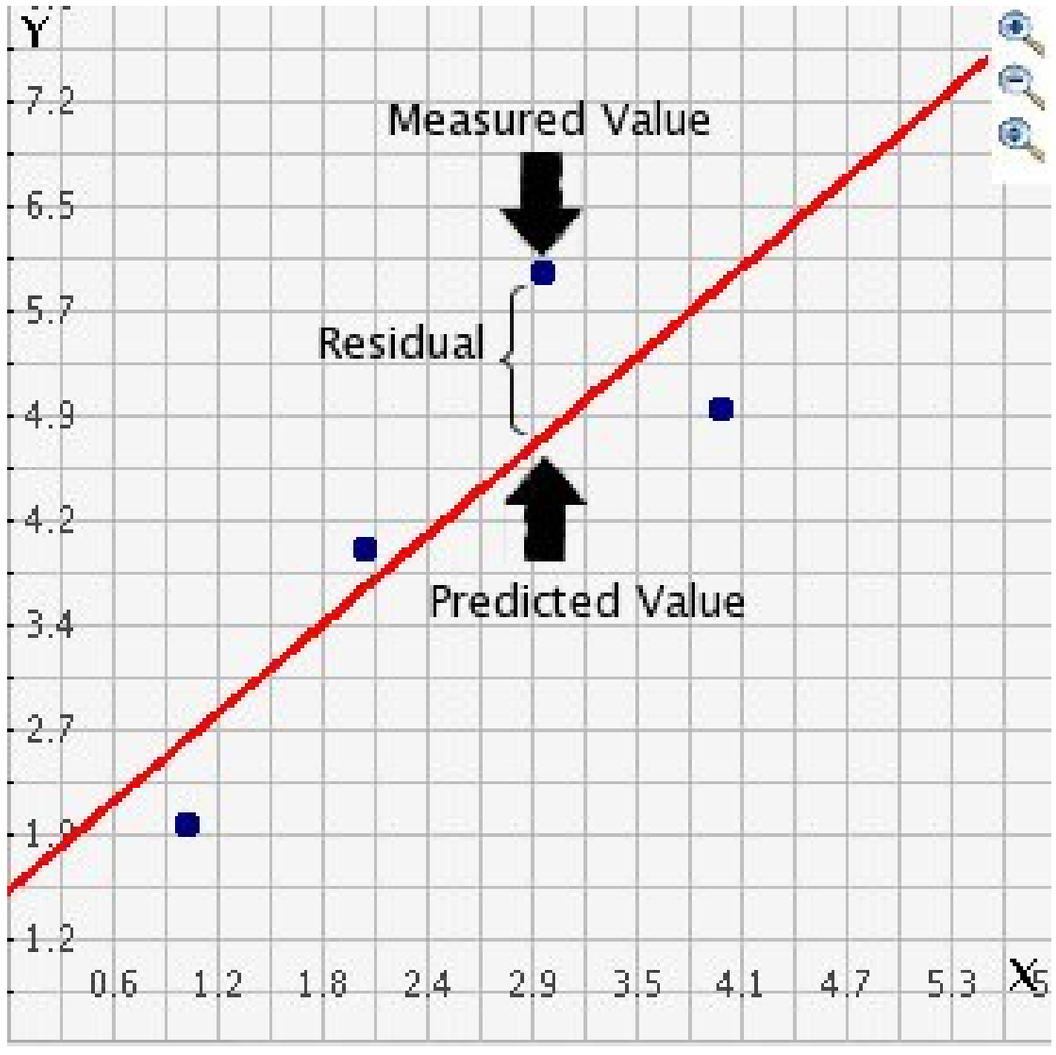
Vícerozměrné modely

- Obecně:

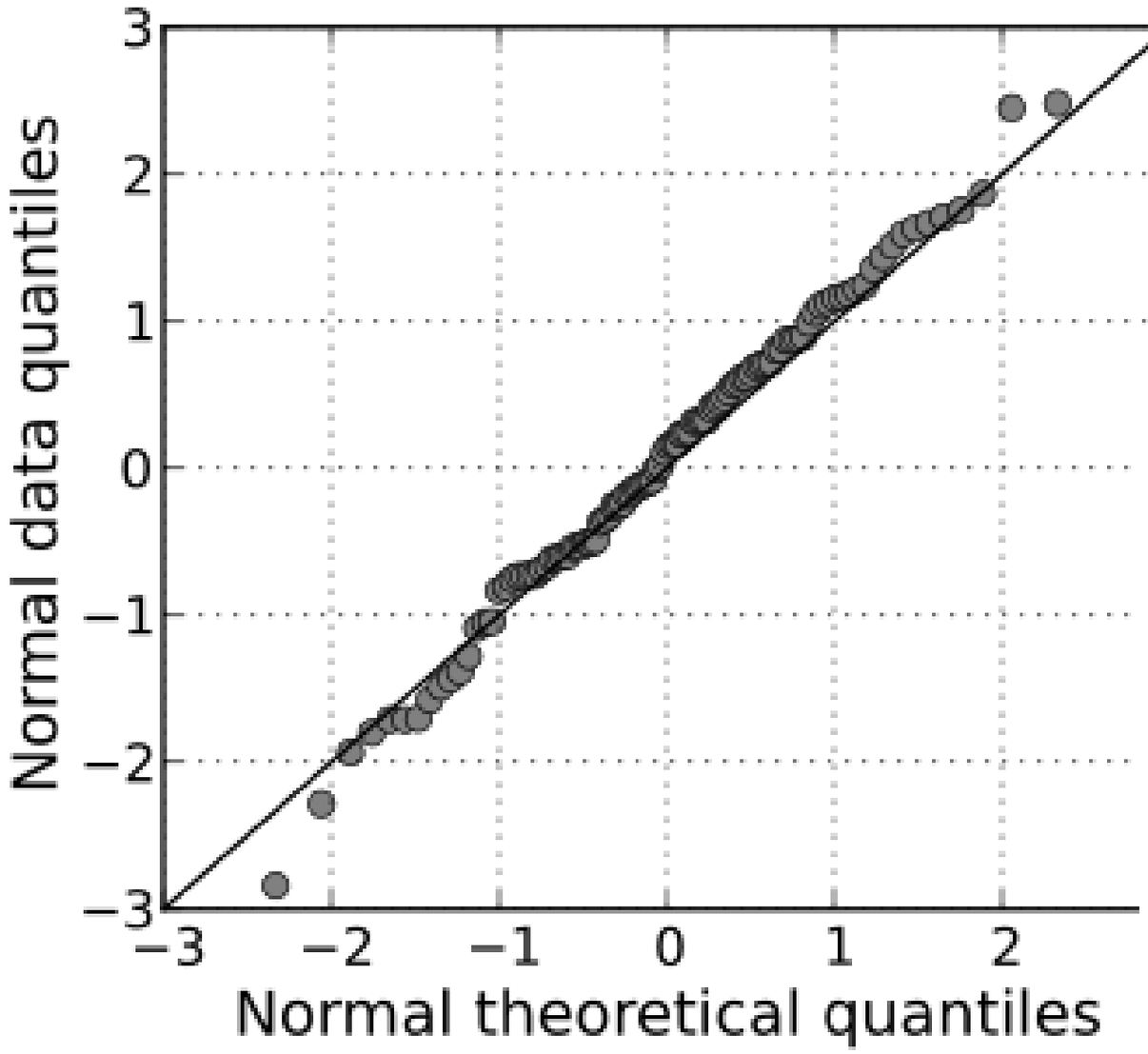
$$Y_i = \sum_{j=1}^{m} X_{ij}\beta_j + \varepsilon_i, \;\; i = 1, \ldots, n,$$

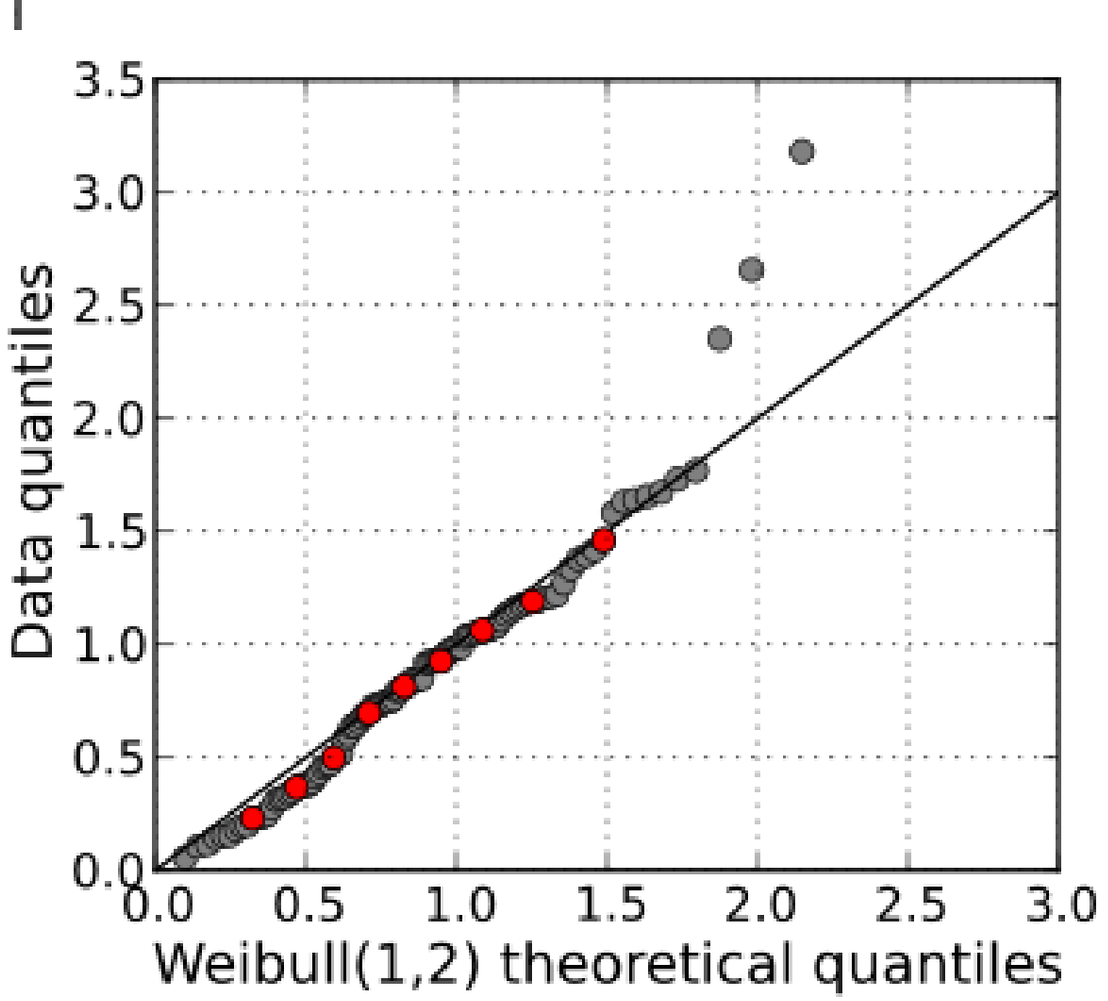
- kde ε je chyba, residuum

- tuto chybu se snažíme minimalizovat a označit jako nezávislou chybu



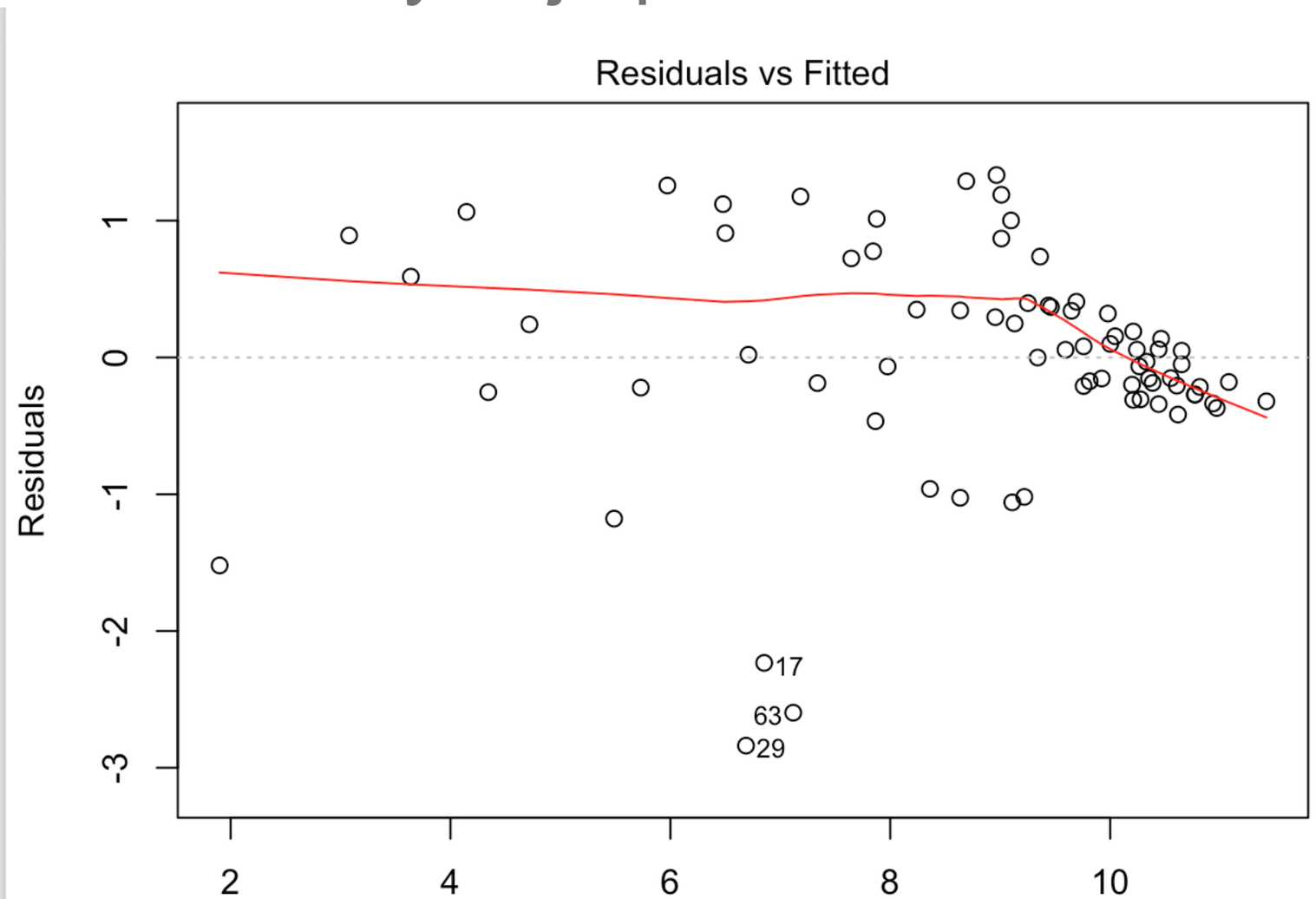
Vizualizace chyby v quantilech – QQ plot





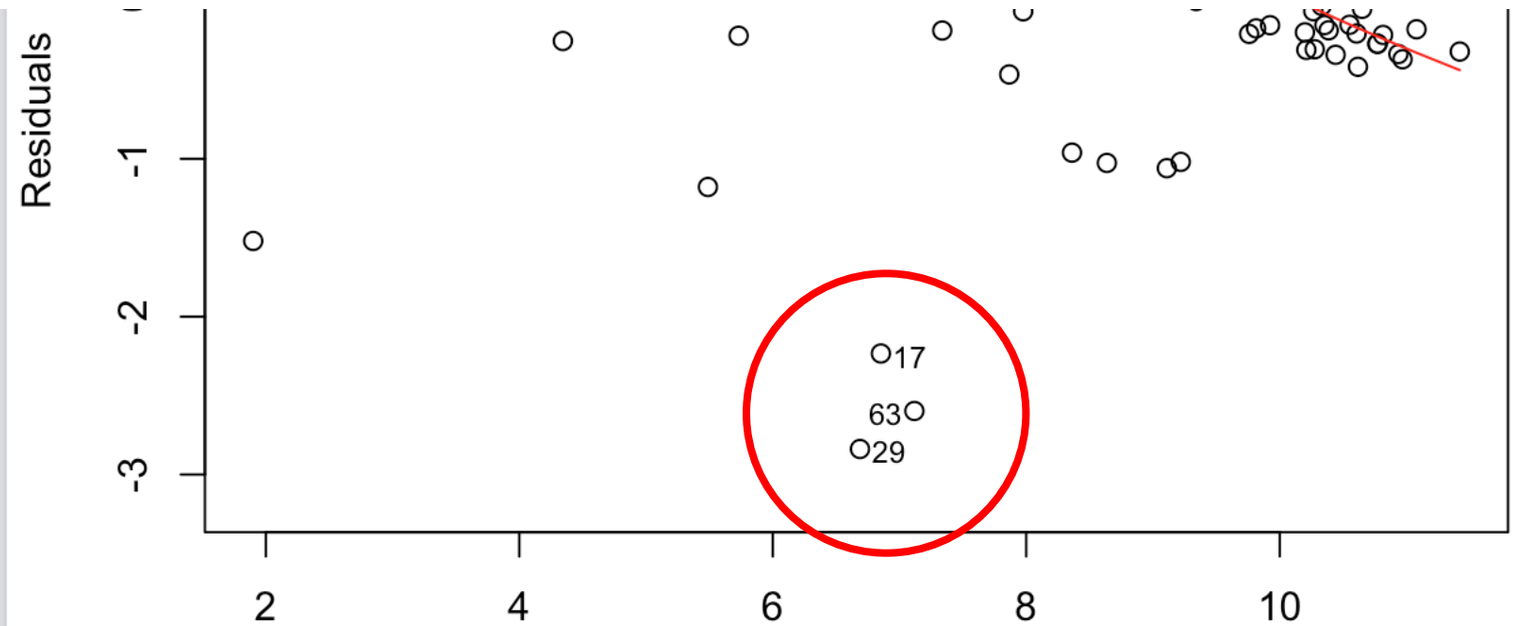
Odlehlé a pákové body

- Jedná se o pozorování nebo měření, která nezapají do modelu.
- Jejich residuální chyba je příliš velká.



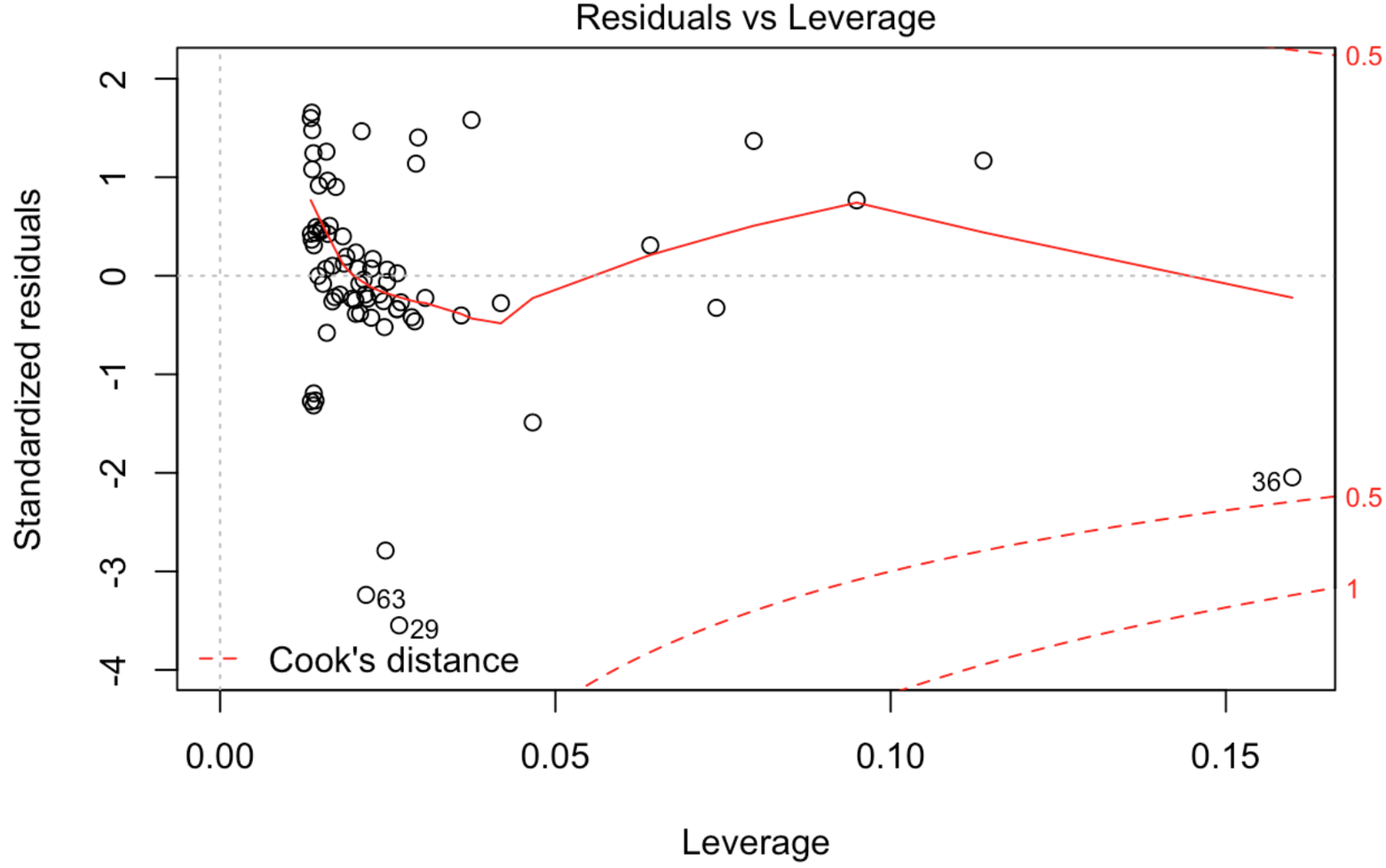
Odlehlé body (outliers)

- Mohou nepříjemně ovlivňovat kvalitu modelu.
- Můžeme jejich hodnoty odstranit nebo přeměřit/ověřit.
  - Můžeme ověřit jestli například nepatří tyto molekuly do stejné skupiny a vyřadit tuto skupinu.
- Detekce na základě velikosti chyby – residua.

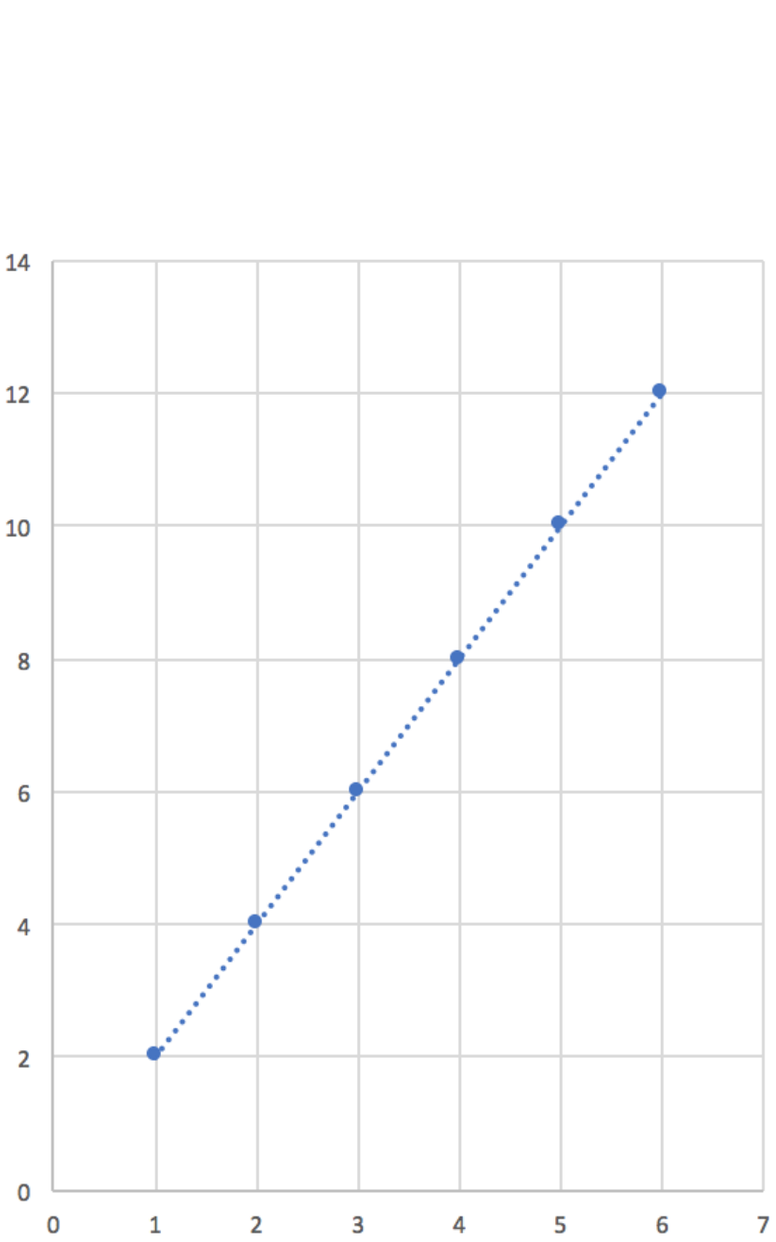


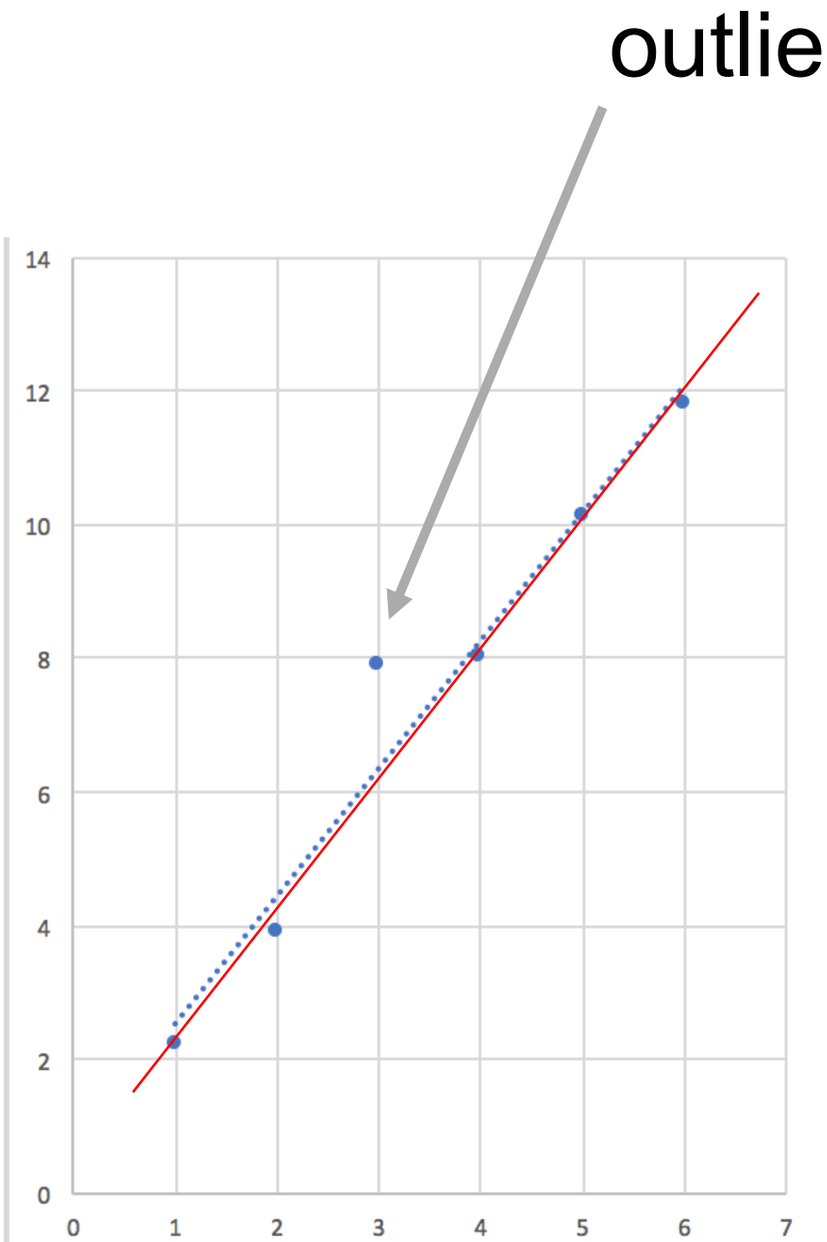
Pákové body (leverage points)

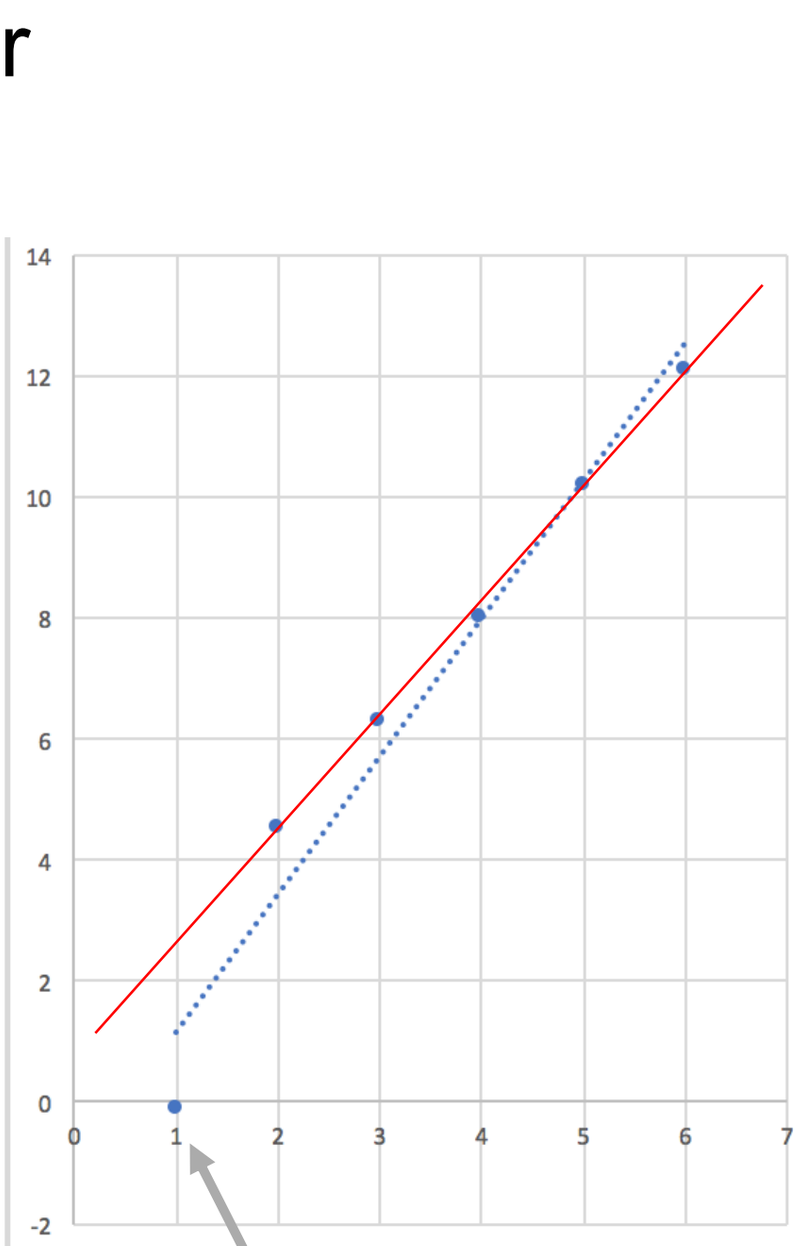
- Podobně jako odlehlé body nezapadají do modelu a mají vysokou hodnotu chyby.
- Tyto hodnoty bohužel velice zásadně ovlivňují kvalitu modelu.
- Detekce pomocí Cookovi vzdálenosti.



Rozdíl mezi odlehlým a pákovým bodem







leverage

Logistická regrese

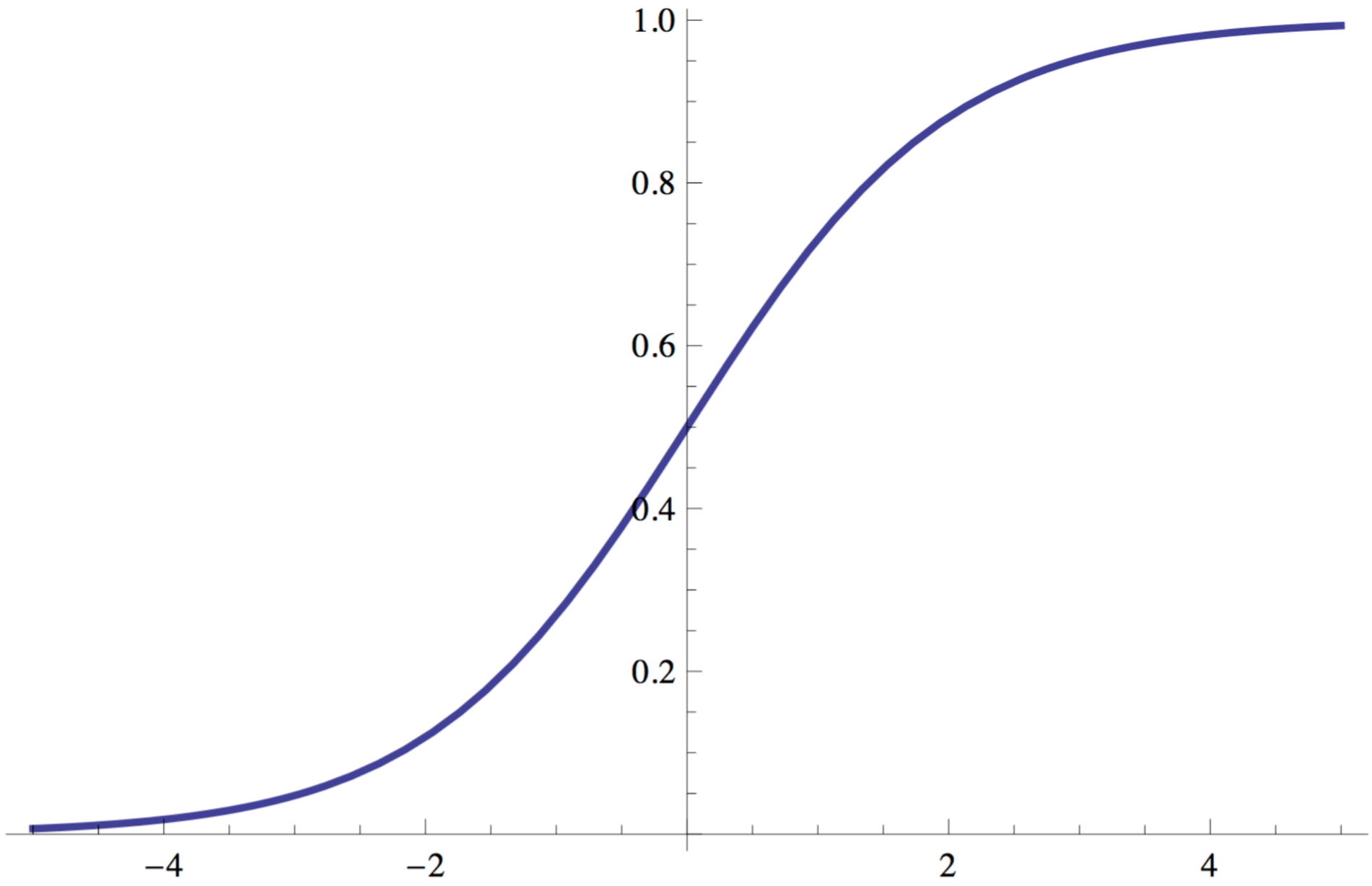
- Výsledek nabývá hodnot v rozsahu <0;1>
- $$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$
- Vícerozměrově

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_m x_m)}$$

- α a β jsou parametry modelu

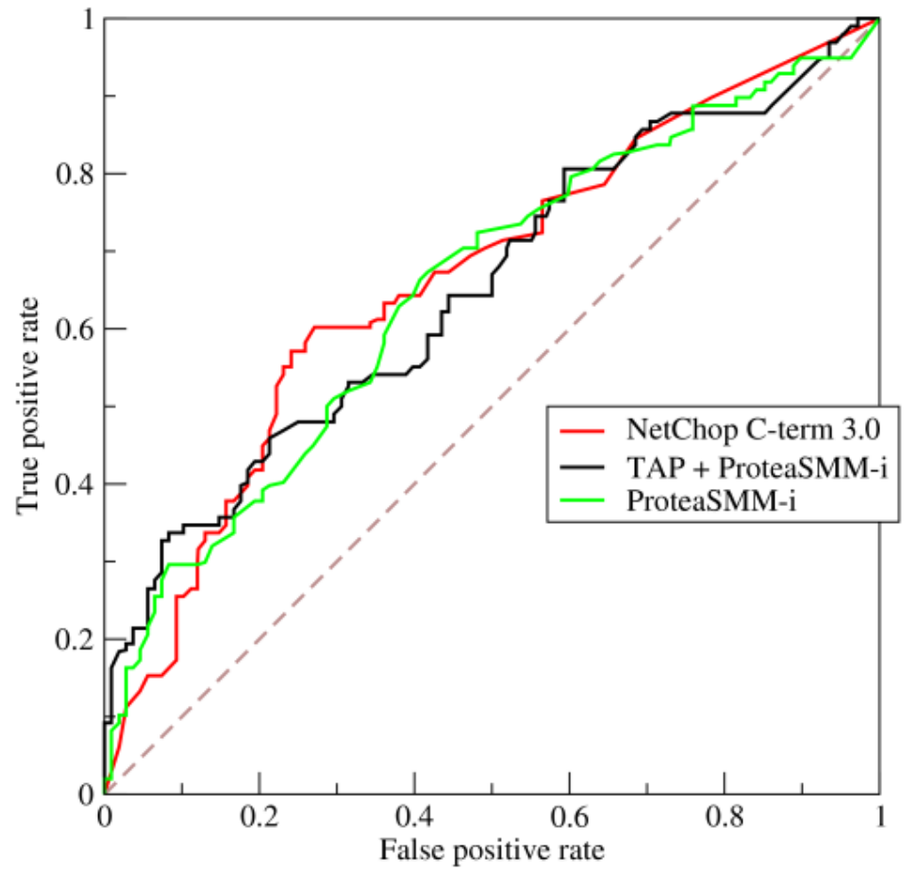
Logistická regrese - graficky

$$\frac{e^{\theta}}{1+e^{\theta}}$$



Kvalita logistické křivky - ROC

- **ROC (Receiver Operating Characteristic) křivka** je nástroj pro hodnocení a optimalizaci binárního klasifikačního systému (testu), který ukazuje vztah mezi specificitou a senzitivitou daného testu nebo detektoru pro všechny přípustné hodnoty prahu.



Specificita a senzitivita

		SKUTEČNÁ HODNOTA (např. počet pacientů s rakovinou tlustého střeva diagnostikovanou endoskopicky)	
		Pozitivní (p)	***Negativní (n)***
ZMĚŘENÁ HODNOTA (např. testem krve ve stolici)	**test pozitivní**	**Skutečně Pozitivní** (TP) = 20	**Falešně Pozitivní** (Chyba I typu) (FP) = 180
	test negativní	**Falešně Negativní** (Chyba II typu) (FN) = 10	**Skutečně Negativní** (TN) = 1820

→ prediktivní hodnota pozitivního testu PPV
= TP / (TP + FP)
= 20 / (20 + 180) = 10 %

→ prediktivní hodnota negativního testu NPV
= TN / (FN + TN)
= 1820 / (10 + 1820) ≈ 99,5 %

P
↓
Senzitivita
= TP / (TP + FN)
= 20 / (20 + 10)
≈ 66,67 %

N
↓
Specificita
= TN / (FP + TN)
= 1820 / (180 + 1820)
= 91 %

Interagující regresory

Interakce regresorů (deskriptorů)

- Y – závislá proměnná
- x1 a x2 nezávislá proměnná
- Y = a□x1 + b□x2 + z

Y = a□x1 + b□x2 + c□**x1**□**x2** + z

Y = a□x1 + c□**x1**□**x2** + z

V Rku:

Y ~ x1 + x2

Y ~ x1*x2

Y ~ x1 + x1:x2