Metodika analýzy dat: Od základů po aplikace metod strojového učení

Regresní modely

Jakub Steinbach, Jan Vrba

Ústav počítačové a řídicí techniky
VŠCHT

2.10.2024

Obsah slajdů I

Lineární regresní model

Vícenásobný regresní model

Definice

Vícenásobný lineární regresní model je definován jako

$$y^{(i)} = h(\mathbf{x}^{(i)}) + \varepsilon^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)} + \varepsilon^{(i)}$$

kde $\varepsilon^{(i)}$ je i-té residuum.
Funkce h se někdy uvádí ve tvaru

$$h(\mathbf{x}) = \sum_{i=0}^{n} \theta_i x_i$$

kde $x_0 = 1$ je tzv. dummy feature.

Vícenásobný regresní model - předpoklady

Předpoklady:

- očekáváná hodnota je lin. kombinací prediktorů, které mají aditivní účinky
- residua mají normální distribuci
- platí, že $E[\varepsilon^{(i)}] = 0$ pro $i = 1, \ldots, m$
- mají homogenní rozptyl, tzn. $D[\varepsilon^{(i)}] = \sigma^2 > 0$ pro $i = 1, \ldots, m$
- nejsou vzájemně korelovaná, tzn. $C(\varepsilon^{(i)}, \varepsilon^{(j)}) = 0$ pro $i \neq j$, $i, j = 1, \ldots, m$

Závěr:

Pokud jsou splněny všechny předpoklady, potom pro predikci platí

$$E[y^{(i)}] = h(\mathbf{x}^{(i)})$$

Odhad parametrů regresního modelu

Problém

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{m} (h(\mathbf{x}^{(i)}) - y^{(i)})^2$$

Metoda obyčejných nejmenších čtverců (ordinary least squares)

- odhad parametrů výpočtem

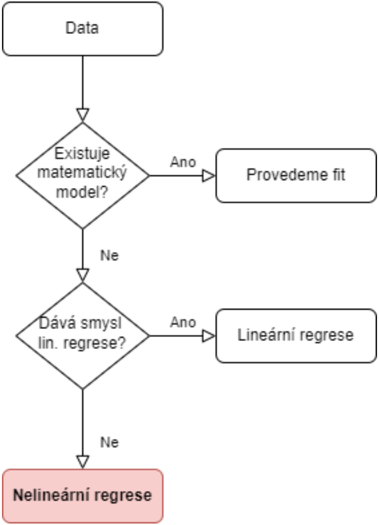
$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Obecně - optimalizační algoritmus

- iterační algoritmy (např. gradientní metody nebo Newtonova metoda)
- heuristické algoritmy
- simplexová metoda

Nelineární regresní analýza

Regresní analýza - volba modelu



Nelineární regrese - formulace problému

- uvažujeme data $\{y_i, x_1^{(i)}, \ldots x_n^{(i)}\}_{i=1}^m$
- stejně jako pro lineární regresi platí vztah

$$y_i = h(x_1^i, \ldots, x_n^i, \boldsymbol{\theta}) + \varepsilon^i$$

- výstup lineárního modelu není lineární kombinací prediktorů

$$h(\mathbf{c} + \boldsymbol{\theta}, \mathbf{x}) \neq h(\mathbf{c}, \mathbf{x}) + h(\boldsymbol{\theta}, \mathbf{x})$$

- odhad parametrů nelze obecně získat v uzavřeném tvaru (tj. $\theta = f(\mathbf{x},\mathbf{y})$ kde f je známá funkce)

Nelineární regrese - příklad

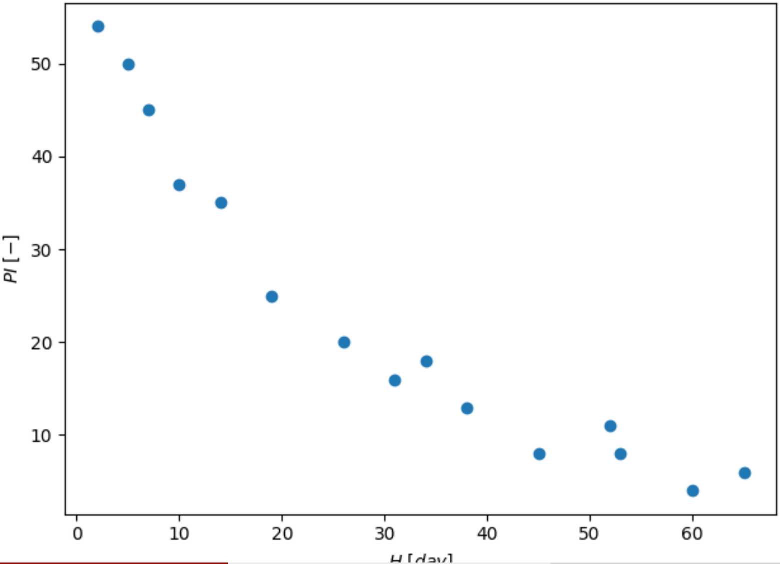
Využití linearizace - příklad

Příklad

Chceme namodelovat prognózu zotavení na základě délky pobytu v nemocnici. Očekáváme, pacienti po dlouhodobých pobytech v nemocnici budou mít obecně problémy s úplnou rekonvalescencí.

H	PI	H	PI
2	54	34	18
5	50	38	13
7	45	45	8
10	37	52	11
14	35	53	8
19	25	60	4
26	20	65	6
31	16		

Využití linearizace - příklad



Využití linearizace - příklad

- nelineární exponenciální regresní model

$$y^{(i)} = \theta_0 \exp(\theta_1 x^{(i)}) + \varepsilon_i$$

- linearizace exponenciálního regresního modelu

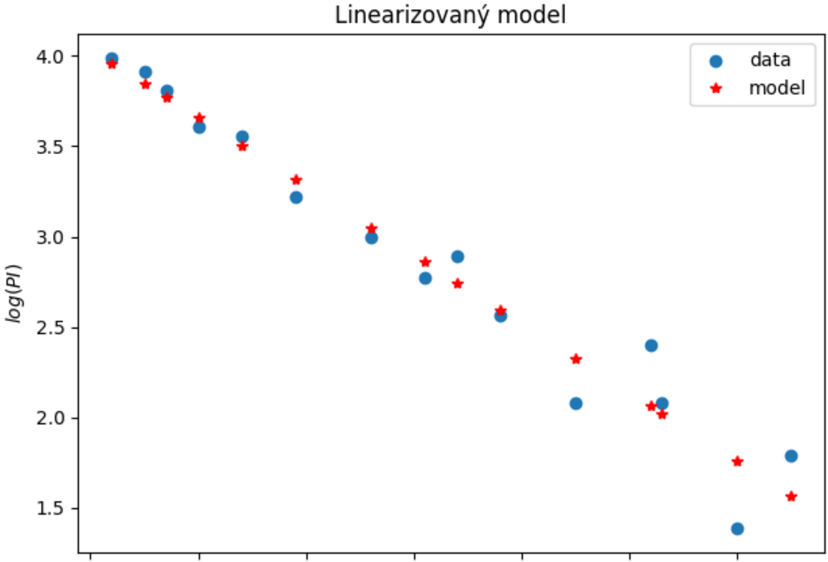
$$\log y^{(i)} = \log \theta_0 + \theta_1 x^{(i)}$$

- výpočet parametrů linearizovaného modelu

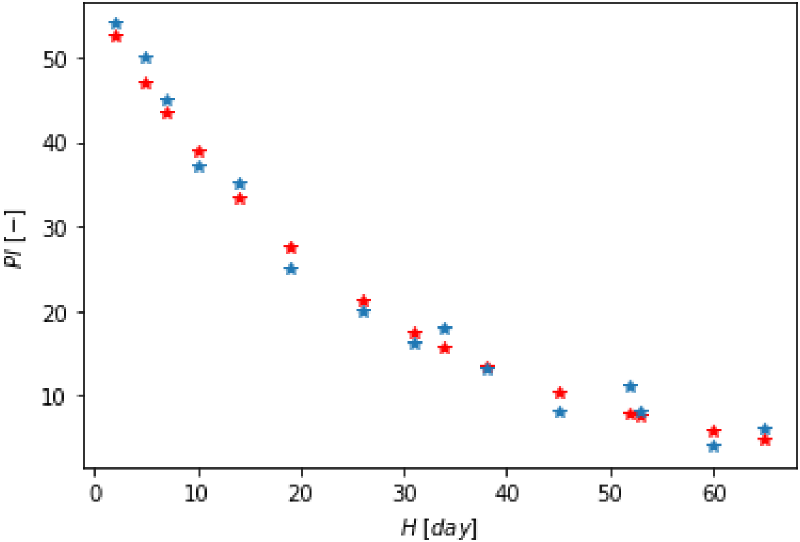
$$\boldsymbol{\theta}_{lin} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \log(\mathbf{Y})$$

$$\boldsymbol{\theta}_{lin} = [4.03715887, -0.03797418]$$

Linearizovaný model



Nelineární model s koeficienty linearizovaného modelu



Využití linearizace - příklad

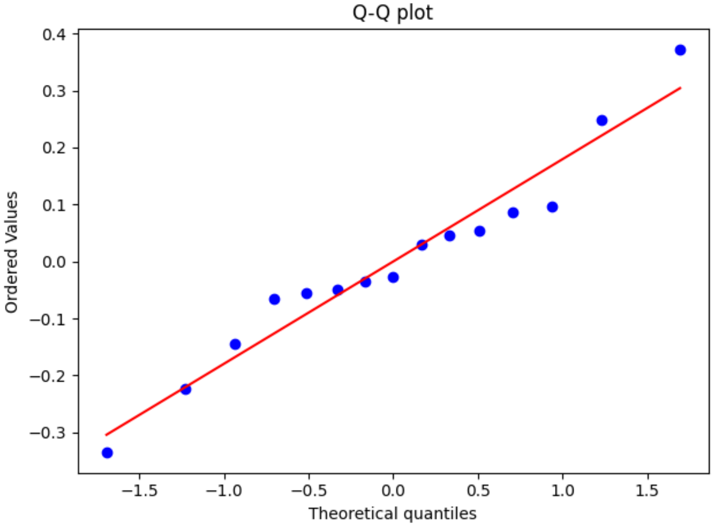
- výpočet reziduí

$$\varepsilon^{(i)} = y^{(i)} - h(\boldsymbol{\theta}, x^{(i)})$$

Test normality residuí:

- frekventistický test
  - D'Agostino-Pearson test (alespoň 20 vzorků)
  - Shapiro Wilks test (méně než 50 vzorků)
  - Kolmogorov-Smirnov test (více než 50 vzorků)
- grafickou metodou
  1. histogram
  2. boxplot
  3. QQ plot

Využití linearizace - příklad



Využití linearizace - příklad

- nepřesvědčivé výsledky normality residuí $\Longrightarrow$ hledání lepšího nelineárního modelu
- výpočet parametrů modelu

$$h(x, \boldsymbol{\theta}) = \theta_0 \exp \theta_1 x$$

- pro počáteční odhad parametrů nelineárního využijeme parametry linearizového modelu

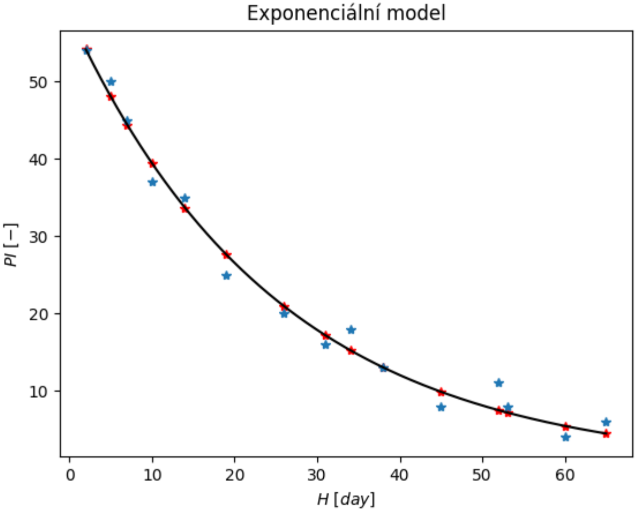
$$\theta_{0,init} = \exp\left(\theta_{0,lin}\right)$$

$$\theta_{1,init} = \theta_{1,lin}$$

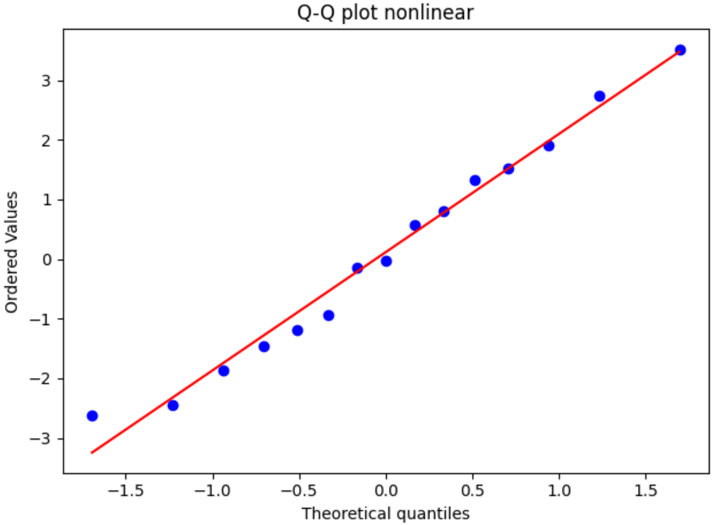
- výsledný odhad

$$\boldsymbol{\theta} = [58.6065651, -0.0395864508]$$

Využití linearizace - příklad



Využití linearizace - příklad



Využití linearizace - příklad

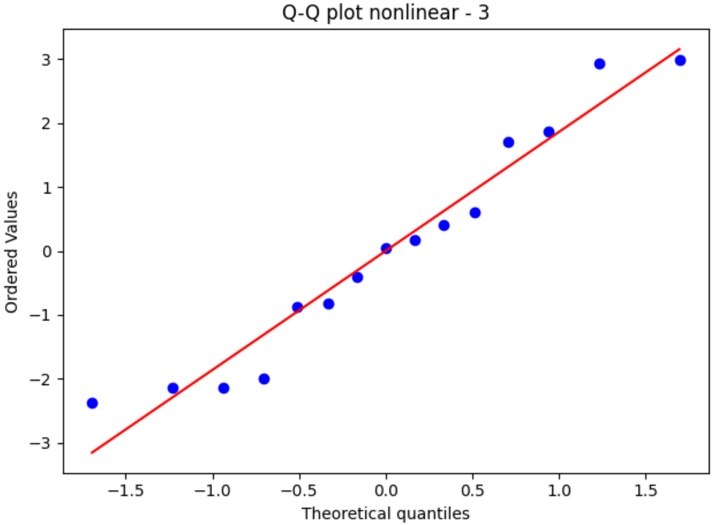
- volba dalšího modelu

$$h(x, \boldsymbol{\theta}) = \theta_0 \exp \theta_1 x + \theta_2$$

- výpočet parametrů modelu
- výsledný odhad

$$\boldsymbol{\theta} = [57.3320853, -0.0446038302, 2.43017740]$$

Využití linearizace - příklad



Akaikovo informační kritérium

- určení relativní kvality modelu (pro porovnání modelů mezi sebou)

$$AIC = 2k - 2\ln \hat{L}$$

- vhodné pro použití v případě že $\frac{m}{k} > 40$ (k - počet parametrů modelu)
- pro nízký počet naměřených dat

$$AICc = AIC + \frac{2k^2 + 2k}{m - k - 1}$$

- pro i.i.d. residua z nulovou střední hodnotou lze v případě, že k nalezení parametrů byla použita metoda LS určit AIC jako

$$AIC = 2k + n\ln RSS = 2k + n\ln\sum_{i=1}^{m}(y^i - h(\mathbf{x}^{(i)}, \boldsymbol{\theta}))^2$$

- nižší AIC $\implies$ lepší model

Porovnání modelů

model	RSS	AICc	$\sum_i \varepsilon_i$
lineární	56.08	65.4	3.73
exponenciální (2 parametry)	49.46	63.51	1.75
exponenciální (3 parametry)	44.78	64.66	$-2.74 \cdot 10^{-7}$

Monte Carlo pro nalezení konfidenčních intervalů parametrů

1. odhad parametrů modelu
2. výpočet standardní odchylky residuí

$$s_{x,y} = \sqrt{\frac{\sum_{i=1}^{m}(y^{(i)} - h(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}))^2}{m-k}}$$

3. vygenerování ideálního datasetu $\widetilde{y}^{(i)} = h(\mathbf{x}^{(i)}, \boldsymbol{\theta})$
4. ke každému ideálnímu bodu $\widetilde{y}_i$ přičteme náhodnou hodnotu z $\mathcal{N} \sim (0, s_{x,y})$
5. provedeme odhad parametrů modelu pro dataset získaný v kroku 4
6. opakujeme kroky 4 a 5 čímž získáme množinu parametrů modelu
7. nalezneme 2.5 a 97.5 hodnoty percentilu velikosti parametrů $\rightarrow$ konfidenční interval

Diskuze o nevhodnosti R^2

$$R^2 = 1 - \frac{RSS}{SST} = 1 - \frac{\sum_{i=1}^{m}(y^{(i)} - \widetilde{y}^{(i)})^2}{\sum_{i=1}^{m}(y^{(i)} - \bar{y})^2}$$

- pouze pro lineární modely
- při výběru modelu podle R^2 je vybrán nejlepší nelineární model v nejvýše 40% (studie)
- pro nelineární modely neplatí $var_{explained} + var_{err} = var_{total}$
- přeučený model má vysoké R^2
- nevypovídá o vhodnosti zvoleného regresního modelu
- nevíme jestli zvolené nezávislé proměnné ovlivňují závisle proměnnou

Regularizace regresních modelů

Regularizace regresních modelů - Lasso

- pro zabránění přeučení se často používá tzv. regularizace
- vhodné pro data na kterých model vykazuje velkou varianci mezi trénovacím a testovacím datasetem
- účelová funkce se rozšíří o další nenulový člen
- **L1 regularizace (Lasso)**

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} (y^{(i)} - h(\mathbf{x}^{(i)}, \boldsymbol{\theta}))^2 + \lambda \sum_{i=0}^{n} |\theta_i|, \ \lambda > 0$$

- optimální hodnoty parametrů $\boldsymbol{\theta}$ je nutné hledat iteračně
- pro některé proměnné může vyjít hodnota $\theta_i = 0$, tzn. že některé příznaky jsou z regresního modelu vynechány $\implies$ feature selection
- pro vícero silně korelovaných proměnných většinou vybere jednu (může být limitace)
- Lasso regrese v Pythonu

```
sklearn.linear_model.Lasso()
```

Regularizace regresních modelů - hřebenová regrese

- **L2 regularizace (Tichonova regularizace, ridge regression)**
- modifikace účelové funkce $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} (y^{(i)} - h(\mathbf{x}^{(i)}, \boldsymbol{\theta}))^2 + \lambda \sum_{i=0}^{n} \theta_i^2, \ \lambda > 0$$

- oproti Lasso regresi, existuje vztah pro výpočet optimálních parametrů

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X} - \lambda I)^{-1} \mathbf{X}^T Y)$$

- pro některé proměnné může vyjít hodnota $\theta_i = 0$, tzn. že některé příznaky jsou z regresního modelu vynechány $\implies$ feature selection
- Hřebenová regrese v Pythonu

```
sklearn.linear_model.Ridge()
```

Regularizace regresních modelů - elastic net

- **Regularizace typu elastic net**
- kombinuje LASSO a Ridge regresi
- modifikace účelové funkce $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m}(y^{(i)} - h(\mathbf{x}^{(i)}, \boldsymbol{\theta}))^2 + \lambda_2 \sum_{i=0}^{n} \theta_i^2 + \lambda_1 \sum_{i=0}^{n} |\theta_i|, \ \lambda_1 > 0, \lambda_2 > 0$$

- častá volba hyperparametrů $\lambda_2 = 0.5\alpha$, $\lambda_1 = 1 - \alpha$
- konvexní problém, eliminace problémů s konvergencí
- Elastic Net v Pythonu

```
sklearn.linear_model.ElasticNet()
```

Vážené nejmenší čtverce

Vážené nejmenší čtverce

- OLS předpokládají i.i.d. residua s konstantním rozptylem
- vážené nejmenší čtverce řeší problém nekorelovaných residuí s různým rozptylem a nulovou střední hodnotou
- kovarianční matice jednotlivých pozorování (měření)

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_m^2 \end{bmatrix}$$

Vážené nejmenší čtverce

- maximalizace věrohodnostní funkce

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \frac{1}{\sqrt{(2\pi)^m|\mathbf{C}|}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right)$$

- logaritmus věrohodnostní funkce

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y}$$

- hledáme $\frac{\partial loglikelihood}{\partial \boldsymbol{\theta}} = 0$

$$2\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y} = 0$$

- výsledný odhad parametrů $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y}$$

Vážené nejmenší čtverce

- inverze $\mathbf{C}^{-1}$

$$\mathbf{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_m^2} \end{bmatrix}$$

- někdy se matice $\mathbf{C}^{-1}$ označuje jako matice vah $\mathbf{W}$
- pro neznámou kovarianční matici $\mathbf{C}$ se nejprve provede fit pomocí LS a z výsledných residuí se odhadne $\mathbf{C}$, kde $w_i = \varepsilon_i^2$