

# BI-ML1.21 přednáška 5

Daniel Vašata

FIT ČVUT

23. 10. 2024

Autoři: Karel Klouda, Daniel Vašata.  
Problémy, návrhy apod. hlaste v [GitLabu](#).  
Verze souboru: 30. října 2024 15:48.

## Co bude v dnešní přednášce

- Představení modelu lineární regrese
- Úvod do problematiky hledání extrémů funkce více proměnných
- Odhad parametrů lineárního modelu pomocí metody nejmenších čtverců
- Diskuse výsledků

## Motivační příklad

- Chceme prodat nemovitost (řekněme v Praze) a netušíme, za kolik ji máme potenciálním zájemcům nabídnout.
- Zároveň si nechceme platit žádného „realitního“ odborníka, který by nám se stanovením ceny poradil.
- Zkusíme tedy udělat vlastní průzkum realitního trhu a vytvořit model, který nám pomůže cenu stanovit.
- Napíšeme skript, který stáhne data z realitních serverů a uloží je v nějaké strukturované podobě (ideálně tabulce či více tabulkách).
- Pro jednoduchost uvažujme, že budeme znát u každé nabídky toto:
  - ▶  $Y$  – cenu, za kterou se prodává,
  - ▶  $X_1$  – užitnou plochu,
  - ▶  $X_2$  – počet místností,
  - ▶  $X_3$  – vzdálenost od nejbližší zastávky metra.
- Jako vysvětlovanou proměnnou  $Y$  jsme označili veličinu, kterou pro naši nemovitost neznáme a chceme ji predikovat.
- Příznaky  $X_1, \dots, X_3$  označují veličiny, které pro naši nemovitost známe a o kterých věříme, že cenu  $Y$  ovlivňují.

## Formalizace úlohy

Obecně tedy chceme na základě  $p$  příznaků  $X_1, \dots, X_p$  predikovat hodnotu vysvětlované proměnné  $Y$ .

V modelu lineární regrese předpokládáme **lineární závislost** vysvětlované proměnné na hodnotách příznaků.

Jelikož nedoufáme, že tato závislost je perfektní v tom smyslu, že pro stejné hodnoty  $x_1, \dots, x_p$  příznaků  $X_1, \dots, X_p$  dostaneme vždy stejnou hodnotu vysvětlované proměnné  $Y$ , modelujeme tuto závislost následovně:

$$Y = w_1x_1 + \dots + w_px_p + \varepsilon,$$

kde  $w_1, \dots, w_p$  jsou nějaké neznámé koeficienty a  $\varepsilon$  je náhodná veličina.

### Poznámky:

- Veličina  $\varepsilon$  odpovídá části  $Y$ , která je nevysvětlitelná pomocí hodnot příznaků a je tedy z našeho pohledu náhodná.
- Do náhodné veličiny  $\varepsilon$  se tak „schovají“ vlivy, které **neznáme** nebo cíleně **nezahrnujeme** do našeho modelu (např. stáří budovy, počet koupelen, počet oken) ale např. i chyby, nekonzistence dat a jiné podivnosti v měření příznaků.

## Model lineární regrese

Obvykle ještě oddělujeme střední hodnotu náhodných vlivů a dostáváme tak:

### Model lineární regrese

Hodnota vysvětlované proměnné  $Y$  v bodě  $(x_1, \dots, x_p)^T$  je

$$Y = w_0 + w_1x_1 + \dots + w_px_p + \varepsilon,$$

kde  $E\varepsilon = 0$ .

- Koeficient  $w_0$  se nazývá **intercept** a odpovídá (očekávané) výchozí hodnotě  $Y$  při nulových příznacích.
- Zavedeme-li nový konstantní příznak  $X_0 = x_0 = 1$  a vektorové značení

$$\mathbf{x} = (x_0, x_1, \dots, x_p)^T \quad \text{a} \quad \mathbf{w} = (w_0, w_1, \dots, w_p)^T,$$

můžeme zkráceně psát

$$Y = \mathbf{w}^T \mathbf{x} + \varepsilon.$$

- Vektor  $\mathbf{w} = (w_0, w_1, \dots, w_p)^T$  koeficientů také někdy nazýváme **vektor vah**.

## Predikce v modelu lineární regrese

Předpokládejme nyní, že už máme odhad  $\hat{\mathbf{w}}$  vektoru koeficientů  $\mathbf{w}$ .

Hodnotu  $Y$  v konkrétním bodě  $\mathbf{x}$  predikujeme vztahem

$$\hat{Y} = \hat{\mathbf{w}}^T \mathbf{x} = \hat{w}_0 + \hat{w}_1 x_1 + \dots + \hat{w}_p x_p.$$

Skutečná hodnota  $Y$  v bodě  $\mathbf{x}$  je přitom určena vztahem

$$Y = \mathbf{w}^T \mathbf{x} + \varepsilon$$

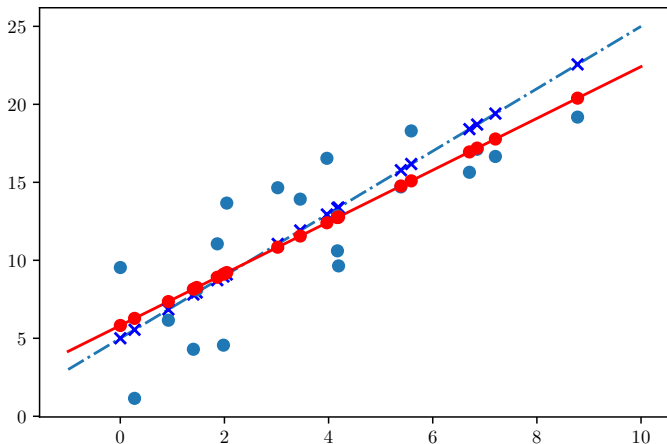
a je tedy náhodnou veličinou.

Z předpokladu  $E\varepsilon = 0$  plyne, že

$$EY = \mathbf{w}^T \mathbf{x}$$

a  $\hat{Y}$  je tedy vlastně bodovým odhadem střední hodnoty  $EY$  v bodě  $\mathbf{x}$ .

## Vizualizace modelu lineární regrese



Modré body jsou body trénovací množiny. Červené body jsou predikce. Modré křížky odpovídají středním hodnotám bodů trénovací množiny,  $(x_i, E Y_i)$ . Modrá čerchovaná čára je skutečná regresní přímka daná rovnicí  $y = w^T x$  a červená čára je přímka  $\hat{y} = \hat{w}^T x$  určující naše predikce.

## Měření chybovosti predikce pomocí ztrátové funkce

Zaměříme se nyní na problematiku odhadu vektoru parametrů modelu  $w$ .

Následující úvahy jsou obecně platné pro supervizované učení nějakého modelu s parametry.

- Naším cílem je najít takovou hodnotu  $w$ , aby chyba modelu byla co nejmenší.
- Tuto hodnotu pak použijeme jako odhad  $\hat{w}$ .
- K tomu musíme specifikovat, **co se myslí chybou modelu** a **v jakém smyslu má být nejmenší**.
- Chybu modelu nejčastěji měříme pomocí nějaké nezáporné funkce  $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ , nazývané **ztrátová funkce** (angl. **loss function**), kterou aplikujeme na skutečnou hodnotou proměnné  $Y$  a odpovídající predikci  $\hat{Y}$ .
- Obvyklou volbou v případě spojité vysvětlované veličiny bývá **kvadratická ztrátová funkce**,

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2.$$



## Metoda nejmenších čtverců

- Velikost chyby modelu v bodě  $\mathbf{x}$  je tedy  $L(Y, \hat{Y})$ , kde  $Y$  je skutečná hodnota vysvětlované  $Y$  v bodě  $\mathbf{x}$  a  $\hat{Y} = \mathbf{w}^T \mathbf{x}$  je predikce v bodě  $\mathbf{x}$ .
- Otázkou je, v jakém bodě  $\mathbf{x}$  bychom měli hodnotu  $L(Y, \hat{Y})$  vyhodnocovat a následně minimalizovat vzhledem k  $\mathbf{w}$ .
- Zřejmě to musí být bod  $\mathbf{x}$  z trénovací množiny, protože jinak bychom neznali skutečnou hodnotu  $Y$  v tomto bodě.
- Abychom co nejvíc využili trénovací data, budeme minimalizovat součet chyb přes všechny body trénovací množiny, tj. přes všechny dvojice  $(\mathbf{x}_i, Y_i)$  pro  $i = 1, \dots, N$ .
- Součet chyb přes všechny tyto body pro kvadratickou ztrátovou funkci je

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^N L(Y_i, \mathbf{w}^T \mathbf{x}_i) = \sum_{i=1}^N (Y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

a nazýváme ho **reziduální součet čtverců** (angl. **residual sum of squares**).

- Minimalizací tohoto výrazu získáme odhad  $\hat{\mathbf{w}}$ . Tento postup se nazývá **metoda nejmenších čtverců** (angl. the method of **least squares**).

## Parciální derivace

Protože  $w$  je vektor, minimalizace  $\text{RSS}(w)$  spadá do problematiky minimalizace funkce více proměnných.

Jak si nyní ukážeme, postupuje se analogicky jako v případě funkce jedné proměnné.

### Definice

Bud'  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  funkce  $d$  proměnných. **Parciální derivaci** funkce  $f(x_1, \dots, x_d)$  podle proměnné  $x_i$  v bodě  $\mathbf{a} = (a_1, \dots, a_d) \in \mathbb{R}^d$  definujeme jako derivaci funkce  $g(x_i) = f(a_1, \dots, a_{i-1}, x_i, a_{i+1}, \dots, a_d)$  v bodě  $a_i$  a značíme

$$\partial_{x_i} f(\mathbf{a}) \quad \text{nebo} \quad \frac{\partial f}{\partial x_i}(\mathbf{a}).$$

Označením  $\partial_{x_i} f$  nebo  $\frac{\partial f}{\partial x_i}$  pak myslíme funkci, která každému bodu, kde je konečná, přiřadí hodnotu parciální derivace podle  $x_i$  v tomto bodě.

Parciální derivace je stejná jako ta „obyčejná“, akorát ostatní proměnné bereme jako konstanty.

Platí např.  $\partial_x(x + y) = 1$  nebo  $\partial_y \sin(y + xy) = (1 + x) \cos(y + xy)$ .

# Gradient funkce

## Definice

Bud'  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  funkce  $d$  proměnných, která má v bodě  $\mathbf{a} \in \mathbb{R}^d$  konečné všechny parciální derivace. **Gradient** funkce  $f$  v bodě  $\mathbf{a}$  definujeme jako vektor

$$\nabla f(\mathbf{a}) = \left( \frac{\partial f}{\partial x_1}(\mathbf{a}), \dots, \frac{\partial f}{\partial x_d}(\mathbf{a}) \right).$$

Označením  $\nabla f$  pak myslíme gradient funkce jakožto zobrazení, které každému bodu, kde to lze, přiřadí gradient v tomto bodě.

Gradient v bodě je tedy vektor z  $\mathbb{R}^d$ , jehož složky jsou jednotlivé parciální derivace.

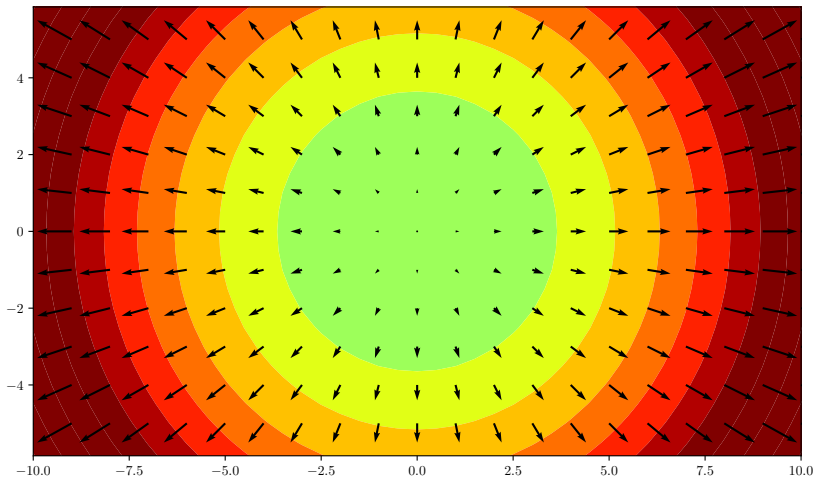
Uvažujme funkci  $f(x, y) = x^2 + y^2$ , která odpovídá parabolické jámě s minimem v počátku  $\mathbf{0}$ . Platí  $\nabla f = (2x, 2y)$ .

Nejdůležitější vlastností gradientu<sup>1</sup> je, že ukazuje **směr maximálního růstu** funkce v daném bodě. Z toho mimo jiné plyne, že je gradient vždy **kolmý na vrstevnici** procházející daným bodem.

---

<sup>1</sup>Za dodatečného předpokladu spojitosti všech jeho složek na okolí daného bodu.

## Vizualizace gradientu



## Hessova matice

Analogicky k jednorozměrnému případu platí, že má-li funkce v nějakém bodě lokální extrém a gradient zde existuje, musí být nulový.

K získání postačující podmínky nám ještě zbývá sestrojít analog druhé derivace.

### Definice

Bud'  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  funkce  $d$  proměnných. **Hessovu matici** funkce  $f$  v bodě  $\mathbf{a} \in \mathbb{R}^d$  definujeme jako

$$\mathbf{H}_f(\mathbf{a}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\mathbf{a}) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(\mathbf{a}) \end{pmatrix},$$

kde  $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}) = \left( \frac{\partial}{\partial x_i} \left( \frac{\partial f}{\partial x_j} \right) \right)(\mathbf{a})$  značí druhou parciální derivaci podle  $x_j$  a  $x_i$ .

Je to tedy matice druhých parciálních derivací podle všech proměnných.

## Postačující podmínka pro existenci lokálního extrému

## Věta

Bud'  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  funkce  $d$  proměnných a bod  $\mathbf{x}^* \in \mathbb{R}^d$  takový, že  $\nabla f(\mathbf{x}^*) = \mathbf{0}$  a  $f$  má spojitě všechny druhé parciální derivace na nějakém okolí bodu  $\mathbf{x}^*$ .

- Jestliže

$$\mathbf{s}^T \mathbf{H}_f(\mathbf{x}^*) \mathbf{s} > 0, \quad \text{pro každé } \mathbf{s} \in \mathbb{R}^d, \mathbf{s} \neq \mathbf{0},$$

nabývá funkce  $f$  v bodě  $\mathbf{x}^*$  ostrého lokálního minima.

- Jestliže pro každé  $\mathbf{x}$  z nějakého okolí bodu  $\mathbf{x}^*$

$$\mathbf{s}^T \mathbf{H}_f(\mathbf{x}) \mathbf{s} \geq 0, \quad \text{pro každé } \mathbf{s} \in \mathbb{R}^d,$$

nabývá funkce  $f$  v bodě  $\mathbf{x}^*$  neostrého lokálního minima.

Tyto vlastnosti se nazývají **pozitivní definitnost** resp. **pozitivní semi-definitnost** Hessovy matice  $\mathbf{H}_f$  v bodě  $\mathbf{x}^*$  resp.  $\mathbf{x}$ .

Pro funkci  $f(x, y) = x^2 + y^2$  je řešením  $\nabla f = (2x, 2y) = \mathbf{0}$  bod  $\mathbf{x}^* = \mathbf{0}$  a Hessova matice je

$$\mathbf{H}_f(\mathbf{x}^*) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2\mathbf{I}.$$

Protože  $\mathbf{s}^T 2\mathbf{I} \mathbf{s} = 2\mathbf{s}^T \mathbf{s} = 2\|\mathbf{s}\|^2 > 0$  pro každé  $\mathbf{s} \neq \mathbf{0}$ , nastává v bodě  $\mathbf{x}^* = \mathbf{0}$  ostré lokální minimum. To pro náš paraboloid skutečně platí.

## Přepis modelu trénovacích dat

Než aplikujeme předchozí metodu na minimalizaci RSS, přepíšme si celkový model pro naše trénovací data do maticového tvaru.

Trénovací data považujeme za náhodný výběr z uvažovaného lineárního modelu provedený v různých bodech  $\mathbf{x}_1, \dots, \mathbf{x}_N$ .

Máme tedy  $N$  párů  $(Y_i, \mathbf{x}_i)$ , kde  $Y_i = \mathbf{w}^T \mathbf{x}_i + \varepsilon_i$ .

Zavedme náhodné vektory  $\mathbf{Y} = (Y_1, \dots, Y_N)^T$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$  a body  $\mathbf{x}_1, \dots, \mathbf{x}_N$  zapišme v řádcích do matice  $\mathbf{X} \in \mathbb{R}^{N, p+1}$ ,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & x_{1;2} & \cdots & x_{1;p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & x_{N;2} & \cdots & x_{N;p} \end{pmatrix}.$$

Při tomto značení můžeme celkový model trénovací množiny zapsat jako

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},$$

kde  $\mathbf{E}\boldsymbol{\varepsilon} = (\mathbf{E}\varepsilon_1, \dots, \mathbf{E}\varepsilon_N)^T = \mathbf{0}$ .

## Minimalizace RSS (1/3)

RSS můžeme vyjádřit jako

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^N (Y_i - \mathbf{w}^T \mathbf{x}_i)^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2.$$

Aplikujme předchozí teorii na minimalizaci  $\text{RSS}(\mathbf{w})$ . Začneme parciálními derivacemi podle  $w_0, \dots, w_p$ ,

$$\frac{\partial \text{RSS}}{\partial w_j} = \sum_{i=1}^N 2(Y_i - \mathbf{w}^T \mathbf{x}_i)(-x_{i,j}).$$

Pro gradient tedy dostáváme

$$\nabla \text{RSS} = - \sum_{i=1}^N 2(Y_i - \mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{w}).$$

Položíme-li  $\nabla \text{RSS} = \mathbf{0}$ , získáme tzv. **normální rovnici** (angl. **normal equation**),

$$\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{0}.$$



## Minimalizace RSS (2/3)

Při výpočtu Hessovy matice použijeme

$$\frac{\partial^2 \text{RSS}}{\partial w_k \partial w_j} = \sum_{i=1}^N 2(-x_{i;k})(-x_{i;j}).$$

Hessova matice je tedy

$$\mathbf{H}_{\text{RSS}}(\mathbf{w}) = 2\mathbf{X}^T \mathbf{X}$$

bez ohledu na konkrétní hodnotu  $\mathbf{w}$ .

Dále pro každé  $\mathbf{s} \in \mathbb{R}^{p+1}$  platí

$$\mathbf{s}^T (\mathbf{X}^T \mathbf{X}) \mathbf{s} = (\mathbf{X} \mathbf{s})^T (\mathbf{X} \mathbf{s}) = \|\mathbf{X} \mathbf{s}\|^2 \geq 0.$$

Hessova matice  $\mathbf{H}_{\text{RSS}}(\mathbf{w})$  je tedy vždy pozitivně semi-definitní.

Podle předchozí věty proto nastává neostré lokální minimum v jakémkoliv bodě  $\mathbf{w}$ , který řeší normální rovnici

$$\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{0}.$$

## Minimalizace RSS (3/3)

Předpokládejme nyní, že  $\mathbf{X}^T \mathbf{X}$  je **regulární matice**.

Normální rovnice

$$\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{Y}$$

má potom jednoznačné řešení

$$\hat{\mathbf{w}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

kde značení vychází z anglického názvu **ordinary least squares** solution.

Pro matici  $\mathbf{X}$  a libovolný vektor  $\mathbf{s} \in \mathbb{R}^{p+1}$  snadno vidíme řetěz implikací

$$\mathbf{X} \mathbf{s} = \mathbf{0} \Rightarrow \mathbf{X}^T \mathbf{X} \mathbf{s} = \mathbf{0} \Rightarrow \mathbf{s}^T \mathbf{X}^T \mathbf{X} \mathbf{s} = 0 \Rightarrow \|\mathbf{X} \mathbf{s}\|^2 = 0 \Rightarrow \mathbf{X} \mathbf{s} = \mathbf{0}.$$

Z regularity  $\mathbf{X}^T \mathbf{X}$  tedy plyne, že pro nenulové  $\mathbf{s}$  nikdy nemůže platit  $\mathbf{s}^T (\mathbf{X}^T \mathbf{X}) \mathbf{s} = 0$ , a tedy Hessova matice je pozitivně definitní.

To znamená, že  $\hat{\mathbf{w}}_{\text{OLS}}$  je bodem ostrého lokálního minima.

Jak uvidíme později, v tomto bodě RSS nabývá dokonce **globálního minima**.

## Shrnutí metody nejmenších čtverců

- Model pro vysvětlovanou proměnnou  $Y$  v bodě  $\mathbf{x}$  je  $Y = \mathbf{w}^T \mathbf{x} + \varepsilon$ .
- Model pro trénovací množinu je  $\mathbf{Y} = \mathbf{X}\mathbf{w} + \varepsilon$ .
- Při trénování minimalizujeme residuální součet čtverců

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^N (Y_i - \mathbf{w}^T \mathbf{x}_i)^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2.$$

- Za předpokladu, že je matice  $\mathbf{X}^T \mathbf{X}$  regulární, existuje jediné řešení minimalizující  $\text{RSS}(\mathbf{w})$ ,

$$\hat{\mathbf{w}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- Predikce  $\hat{Y}$  v bodě  $\mathbf{x}$  je potom

$$\hat{Y} = \hat{\mathbf{w}}_{\text{OLS}}^T \mathbf{x} = \mathbf{x}^T \hat{\mathbf{w}}_{\text{OLS}} = \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

## Různé poznámky

- Lineární regrese je krásným příkladem diskriminativní metody, kdy přímo odhadujeme  $P(Y|\mathbf{X} = \mathbf{x})$ . Resp. přímo  $E(Y|\mathbf{X} = \mathbf{x})$ .
- Lineární regrese je ve vztahu ke problémům dimenzionality poměrně rezistentní.
- Důvodem je, že se jedná o parametrickou metodu. V ideálním případě nám tedy pro  $p$  příznaků +1 intercept může k určení přesného modelu stačit přesně  $p + 1$  bodů trénovací množiny.
- Problémy nastávají, když jsou v důsledku malé trénovací množiny nebo špatných příznaků, které jsou např. silně korelované, sloupce matice  $\mathbf{X}$  (skoro) lineárně závislé.
- Potom již nelze snadno provést inverzi matice  $\mathbf{X}^T \mathbf{X}$  a případně je numericky nestabilní.
- V důsledku se tento problém typicky projeví přeučením modelu, které znamená, že se model příliš přizpůsobí trénovací množině a nebude schopen dobře predikovat nové body.