Teorie informací a hodnocení statistických modelů

Bakalářská Práce

Tomáš Petit
505485@mail.muni.cz

Přírodovědecká fakulta, Masarykova Univerzita

27-06-2023

Motivace

Výběr modelů/pod-modelů v lineární regresi

- Výběr proměnných pro účel predikce v regresním modelu
- Snaha najít "spravedlivé" kritérium
- $y_i = \beta_0 + x_1\beta_1 + \ldots + x_m\beta_m + \varepsilon_i \quad i = 1, \ldots, n$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}$$

Cíle práce

- Představit v krátkosti základy teorie informace
- Poskytnout důkladné odvození Akaikeho informačního kritéria v plné obecnosti
- Předvést praktické využití pro širokou třídu statistických modelů

Obsah Práce

- Teorie Informace
  - Shannonova Entropie
  - Relativní Entropie
- Akaikeho informační kritétium
  - Metoda maximální věrohodnosti
  - Asymptotické vlastnosti věrohodnosti
  - Střední hodnota logaritmu věrohodnostní funkce
  - Vychýlení logaritmu věrohodnostní funkce
  - AIC
- Modelování pomocí AIC
  - Lineární a polynomiální regrese
  - Histogramy
  - Rovnost dvou diskrétních distribucí
  - Rovnost středních hodnot a rozptylů
  - Mallowsovo C_p
  - Analýza hlavních komponent

Kullback-Leiblerova divergence

Pravděpodobnostní distribuce P_X a Q_X náhodné veličiny X na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, \mu)$

$$I(P_X; Q_X) = E_{P_X}\left[\ln\left(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\right)\right] = \int_\Omega P_X \ln\left(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\right)\mathrm{d}P_X$$

- Není metrikou
- Množství informace ztracené při nahrazení P_X za Q_X
- Problém praktického výpočtu $\rightarrow$ potřeba odhadu

Expected likelihood

$$E_{f(x)}[\ln(g(x))] = \int_{-\infty}^{\infty} f(x)\ln(g(x))dx \tag{1}$$

$$= \sum_{\alpha=1}^{n} \hat{f}(x_\alpha)\ln(g(x_\alpha)) \tag{2}$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n} \ln(g(x_\alpha)) \tag{3}$$

Tedy

$$n\int_{-\infty}^{\infty} f(x)\ln(g(x))dx = \sum_{\alpha=1}^{n} \ln(g(x_\alpha)) \tag{4}$$

Informační kritéria

Takeuchiho informační kritérium

$$\text{TIC} = -2\sum_{\alpha=1}^{n} \ln f(X_\alpha|\hat{\boldsymbol{\theta}}) + 2\text{tr}\{J(\boldsymbol{\theta})I(\boldsymbol{\theta})^{-1}\} \tag{5}$$

Akaikeho informační kritérium

$$\text{AIC} = -2\sum_{\alpha=1}^{n} \ln f(X_\alpha|\hat{\boldsymbol{\theta}}) + 2p \tag{6}$$

Aplikace

- Hodnocení statistických modelů
  - Lineární regresní modely (polynomiální, ANOVA etc.)
  - PCA
  - Histogramy
  - Ekvivalence množin kategoriálních dat
  - Časové řady
  - Testování rovnosti středních hodnot

Příklad

Volně dostupná data **Swiss**

- Model 1 - Fertility $\sim$ Catholic
- Model 2 - Fertility $\sim$ Catholic + Education + Agriculture
- Model 3 - Fertility $\sim$ Catholic + Education + Agriculture + Infant.Mortality
- Model 4 - Full model

Model	**AIC**	R^2	R^2_{adj}
1	364.3479	0.215	0.1976
2	331.4126	0.6423	0.6173
3	325.2408	0.6993	0.6707
4	326.0716	0.7067	0.671

Table: Srovnání AIC, R^2 a R^2_{adj}

Shrnutí

- Základ v teorii maximální věrohodnosti
- Lze provést bodové odhady i intervaly spolehlivosti
- Lze použít jako základ pro statistiku

Děkuji za pozornost!

Otázka 1.1

Jaká je souvislost mezi p v (2.59) a k ve vzorci pro AICc?

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}$$

- Překlep

Otázka 1.2

V odstavci před Příkladem 3.1.2 zmiňujete, že automatizace výběru modelu se nedoporučuje jako hlavní nástroj. Jaký postup byste Vy, resp. použité zdroje, doporučil?

- Partial least squares
- LASSO
  - $\min_{\beta_0, \boldsymbol{\beta}}\{||y - \beta_0 - \boldsymbol{X}\boldsymbol{\beta}||_2^2\}$ za podmínky $||\boldsymbol{\beta}||_1 \leq t$
- Least Angle Regression
- Znalost dat a problematiky

Otázka 2.1

Ve třetí kapitole uvádíte dva přístupy výběru modelu pomocí AIC (forward-selection a backward-selection). Po použití obou těchto přístupů dostáváte stejný výsledný model. Funguje to tak vždy? Existuje ještě nějaký další přístup?

- Nemusí to vždy platit
- Bidirectional elimination

Otázka 2.2

Data z Příkladu 3.1.2 modelujete pomocí polynomiální regrese čtvrtého řádu. Je tento model pro uvedený datový soubor vhodný? Pokud ne, jaký jiný model byste použil?

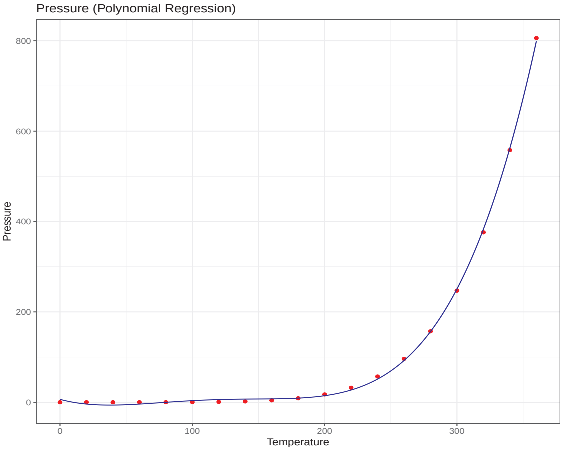

Pressure (Polynomial Regression)
800
600
400
200
0
Pressure
0
100
200
300
Temperature

Table: Výsledek regrese

	Dependent variable:
	pressure
temperature	-0.799^{***} $(-1.170, -0.428)$
temperature2	0.016^{***} $(0.012, 0.020)$
temperature3	-0.0001^{***} $(-0.0001, -0.0001)$
temperature4	0.00000^{***} $(0.00000, 0.00000)$
Constant	6.453 $(-2.650, 15.557)$
Observations	19
R^2	1.000
Adjusted R^2	0.999
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

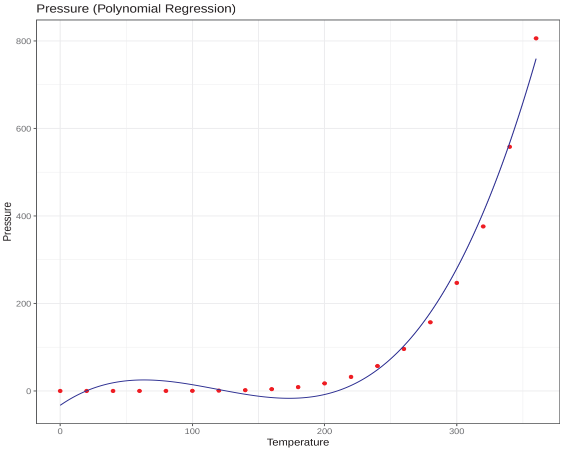
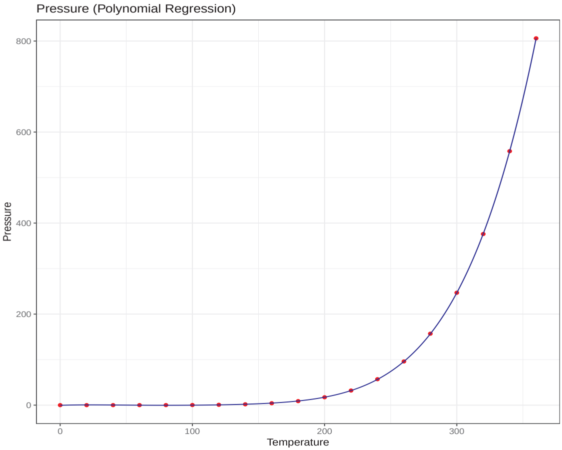

Pressure (Polynomial Regression)
Pressure
Temperature
800
600
400
200
0
0
100
200
300

Table: Výsledek regrese

	Dependent variable:
	pressure
temperature	2.086*** (1.139, 3.032)
temperature2	−0.022*** (−0.029, −0.016)
temperature3	0.0001*** (0.0001, 0.0001)
Constant	−32.847 (−71.135, 5.441)
Observations	19
R^2	0.989
Adjusted R^2	0.987
Note:	*p<0.1; **p<0.05; ***p<0.01



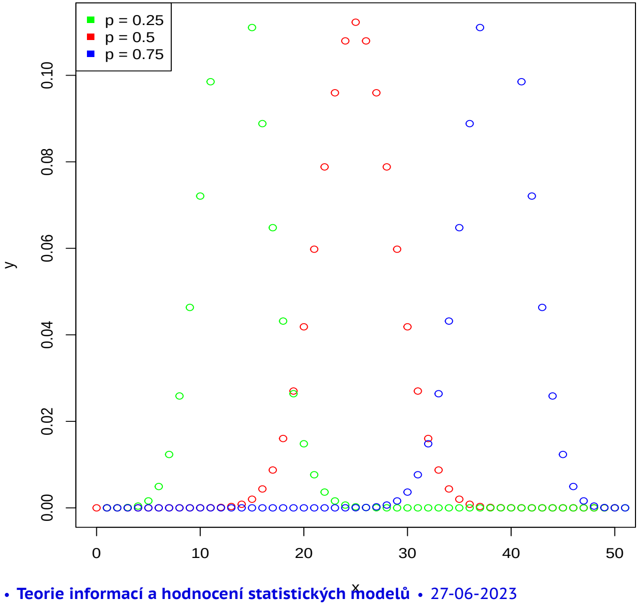
Pressure (Polynomial Regression)
800
600
400
200
0
Pressure
0
100
200
300
Temperature

Table: Výsledek regrese

	Dependent variable:
	pressure
temperature	0.106^{***} (0.062, 0.150)
temperature2	-0.004^{***} (-0.004, -0.003)
temperature3	0.00004^{***} (0.00004, 0.0001)
temperature4	-0.00000^{***} (-0.00000, -0.00000)
temperature5	0.000^{***} (0.000, 0.000)
Constant	-0.406 (-1.125, 0.312)
Observations	19
R^2	1.000
Adjusted R^2	1.000
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

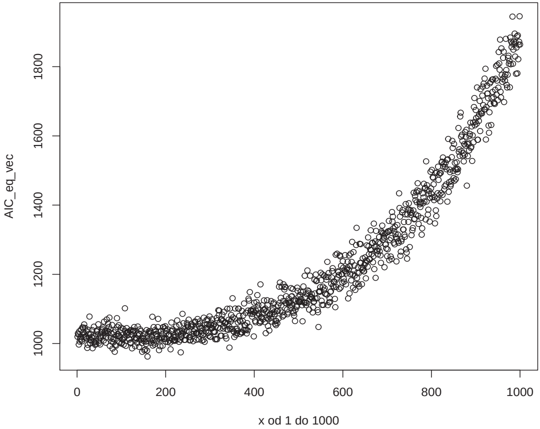
Otázka 2.3

Na str. 28 aproximujete data z normálního rozdělení binomickým rozdělením s parametrem $p = 1/2$. Proč právě $p = 1/2$ dává nejlepší aproximaci?


p = 0.25
p = 0.5
p = 0.75
y
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50
x

Otázka 2.4

Co způsobuje oscilaci na Obrázcích 3.6 a 3.7?


AIC_eq_vec
1000
1200
1400
1600
1800
0
200
400
600
800
1000
x od 1 do 1000

Otázka 2.5

Jak bychom mohli zlepšit odhady rozptylů v Tabulce 3.6, kde je počet pozorování $n = 10$.

MASARYK
UNIVERSITY