

**MASARYKOVA UNIVERZITA**  
**PŘÍRODOVĚDECKÁ FAKULTA**  
**ÚSTAV MATEMATIKY A STATISTIKY**

# **Bakalářská práce**

**BRNO 2023**

**TOMÁŠ PETIT**



# Teorie informace a hodnocení statistických modelů

Bakalářská práce

**Tomáš Petit**



# Bibliografický záznam

<b>Autor:</b>	Tomáš Petit Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
<b>Název práce:</b>	Teorie informace a hodnocení statistických modelů
<b>Studijní program:</b>	Matematika
<b>Studijní obor:</b>	Statistika a analýza dat
<b>Vedoucí práce:</b>	Mgr. Ondřej Pokora, Ph.D.
<b>Akademický rok:</b>	2022/2023
<b>Počet stran:</b>	vii + 53
<b>Klíčová slova:</b>	Teorie Informace; AIC; Informační kritéria; Akaikeho informační kritérium; Hodnocení statistických modelů; Výběr modelu; Jazyk R



# Bibliographic Entry

<b>Author:</b>	Tomáš Petit Faculty of Science, Masaryk University
<b>Title of Thesis:</b>	Information theory and evaluation of statistical models
<b>Degree Programme:</b>	Mathematics
<b>Field of Study:</b>	Statistics and data analysis
<b>Supervisor:</b>	Mgr. Ondřej Pokora, Ph.D.
<b>Academic Year:</b>	2022/2023
<b>Number of Pages:</b>	vii + 53
<b>Keywords:</b>	Information theory; AIC; Akaike information criterion; Evaluation of statistical models; Model selection; R language





# Abstrakt

V této bakalářské práci se věnujeme informačním kritériím a jejich použití při hodnocení různých statistických modelů, konkrétně se zaměřujeme na Akaikeho informační kritérium (AIC). Je zde provedeno důkladné teoretické odvození AIC. Věnujeme se také aplikacím na různých statistických modelech. Všechny numerické výpočty jsou řešeny pomocí jazyka R.

# Abstract

In this thesis we study information criteria and their application in the evaluation of different statistical models, specifically Akaike's information criterion (AIC). We rigorously derive the AIC and in the following chapter we show how to apply it on multiple statistical models applications. All the numerical results were obtained with the R statistical language.



ZADÁNÍ  
BAKALÁŘSKÉ PRÁCE

Akademický rok: 2022/2023

Ústav:	Ústav matematiky a statistiky
Student:	Tomáš Petit
Program:	Matematika
Specializace:	Statistika a analýza dat

Ředitel ústavu PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s názvem:

Název práce:	Teorie informace a hodnocení statistických modelů
Název práce anglicky:	Information theory and evaluation of statistical models
Jazyk závěrečné práce:	čeština

**Oficiální zadání:**

Cílem práce je podat přehled o základních pojmech z teorie informace a jejich aplikacích pro hodnocení statistických modelů. V práci budou mj. probrány teoretické aspekty Akaikeova kritéria AIC, včetně potřebných pojmů z teorie informace a matematické statistiky. Kromě teoretického popisu bude práce obsahovat i praktickou část s příklady využití AIC kritéria pro hodnocení statistických (např. regresních) modelů. Výpočty budou implementovány ve vhodném matematicko-statistickém softwaru.

**Literatura:**

COVER, T. M. a Joy A. THOMAS. *Elements of information theory*. 2nd ed. Hoboken, N.J.: Wiley-Interscience, 2006. xxiii, 748. ISBN 0471241954.

ASH, Robert B. *Information theory*. New York: Dover Publications, 1990. xi, 339. ISBN 0486665216.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer.

Vedoucí práce:	Mgr. Ondřej Pokora, Ph.D.
Datum zadání práce:	5. 5. 2022
V Brně dne:	30. 10. 2022

Zadání bylo schváleno prostřednictvím IS MU.

Tomáš Petit, 17. 10. 2022

Mgr. Ondřej Pokora, Ph.D., 20. 10. 2022

RNDr. Jan Vondra, Ph.D., 26. 10. 2022



# Poděkování

Na tomto místě bych chtěl poděkovat vedoucímu své práce, Mgr. Ondřeji Pokorovi, Ph.D., za veškeré přínosné rady a pomoc věnované při konzultacích. Dále bych chtěl poděkovat svým rodičům za podporu v mém studiu.

# Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracoval samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 31. března 2023

.....  
Tomáš Petit



# Obsah

<b>Úvod</b> .....	<b>1</b>
<b>Kapitola 1. Teorie informace</b> .....	<b>3</b>
1.1 Shannonova entropie .....	3
1.2 Relativní entropie .....	5
<b>Kapitola 2. Akaikeho informační kritérium</b> .....	<b>9</b>
2.1 Metoda maximální věrohodnosti .....	9
2.2 Asymptotické vlastnosti metody maximální věrohodnosti .....	11
2.3 Střední hodnota logaritmu věrohodnostní funkce .....	13
2.4 Vychýlení logaritmu věrohodnostní funkce .....	15
2.5 AIC .....	19
<b>Kapitola 3. Modelování pomocí AIC</b> .....	<b>23</b>
3.1 Lineární a polynomiální regrese .....	23
3.2 Histogramy .....	28
3.3 Rovnost dvou diskretních distribucí .....	30
3.4 Rovnost středních hodnot a rozptylů .....	32
3.5 Mallowsovo $C_p$ .....	38
3.6 Analýza hlavních komponent .....	40
<b>Závěr</b> .....	<b>47</b>
<b>Příloha</b> .....	<b>49</b>
<b>Seznam použité literatury</b> .....	<b>51</b>





# Úvod

Teorie informace je jedním z matematických oborů, o které se dá říct, že vznikla na díle jednoho autora. Její počátky můžeme sledovat již v roce 1924 v článku *Certain Factors Affecting Telegraph Speed* [21] Harryho Nyquista nebo v článku, který vyšel 4 roky po něm, *Transmission of Information* [13] Ralpa Hartleyho, kde je poprvé užito slovo *informace* ve smyslu technickém, tj. měřitelnou kvantitu, jejíž množství definoval jako  $H = \log(S)^n$ , kde  $S$  bylo množství všech možných symbolů v signálu a  $n$  počet symbolů v signálu.

Článek který stal za zrodem teorie informace jako samostatným matematickým oborem je článek z roku 1948 Clauda Shannona *A Mathematical Theory of Communication* [24] (později přejmenovaný na *The Mathematical Theory of Communication*). V tomto článku se objevuje definice informační entropie jakožto míry neurčitosti pro diskrétní i spojitě náhodné veličiny (entropie samotná je koncept známý ze statistické mechaniky již v 19. století) spolu s názvem pro její jednotku (pro logaritmus o základu 2) – *bity*. Jak název napovídá, článek se primárně zajímal o komunikaci mezi dvěma a vícero kanály a množstvím informace, které lze poslat jako signál mezi nimi.

Klíčová pro nás je definice Kullback-Leiblerovy divergence v článku [18] publikovaném v roce 1951. Její použití je klíčové nejen pro informační kritéria celkově, ale i například v podoboru diferenciální geometrie známém jako informační geometrie (anglicky *information geometry*). Už při odvození Kullbackovy-Leiblerovy divergence je vidět nejen silná spojitost s teorií maximální věrohodnosti, ale i myšlenka, že bychom ji mohli použít pro ohodnocení statistického modelu.

Japonský statistik Hirotugu Akaike formuloval „*an information criterion*” na sympoziu roku 1971, kdy sborník byl publikován až v roce 1973, [2]. Jednalo se pouze o neformální prezentaci a Akaikeho odborný článek vyšel následně o rok později, [3]. Dnes se jedná o jeden z nejvíce citovaných článků v historii. Akaikeho původní článek pracoval s poměrně silnými předpoklady, a proto japonský statistik Takeuchi v roce 1976, [27], ukázal, že tyto předpoklady lze zeslabit a dosáhneme požadovaného výsledku. Vzhledem k tomu, že Takeuchiho článek vyšel pouze v japonštině, tak se mu nedostalo příliš velké pozornosti. Už poměrně brzy pak začaly vycházet další články, které se věnovaly AIC a jeho vlastnostem – například [25] a [15].

Obecný přehled, který pomohl popularizovat problematiku selekce a hodnocení modelů pomocí informačních kritérií, byl *Model Selection and Multimodel Inference: A practical information-theoretic approach* [8] autorů Burnham & Anderson, který vyšel v roce 2002. Od té doby vyšlo bohaté množství publikací na tuto tématiku s aplikacemi v mnoha různých oborech - například [9], [17], [1], [29] nebo [16].

Cílem této práce je pochopitelně nejen odvodit AIC a ukázat jeho využití u různých statistických modelů, ale hlavně seznámit čtenáře s informačně-teoretickým přístupem a

ukázat, že se nejedná o nic jiného než rozšíření teorie maximální věrohodnosti, která je obecně známa.

# Kapitola 1

## Teorie informace

### 1.1 Shannonova entropie

Zavedeme koncept entropie jako míry neurčitosti pro náhodné veličiny.

**Definice 1.1.1.** Buď  $X$  diskrétní náhodná veličina s oborem hodnot  $M$  a pravděpodobnostní funkcí  $p(x)$  pro  $x \in M$ . Entropie diskrétní náhodné veličiny se definuje jako

$$H(X) = - \sum_{x \in M} p(x) \log p(x) \quad (1.1)$$

Podobně lze rozšířit definici entropie pro spojité náhodné veličiny.

**Definice 1.1.2.** Buď  $X$  spojitá náhodná veličina s hustotou  $f(x)$ . Potom se její entropie definuje jako

$$h(f) = - \int_S f(x) \log f(x) dx \quad (1.2)$$

kde  $S$  je nosičem hustoty, neboli  $S = \{x \in \mathbb{R} \mid f(x) > 0\}$ . Tuto veličinu nazýváme diferenciální entropií.

Budeme se držet konvence, že  $0 \log 0 = 0$ , kterou lze lehce odůvodnit existencí limity  $\lim_{x \rightarrow 0^+} x \log x = 0$ . Použijeme-li logaritmus o základu  $b$ , tak označíme entropii jako  $H_b$  pro přehlednost. Dokud není jinak řečeno, tak používáme logaritmus o základu 2, kdy jednotky se jmenují *bity*. Pro přirozený logaritmus se jednotky nazývají *naty*.

**Příklad 1.1.1.** Uvažujme klasickou šestihrannou hrací kostku. Předpokládáme, že kostka je spravedlivá a nechť má rovnoměrné pravděpodobnostní rozdělení, tj. platí, že  $p(x) = \frac{1}{6}$  pro  $x \in M$ , kde  $M = \{1, \dots, 6\}$ . Potom se její entropie rovná

$$H(X) = - \sum_{x \in M} \frac{1}{6} \log \frac{1}{6} = 2,584 \text{ bitů} \quad (1.3)$$

Nyní můžeme popsat některé základní vlastnosti entropie diskrétní náhodné veličiny.

**Lemma 1.1.1.** Buď  $X$  diskrétní náhodná veličina a  $p(x)$  její pravděpodobnostní funkce. Pak platí následující:

$$(i) H(X) \geq 0$$

$$(ii) H_b(X) = \log_b(a) H_a(X)$$

*Důkaz.* (i) Je zřejmé, že pro  $x \in M$  platí, že  $0 \leq p(x) \leq 1$  a zároveň lze přepsat  $-\log p(x)$  na  $\log \frac{1}{p(x)}$ . Ve výsledku sčítáme nezáporná čísla a tedy  $H(X) \geq 0$ .

(ii) Dosazením získáváme

$$H(X)_b = - \sum_{x \in M} p(x) \log_b p(x) = - \sum_{x \in M} \log_b(a) p(x) \log_a p(x) = \log_b(a) H_a(X).$$

□

Bod (i) nemusí platit pro spojité náhodné veličiny, jak dokazuje následující příklad.

**Příklad 1.1.2.** Mějme spojitou náhodnou veličinu  $X$  s rovnoměrným rozdělením na intervalu  $[0, a]$  a hustotou  $f(x) = \frac{1}{a}$  na uvedeném intervalu a  $f(x) = 0$  jinde. Její diferenciální entropie je tedy

$$h(f) = - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a. \quad (1.4)$$

Je zřejmé, že pro  $a < 1$  je  $\log a < 0$  a tedy diferenciální entropie je záporná.

**Příklad 1.1.3.** Nechť diskrétní náhodná veličina je definována jako

$$X = \begin{cases} 1 & \text{s pravděpodobností } \theta, \\ 0 & \text{s pravděpodobností } 1 - \theta. \end{cases}$$

Pak její entropie je

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta). \quad (1.5)$$

Je lehké vidět, že v případě, kdy se  $\theta = 0$  nebo  $1$ , tak máme entropii rovnu  $0$ . To lze interpretovat tak, že když ve veličině není žádná náhoda, tak pochopitelně v ní nemůže být ani žádná neurčitost. Naopak největší neurčitosti dosáhneme s  $\theta = \frac{1}{2}$ , kdy máme entropii rovnu  $1$  bitu.

**Příklad 1.1.4 (Normální rozdělení).** Nechť  $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ , neboli normální rozdělení s parametry  $\mu = 0$  a  $\sigma = 1$ . Pak její diferenciální entropie v natech je

$$\begin{aligned} h(f) &= - \int_{-\infty}^{\infty} \phi \ln \phi dx \\ &= - \int_{-\infty}^{\infty} \phi(x) \left[ -\frac{x^2}{2\sigma^2} - \ln(\sqrt{2\pi\sigma^2}) \right] dx \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \\ &= \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2), \end{aligned} \quad (1.6)$$

kde  $E[X^2]$  je druhý moment veličiny  $X$ . Pokud by nás zajímal případ se střední hodnotou  $\mu \neq 0$ , tak se jedná pouze o výpočet diferenciální entropie pro funkci  $\phi(x - \mu)$  a tedy

$$\begin{aligned} h(f) &= - \int_{-\infty}^{\infty} \phi(x - \mu) \ln \phi(x - \mu) dx \\ &= - \int_{-\infty}^{\infty} \phi(x) \ln \phi(x) dx \\ &= \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2). \end{aligned} \quad (1.7)$$

Vidíme, že entropie jednorozměrného normálního rozdělení nezávisí na jeho střední hodnotě, ale pouze na jeho rozptylu.

**Příklad 1.1.5.** Nechť  $X$  je náhodná veličina s exponenciálním rozdělením a hustotou  $f(x) = \lambda e^{-\lambda x}$  pro  $x \geq 0$  a 0 jinde. Její diferenciální entropie je

$$\begin{aligned} h(f) &= - \int_0^{\infty} \lambda e^{-\lambda x} \log(\lambda e^{-\lambda x}) dx \\ &= - \left( \int_0^{\infty} (\log \lambda) \lambda e^{-\lambda x} dx + \int_0^{\infty} (-\lambda x) \lambda e^{-\lambda x} dx \right) \\ &= -\log \lambda \int_0^{\infty} f(x) dx + \lambda E[X] \\ &= -\log \lambda + 1 \text{ natů.} \end{aligned} \quad (1.8)$$

## 1.2 Relativní entropie

V mnoha úlohách z pravděpodobnosti a statistiky potřebujeme vybrat mezi dvěma a více pravděpodobnostními distribucemi a vybrat tu „nejlepší“ z nich. Pro definování kvality výběru se často používá nějaká míra vzdálenosti mezi pravděpodobnostními distribucemi (za předpokladu, že je známe). Jednou z takových měr vzdálenosti je *relativní entropie* neboli *Kullbackova-Leiblerova divergence*. V dalších definicích se budeme držet konvence značení v [17].

**Definice 1.2.1.** Pro pravděpodobnostní distribuce  $P_X$  a  $Q_X$  náhodné veličiny  $X$  definované na pravděpodobnostním prostoru  $(\Omega, \mathcal{F})$  se Kullbackova-Leiblerova divergence (dále K-L informace) definuje jako

$$I(P_X; Q_X) = E_{P_X} \left[ \log \left\{ \frac{dP_X}{dQ_X} \right\} \right], \quad (1.9)$$

kde  $\frac{dP_X}{dQ_X}$  je Radanova-Nikodymova derivace  $P_X$  vzhledem k  $Q_X$  a  $E_{P_X}$  je střední hodnota vzhledem k distribuci  $P_X$ , tj.

$$E_{P_X} \left[ \log \left\{ \frac{dP_X}{dQ_X} \right\} \right] = \int_{\Omega} \log \left\{ \frac{dP_X}{dQ_X} \right\} dP_X. \quad (1.10)$$

Pro dvě hustoty  $f(x)$  a  $g(x)$  spojité náhodné veličiny  $X$  lze K-L informaci zapsat jako

$$I(g; f) = \int_{-\infty}^{\infty} \log \left\{ \frac{g(x)}{f(x)} \right\} g(x) dx. \quad (1.11)$$

Naopak pro dvě pravděpodobnostní funkce diskrétní náhodné veličiny  $X$  máme

$$I(p; q) = \sum_{i=1}^{\infty} p(x_i) \log \left\{ \frac{p(x_i)}{q(x_i)} \right\}. \quad (1.12)$$

*Poznámka.* V obou případech používáme konvenci  $0 \log \frac{0}{0} = 0$  a dále uvažujeme, že  $0 \log \frac{0}{f} = 0$  a  $g \log \frac{g}{0} = 0$ .

Důvod, proč lze použít výraz K-L informace je takový, že tuto veličinu můžeme chápat jako množství informace, které ztratíme, když namísto rozdělení  $P_X$  použijeme  $Q_X$ , respektive, když nahradíme  $P_X$  za  $Q_X$ . Tato intuice bude důležitá pro pochopení toho, co vlastně Akaikeho informační kritérium vyjadřuje.

Definici K-L informace si můžeme přiblížit následujícím způsobem. Z matematické statistiky je známo, že Neymannovo-Pearsonovo lemma nás informuje o existenci testů hypotéz s největší silou na zvolené  $\alpha$  úrovni, a jak takové testy konstruovat. Velice často konstrukce probíhá pomocí podílu logaritmu věrohodnostních funkcí, tj. máme-li náhodnou veličinu  $Y$  a testujeme například hypotézy  $H_0 : \mu = \mu_0$  a  $H_1 : \mu = \mu_1$ , zkoumali bychom rozdíl  $\log(P_Y) - \log(Q_Y)$ , kdy  $P_Y$  a  $Q_Y$  jsou pravděpodobnostní distribuce s parametry  $\mu_0$  a  $\mu_1$ , respektive. Pak K-L informace je střední hodnotou tohoto rozdílu vzhledem k  $P_Y$  (viz [8]).

Je lehké si ověřit, že K-L informace není symetrická. Podíváme-li se na rozdíl například dvou hustot, tak máme

$$I(g; f) - I(f; g) = \int_{-\infty}^{\infty} (f(x) + g(x)) \log \left\{ \frac{g(x)}{f(x)} \right\} dx,$$

a je jasné, že pravá strana nulová být nemusí. Následující lemma nám říká, v jakém případě je K-L informace nulová.

**Věta 1.2.1.** K-L informace má následující vlastnosti:

- (i)  $I(g; f) \geq 0$ ,
- (ii)  $I(g; f) = 0 \iff g(x) = f(x)$ .

*Důkaz.* Definujme funkci  $K(t) = \log t - t + 1$  pro všechna  $t > 0$ . Její derivace  $K'(t) = \frac{1}{t} - 1$  splňuje podmínky  $K'(1) = 0$  a  $K(t)$  dosahuje svého maxima  $K(1) = 0$  v  $t = 1$ . Tedy pro všechna  $t > 0$  platí nerovnost  $K(t) \leq 0$  a rovnost je splněna pouze v případě, kdy se  $t = 1$ . Z toho lze odvodit, že vztah

$$\log t \leq t - 1 \quad (1.13)$$

platí. Pro spojitý případ uvažujme substituci  $t = \frac{f(x)}{g(x)}$  a dostaneme

$$\log \frac{f(x)}{g(x)} \leq \frac{f(x)}{g(x)} - 1. \quad (1.14)$$

Vynásobením obou stran  $g(x)$  a integrováním získáme

$$\int_{-\infty}^{\infty} \log \left\{ \frac{f(x)}{g(x)} \right\} g(x) dx \leq \int_{-\infty}^{\infty} \left\{ \frac{f(x)}{g(x)} - 1 \right\} g(x) dx, \quad (1.15)$$

kdy se pravá strana rovná

$$\int_{-\infty}^{\infty} f(x) dx - \int_{-\infty}^{\infty} g(x) dx = 0. \quad (1.16)$$

Ve výsledku dostáváme

$$\int_{-\infty}^{\infty} \log \left\{ \frac{g(x)}{f(x)} \right\} g(x) dx = - \int_{-\infty}^{\infty} \log \left\{ \frac{f(x)}{g(x)} \right\} g(x) dx \geq 0, \quad (1.17)$$

a tím je dokázána vlastnost (i). Tato rovnost je splněna pouze v případě, kdy se  $f(x) = g(x)$  a tímto je důkaz hotov. Diskrétní případ se dokazuje analogicky.  $\square$

Stejně tak lze ověřit, že K-L informace nesplňuje trojúhelníkovou nerovnost, a tedy se nejedná o metriku, ale pouze o divergenci.

**Příklad 1.2.1.** Mějme prostor  $M = \{0, 1\}$  a na něm definovaná tři diskrétní pravděpodobnostní rozdělení pomocí

$$\begin{aligned} p(x) &= \begin{cases} \frac{1}{2} & \text{pro } x = 0, \\ \frac{1}{2} & \text{pro } x = 1, \end{cases} \\ q(x) &= \begin{cases} \frac{1}{4} & \text{pro } x = 0, \\ \frac{3}{4} & \text{pro } x = 1, \end{cases} \\ z(x) &= \begin{cases} \frac{1}{10} & \text{pro } x = 0, \\ \frac{9}{10} & \text{pro } x = 1, \end{cases} \end{aligned}$$

Dosazením do vztahu

$$I(p; z) \leq I(p; q) + I(q; z)$$

dostáváme přibližné hodnoty

$$0,51 \leq 0,14 + 0,09,$$

které viditelně nesplňují trojúhelníkovou nerovnost.

Existuje mnoho dalších různých měř vzdálenosti mezi dvěma distribucemi. Například

$$\chi^2(g; f) = \sum_{i=1}^k \frac{g_i^2}{f_i^2} - 1 \quad \text{je tzv. } \chi^2 \text{ statistika,} \quad (1.18)$$

$$I_\lambda(g; f) = \frac{1}{\lambda} \int_{-\infty}^{\infty} \left\{ \left( \frac{g(x)}{f(x)} \right)^\lambda - 1 \right\} g(x) dx \quad \text{je tzv. zobecněná informace,} \quad (1.19)$$

$$D(g; f) = \int_{-\infty}^{\infty} u(x) \left( \frac{g(x)}{f(x)} \right) g(x) dx \quad \text{je divergence.} \quad (1.20)$$

Je zřejmé, že volbou  $u(x) = \log x$  u divergence  $D(g; f)$  dostaneme K-L informaci. Stejně tak volba  $u(x) = \lambda^{-1}(x^\lambda - 1)$  nám dá zobecněnou informaci  $I_\lambda(g; f)$ . A pokud u  $I_\lambda(g; f)$  necháme  $\lambda \rightarrow 0$ , tak tato limita bude rovna K-L informaci. Nyní si uvedeme další dva příklady.

**Příklad 1.2.2.** Předpokládejme, že dvě hrací kostky mají následující pravděpodobnostní funkce hodů čísel od 1 do 6:

$$p(x) = \{0, 2; 0, 15; 0, 15; 0, 16; 0, 24; 0, 1\}, \quad (1.21)$$

$$q(x) = \{0, 11; 0, 09; 0, 22; 0, 18; 0, 25; 0, 15\}. \quad (1.22)$$

Zajímá nás, která z těchto kostek je spravedlivější? Víme, že spravedlivá má rovnoměrné rozdělení, které můžeme popsat jako  $z(x) = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$ . Nižší hodnota K-L informace nám řekne, které z rozdělení je blíže pravému rozdělení popsané pomocí  $z$ , a tedy která kostka je spravedlivější. Když spočítáme K-L informaci pomocí

$$I(z; p) = \sum_{i=1}^6 g_i \log \frac{z_i}{p_i}, \quad (1.23)$$

dostáváme, že  $I(z; p) = 0,517$  a  $I(z; q) = 0,090$  a tedy kostka s rozdělením  $q(x)$  je spravedlivější.

**Příklad 1.2.3.** Mějme dvě exponenciální rozdělení o parametrech  $\lambda$  a  $\gamma$  s hustotami

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x \geq 0, \\ 0 & \text{pro } x < 0, \end{cases}$$

a

$$f(x) = \begin{cases} \gamma e^{-\gamma x} & \text{pro } x \geq 0, \\ 0 & \text{pro } x < 0. \end{cases}$$

Pak K-L informace hustoty  $f(x)$  s ohledem na  $g(x)$  v natech je

$$\begin{aligned} I(g; f) &= \int_0^\infty g(x) \ln \left\{ \frac{g(x)}{f(x)} \right\} dx \\ &= \int_0^\infty (\lambda e^{-\lambda x}) \ln \left\{ \frac{\lambda e^{-\lambda x}}{\gamma e^{-\gamma x}} \right\} dx \\ &= \int_0^\infty (\lambda e^{-\lambda x}) \left\{ \ln \left\{ \frac{\lambda}{\gamma} \right\} + x(\gamma - \lambda) \right\} dx \\ &= \ln \left\{ \frac{\lambda}{\gamma} \right\} \int_0^\infty \lambda e^{-\lambda x} dx + (\gamma - \lambda) \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \ln \left\{ \frac{\lambda}{\gamma} \right\} + \frac{\gamma - \lambda}{\lambda} \text{ natů.} \end{aligned} \quad (1.24)$$



# Kapitola 2

## Akaikeho informační kritérium

### 2.1 Metoda maximální věrohodnosti

Po zbytek práce bude log označovat přirozený logaritmus, pokud nebude třeba specifikovat, o jaký základ se jedná.

**Definice 2.1.1.** Nechť  $\Theta$  je parametrický prostor hodnot  $p$ -rozměrného vektoru parametrů  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ . Označme sdruženou hustotu pravděpodobnosti náhodného vektoru  $\mathbf{X}$  s pravděpodobnostní funkcí  $f(x)$  (nebo hustotu) následovně

$$L(\boldsymbol{\theta}; x_1, \dots, x_n) = L(\theta_1, \dots, \theta_p; x_1, \dots, x_n) \quad (2.1)$$

$$= \prod_{i=1}^n f(x_i; \boldsymbol{\theta}), \quad (2.2)$$

a nazveme ji věrohodnostní funkcí náhodného výběru. Odhad  $\hat{\boldsymbol{\theta}}_{MLE}$  nazveme maximálně věrohodným, jestliže pro každé  $\boldsymbol{\theta} \in \Theta$  platí

$$L(\hat{\boldsymbol{\theta}}_{MLE}; x_1, \dots, x_n) \geq L(\boldsymbol{\theta}; x_1, \dots, x_n). \quad (2.3)$$

Často je mnohem jednodušší pracovat s logaritmem funkce  $L(\boldsymbol{\theta}; x_1, \dots, x_n)$ , který budeme značit jako  $l(\boldsymbol{\theta}; x_1, \dots, x_n)$ . Za předpokladů, že logaritmus věrohodnostní funkce je spojitě diferencovatelný a Hessova matice je negativně definitní, maximum věrohodnostní funkce, a tedy  $\hat{\boldsymbol{\theta}}_{MLE}$ , lze najít jako řešení soustavy rovnic

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, p. \quad (2.4)$$

**Příklad 2.1.1.** Mějme  $n$  množin dat  $\{y_\alpha, x_{\alpha,1}, x_{\alpha,2}, \dots, x_{\alpha,p}\}$  pro závislou proměnnou  $y$  a  $p$  nezávislých proměnných  $\{x_1, x_2, \dots, x_p\}$ . Uvažujme následující lineární normální regresní model

$$y_\alpha = \mathbf{x}_\alpha^T \boldsymbol{\beta} + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \quad \alpha = 1, 2, \dots, n, \quad (2.5)$$

kde  $\mathbf{x}_\alpha = (1, x_{\alpha,1}, x_{\alpha,2}, \dots, x_{\alpha,p})^T$  a  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ . Víme, že hustota lineárního regresního modelu v proměnné  $y_\alpha$  je

$$f(y_\alpha | \mathbf{x}_\alpha; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} (y_\alpha - \mathbf{x}_\alpha^T \boldsymbol{\beta})^2 \right\}. \quad (2.6)$$

Pak můžeme jednoduše vyjádřit logaritmus věrohodnostní funkce jako

$$\begin{aligned}
 l(\boldsymbol{\theta}) &= \sum_{\alpha=1}^n \log f(y_{\alpha} | \mathbf{x}_{\alpha}; \boldsymbol{\theta}) \\
 &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{\alpha=1}^n (y_{\alpha} - \mathbf{x}_{\alpha}^T \boldsymbol{\beta})^2 \\
 &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}), \tag{2.7}
 \end{aligned}$$

kde  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$  a  $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$ . Zderivujeme-li 2.7 vzhledem k parametru  $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)$ , tak dostaneme

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} (-2X^T \mathbf{y} + 2X^T X \boldsymbol{\beta}) = \mathbf{0}, \tag{2.8}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) = 0, \tag{2.9}$$

kde  $\mathbf{0}$  je nulový vektor. Řešením této soustavy rovnic nám poskytuje odhady pro  $\boldsymbol{\beta}$  a  $\sigma^2$  ve tvaru

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}, \quad \hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T (\mathbf{y} - X\hat{\boldsymbol{\beta}}). \tag{2.10}$$

Z teorie lineárních regresních modelů je známo, že  $\hat{\sigma}^2$  v tomto tvaru není nestranným odhadem, narozdíl od  $\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T (\mathbf{y} - X\hat{\boldsymbol{\beta}})$ . To nicméně není v našem případě problém, jelikož, jak se dozvíme v další sekci, tento odhad splňuje asymptotickou nestrannost.

**Příklad 2.1.2.** Mějme náhodnou veličinu  $X$  s normálním rozdělením  $N(\mu, \sigma^2)$  a soubor  $n$  dat  $\{x_1, x_2, \dots, x_n\}$ . Logaritmus věrohodnostní funkce pro tento model je

$$l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{\alpha=1}^n (x_{\alpha} - \mu)^2, \tag{2.11}$$

kde zderivováním dle obou parametrů dostaneme

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{\alpha=1}^n (x_{\alpha} - \mu) = 0$$

a

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\alpha=1}^n (x_{\alpha} - \mu)^2 = 0.$$

Řešením těchto jsou maximálně věrohodné odhady pro  $\mu$  a  $\sigma^2$  ve tvaru

$$\hat{\mu} = \frac{1}{n} \sum_{\alpha=1}^n x_{\alpha}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{\alpha=1}^n (x_{\alpha} - \mu)^2.$$

Vidíme, že na těchto dvou vybraných modelech jsme dosáhli explicitního vzorce pro naše odhady, nicméně ve většině případů takové štěstí nemáme a musíme použít numerických metod pro aproximaci řešení dané soustavy rovnic. Stručně zde popíšeme jednu metodu takové aproximace. Začneme s počáteční aproximací parametrického vektoru  $\boldsymbol{\theta}$  a postupně budeme vytvářet posloupnost  $\{\boldsymbol{\theta}_i\}_{i=1}^{\infty}$ , která bude konvergovat k hodnotě  $\hat{\boldsymbol{\theta}}$ . Začneme tedy s Taylorovým rozvojem  $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$  v okolí  $\boldsymbol{\theta}_k$ ,

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \frac{\partial l(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 l(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} (\boldsymbol{\theta} - \boldsymbol{\theta}_k). \quad (2.12)$$

Dále označme

$$g(\boldsymbol{\theta}) = \left\{ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_2}, \dots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right\}^T,$$

$$H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \left\{ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\}, \quad i, j = 1, 2, \dots, p.$$

Pro  $\boldsymbol{\theta}$ , které splňují  $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$  dostáváme

$$\mathbf{0} = g(\boldsymbol{\theta}) \approx g(\boldsymbol{\theta}_k) + H(\boldsymbol{\theta}_k)(\boldsymbol{\theta} - \boldsymbol{\theta}_k), \quad (2.13)$$

kde  $g(\boldsymbol{\theta}_k)$  je vektor gradientu a  $H(\boldsymbol{\theta}_k)$  je Hessova matice. Ze vztahu 2.13 je patrné, že  $\boldsymbol{\theta} \approx \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1} g(\boldsymbol{\theta}_k)$ . Získali jsme tedy posloupnost ve tvaru

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1} g(\boldsymbol{\theta}_k). \quad (2.14)$$

Tato metoda se nazývá *Newtonova-Raphsonova metoda*, o které je známo, že konverguje rychle v blízkosti kořenu, pokud máme vhodně zvolenou počáteční aproximaci.

## 2.2 Asymptotické vlastnosti metody maximální věrohodnosti

V rámci této sekce se podíváme blíže na asymptotické vlastnosti metody maximální věrohodnosti a na vlastnosti jejího odhadu. Zároveň ukážeme spojitost s K-L informací a blíže prozkoumáme jejich vztah. Následující věty a jejich důkazy najdeme v [17].

**Věta 2.2.1.** Předpokládejme, že následující podmínky regularity platí pro hustotu  $f(x|\boldsymbol{\theta})$ :

- (1) Funkce  $\log f(x|\boldsymbol{\theta})$  je třikrát spojitě diferencovatelné s ohledem na  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ .
- (2) Existují integrovatelné funkce  $A(x)$ ,  $B(x)$  a  $C(x)$  takové, že

$$\int_{-\infty}^{\infty} C(x) f(x|\boldsymbol{\theta}) dx < M, \quad (2.15)$$

pro vhodnou hodnotu konstanty  $M$  a jsou splněny následující nerovnosti pro všechny  $\boldsymbol{\theta} \in \Theta$ :

$$\left| \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \right| < A(x), \quad \left| \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| < B(x), \quad (2.16)$$

a

$$\left| \frac{\partial^3 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| < C(x), \quad i, j, k = 1, 2, \dots, p.$$

(3) Následující nerovnost je splněna pro jakékoliv  $\boldsymbol{\theta} \in \Theta$ :

$$0 < \int_{-\infty}^{\infty} f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} dx < \infty, \quad i, j = 1, \dots, p. \quad (2.17)$$

Dále předpokládejme, že  $\boldsymbol{\theta}_0$  je řešením rovnice

$$\int_{-\infty}^{\infty} f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx = \mathbf{0}, \quad (2.18)$$

a že množina dat  $\mathbf{x}_n = \{x_1, x_2, \dots, x_n\}$  je generována náhodnou veličinou mající hustotu  $f(x|\boldsymbol{\theta})$ . Poté platí následující vlastnosti:

(i) Věrohodnostní rovnice

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{\alpha=1}^n \frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \quad (2.19)$$

má řešení, které konverguje k  $\boldsymbol{\theta}_0$ .

(ii) Maximálně věrohodný odhad  $\hat{\boldsymbol{\theta}}_n$  konverguje v pravděpodobnosti k  $\boldsymbol{\theta}_0$  pro  $n \rightarrow \infty$ .

(iii) Maximálně věrohodný odhad  $\hat{\boldsymbol{\theta}}_n$  je asymptoticky normální, tedy rozdělení veličiny  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  konverguje k  $p$ -rozměrnému normálnímu rozdělení  $N_p(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$ , kde  $I(\boldsymbol{\theta}_0)$  je hodnota matice  $I(\boldsymbol{\theta})$  v  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ , která je daná vztahem

$$I(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx. \quad (2.20)$$

Tato matice se nazývá *Fisherova informační matice*.

V předchozí větě jsme předpokládali existenci  $\boldsymbol{\theta}_0 \in \Theta$ , které splňuje náš předpoklad  $g(x) = f(x|\boldsymbol{\theta}_0)$ . Pokud takové  $\boldsymbol{\theta}_0$  neexistuje, lze odvodit podobné podmínky.

**Věta 2.2.2.** Nechť tedy  $\boldsymbol{\theta}_0$  je řešením

$$\int_{-\infty}^{\infty} g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx = \mathbf{0} \quad (2.21)$$

a mějme množinu dat  $\mathbf{x}_n = \{x_1, x_2, \dots, x_n\}$ , která je generovaná veličinou mající hustotu  $g(x)$ . Za použití předpokladů 1–3 z věty 2.2.1 platí následující tvrzení pro odhad  $\hat{\boldsymbol{\theta}}_n$ :

(i) Maximálně věrohodný odhad  $\hat{\boldsymbol{\theta}}_n$  konverguje v pravděpodobnosti k  $\boldsymbol{\theta}_0$  pro  $n \rightarrow \infty$ .

(ii) Rozdělení veličin  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  konverguje k  $p$ -rozměrnému normálnímu rozdělení  $N_p(\mathbf{0}, J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0))$  pro  $n \rightarrow \infty$ , kde  $I(\boldsymbol{\theta}_0)$  a  $J(\boldsymbol{\theta}_0)$  jsou  $p \times p$  matice vyhodnocené v  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$  a jsou dané následujícími rovnicemi

$$I(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx \quad (2.22)$$

$$= \left( \int_{-\infty}^{\infty} g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} dx \right). \quad (2.23)$$

$$J(\boldsymbol{\theta}) = - \int_{-\infty}^{\infty} g(x) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} dx \quad (2.24)$$

$$= - \left( \int_{-\infty}^{\infty} g(x) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} dx \right), \quad i, j = 1, 2, \dots, p. \quad (2.25)$$

Důkaz tohoto tvrzení můžeme najít v [17]. Podívejme se blíže na vztah matic  $I(\boldsymbol{\theta})$  a  $J(\boldsymbol{\theta})$ . Začneme s následující řadou rovností

$$\begin{aligned} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_i} \left\{ \frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta}) \right\} \\ &= \frac{\partial}{\partial \theta_i} \left\{ \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta}) \right\} \\ &= \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) - \frac{1}{f(x|\boldsymbol{\theta})^2} \frac{\partial}{\partial \theta_i} f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta}) \\ &= \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) - \frac{\partial}{\partial \theta_i} \log f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta}). \end{aligned} \quad (2.26)$$

Nyní se podívejme na střední hodnotu 2.26 obou stran s ohledem na rozdělení  $P_X$

$$E_{P_X} \left[ \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) \right] - E_{P_X} \left[ \frac{\partial}{\partial \theta_i} \log f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta}) \right].$$

Lze vidět, že obecně  $I(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$ , ovšem pokud existuje vektor  $\boldsymbol{\theta}_0 \in \Theta$  takový, že  $g(x) = f(x|\boldsymbol{\theta}_0)$ , pak máme

$$\begin{aligned} E_{P_X} \left[ \frac{1}{f(x|\boldsymbol{\theta}_0)} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}_0) \right] &= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}_0) dx \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_{-\infty}^{\infty} f(x|\boldsymbol{\theta}_0) dx = 0, \end{aligned}$$

a z toho plyne rovnost  $I_{ij}(\boldsymbol{\theta}_0) = J_{ij}(\boldsymbol{\theta}_0)$  pro  $i, j = 1, 2, \dots, p$ , a tedy  $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ .

## 2.3 Střední hodnota logaritmu věrohodnostní funkce

Vraťme se nyní zpět ke K-L informaci. Víme, že ji lze použít k evaluaci korektnosti statistického modelu (viz Příklad 1.2.1). Nicméně tento přístup je limitovaný tím, že ve většině případů neznáme skutečnou distribuci svého modelu. Rozepišme K-L informaci jako

$$I(P_X; Q_X) = E_{P_X} \left[ \log \left\{ \frac{P_X}{Q_X} \right\} \right] = E_{P_X} [\log P_X] - E_{P_X} [\log Q_X], \quad (2.27)$$

první člen na pravé straně je konstantní, jelikož závisí pouze na našem skutečném rozdělení  $P_X$ . Proto pokud chceme porovnávat modely, je přirozené se podívat na druhý člen. Tento člen je střední hodnota logaritmu věrohodnostní funkce (anglicky *expected log-likelihood*).

Je jasné, že čím větší bude jeho hodnota, tím menší bude K-L informace, a tím přesnější bude náš model. Vyjádřením této veličiny v následující podobě

$$\begin{aligned} E_{P_X}[\log Q_X] &= \int_{\Omega} \log Q_X dP \\ &= \int_{-\infty}^{\infty} g(x) \log f(x) dx \quad \text{pro spojitou distribuci,} \\ &= \sum_{i=1}^{\infty} p(x_i) \log q(x_i) \quad \text{pro diskretní distribuci,} \end{aligned} \quad (2.28)$$

je patrné, že je stále závislá na skutečném rozdělení  $P_X$ , a proto ji nelze spočítat explicitně. Přirozeně se nabízí tuto veličinu aproximovat nějakým odhadem. Mějme tedy  $\mathbf{x}_n = \{x_1, x_2, \dots, x_n\}$  množinu dat generovaných skutečným rozdělením  $P_X$ . Jako odhad neznámého rozdělení  $P_X$  v (2.28) lze použít empirické rozdělení  $\hat{P}_X$ . Empirické pravděpodobnostní rozdělení má pravděpodobnostní funkci ve tvaru  $\hat{p}(x_\alpha) = 1/n$  pro  $\alpha = 1, 2, \dots, n$ . Dosazením do 2.28 dostaneme

$$\begin{aligned} E_{\hat{P}_X}[\log f(X)] &= \int_{\Omega} \log f(x) d\hat{P} \\ &= \sum_{i=1}^n \hat{p}(x_\alpha) \log f(x_\alpha) \\ &= \frac{1}{n} \sum_{i=1}^n \log f(x_\alpha). \end{aligned} \quad (2.29)$$

Aplikací zákona velkých čísel dospějeme k tomu, že s rostoucím počtem dat  $n \rightarrow \infty$  konverguje výběrový průměr náhodné veličiny  $Y_\alpha = \log f(X_\alpha)$ , pro  $\alpha = 1, 2, \dots, n$ , v pravděpodobnosti k její střední hodnotě, tj.

$$\frac{1}{n} \sum_{i=1}^n \log f(X_\alpha) \rightarrow E_{P_X}[\log f(X)] \quad \text{pro } n \rightarrow \infty. \quad (2.30)$$

Tímto jsme získali odhad pro střední hodnotu logaritmu věrohodnostní funkce. Přenásobením odhadu  $n$  dostaneme

$$n \int_{\Omega} \log f(x) d\hat{P} = \sum_{i=1}^n \log f(x_\alpha). \quad (2.31)$$

Toto implikuje, že logaritmus věrohodnostní funkce slouží jako odhad K-L informace, a proto ji lze použít k definici kritéria pro posuzování statistických modelů.

V rámci této práce nás bude zajímat hodnocení statistických modelů s ohledem na predikci, a nikoliv inferenci. Budeme postupovat stejně jako dosud, tj. chceme odhadnout kvalitu modelu  $f(z|\hat{\theta})$ , který použijeme pro predikci nezávislých budoucích dat  $Z = z$  generované neznámou skutečnou distribucí  $g(z)$ . Stejným způsobem dojdeme k tomu, že odhad pro střední hodnotu logaritmu věrohodnostní funkce je

$$E_{\hat{G}}[\log f(Z|\hat{\theta})] = \int_{\Omega} \log f(z|\hat{\theta}) d\hat{G}(z) \quad (2.32)$$

$$= \frac{1}{n} \sum_{\alpha=1}^n \log f(x_\alpha|\hat{\theta}), \quad (2.33)$$

kde  $\mathbf{x}_n = \{x_1, x_2, \dots, x_n\}$  jsou data, která máme k dispozici v době vytváření statistického modelu. Analogicky dojdeme k tomu, že odhad střední hodnoty  $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$  je  $n^{-1}l(\hat{\boldsymbol{\theta}})$  a že logaritmus věrohodnostní funkce  $l(\hat{\boldsymbol{\theta}})$  je odhadem  $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ .

## 2.4 Vychýlení logaritmu věrohodnostní funkce

Předpokládejme, že máme množinu modelů  $\{f_j(z|\boldsymbol{\theta}_j); j = 1, 2, \dots, m\}$  a pro parametry  $\boldsymbol{\theta}_j$  mějme jejich příslušné odhady  $\hat{\boldsymbol{\theta}}_j$ . Z předešlé diskuze bychom mohli říct, že kvalita modelu se dá jednoznačně určit na základě hodnoty maxima  $l_j(\hat{\boldsymbol{\theta}}_j)$ , respektive porovnáním těchto hodnot mezi modely. Bohužel tento přístup nám nedává spravedlivé srovnání modelů, jelikož veličina  $l_j(\hat{\boldsymbol{\theta}}_j)$  je vychýleným odhadem střední hodnoty logaritmu věrohodnostní funkce  $nE_G[\log f_j(z|\hat{\boldsymbol{\theta}}_j)]$  a velikost vychýlení závisí na počtu parametrů.

Tento poznatek se může zdát na první pohled poněkud zvláštní, když tvrdíme, že  $l(\boldsymbol{\theta})$  je dobrým odhadem  $nE_G[\log f(Z|\boldsymbol{\theta})]$ . Nicméně pokud se podíváme zpátky na 2.33, vidíme, že jsme odvodili logaritmus věrohodnostní funkce pomocí jeho střední hodnoty na základě dat, která jsme ale použili i k odhadu vektoru parametrů. Toto opakované použití té samé množiny dat k odhadu parametrů a střední hodnoty věrohodnostní funkce nám zde vytváří vychýlení (anglicky *bias*). Podívejme se na vztah mezi logaritmem věrohodnostní funkce a jeho střední hodnotou pro model  $f(x|\theta)$  s jednorozměrným parametrem  $\theta$ . Skutečný parametr  $\theta_0$  je dán jako maximum střední hodnoty logaritmu věrohodnostní funkce. Na druhou stranu, maximálně věrohodný odhad  $\hat{\theta}(\mathbf{x}_n)$  je hodnota, která je maximem logaritmu věrohodnostní funkce  $l(\theta)$ . Kvalitu modelu chceme posoudit dle hodnoty  $E_G[\log f(Z|\hat{\theta})]$ , kterou ovšem vyhodnocujeme pomocí  $l(\hat{\theta})$ , kterou lze vypočítat na základě našich dat. Kritérium pro srovnání modelů musí tedy dávat  $E_G[\log f(Z|\hat{\theta})] \leq E_G[\log f(Z|\theta_0)]$ . Pro logaritmus věrohodnostní funkce ovšem vždy platí nerovnost opačná, tj.  $l(\hat{\theta}) \geq l(\theta_0)$ .

Přestože hodnoty logaritmu věrohodnostní funkce jsou závislé na našich datech, přechází dvě nerovnosti vždy platí. Chceme-li korektně definovat kritérium na srovnání modelů, potřebujeme přidat míru vychýlení (nějakou formu penalizace) do své definice, abychom kompenzovali tyto nerovnosti.

**Definice 2.4.1.** Předpokládejme, že  $n$  pozorování  $\mathbf{x}_n$  generovaná skutečnou distribucí  $G(x)$  nebo  $g(x)$ , jsou realizace náhodné veličiny  $\mathbf{X}_n = (X_1, X_2, \dots, X_n)^T$  a necht

$$l(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^n \log f(x_\alpha|\hat{\boldsymbol{\theta}}(\mathbf{x}_n)) = \log f(\mathbf{x}_n|\hat{\boldsymbol{\theta}}(\mathbf{x}_n)) \quad (2.34)$$

reprezentuje logaritmus věrohodnostní funkce statistického modelu  $f(z|\hat{\boldsymbol{\theta}}(\mathbf{x}_n))$ . Vychýlení logaritmu věrohodnostní funkce jako odhad jeho střední hodnoty je definováno jako

$$b(G) = E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n|\hat{\boldsymbol{\theta}}(\mathbf{X}_n)) - nE_{G(z)} \left[ \log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}_n)) \right] \right], \quad (2.35)$$

kde střední hodnota  $E_{G(\mathbf{x}_n)}$  je brána s ohledem na sdruženou distribuci  $\prod_{\alpha=1}^n G(x_\alpha) = G(\mathbf{x}_n)$  vektoru realizací  $\mathbf{X}_n$  a  $E_{G(z)}$  je střední hodnota skutečné distribuce  $G(z)$ .

Vidíme, že obecná forma informačního kritéria je

$$\begin{aligned} IC(\mathbf{X}_n; \hat{G}) &= -2(\text{logaritmus věrohodnostní funkce modelu} - \text{odhad vychýlení}) \\ &= -2 \sum_{\alpha=1}^n \log f(X_\alpha | \hat{\boldsymbol{\theta}}) + 2\{\text{odhad pro } b(G)\}. \end{aligned} \quad (2.36)$$

Obecně může  $b(G)$  nabývat vícero forem v závislosti na vztahu mezi skutečnou distribucí a modelem a na metodě použité ke konstrukci statistického modelu. My se budeme hlavně zajímat a odvodíme informační kritérium pro statistické modely založené na metodě maximalizace věrohodnostní funkce.

Nyní si odvodíme explicitní vyjádření vychýlení pro případ, kdy střední hodnota logaritmu věrohodnostní funkce je odhadována pomocí logaritmu věrohodnostní funkce daného modelu. Všimněme si, že pokud pro  $\boldsymbol{\theta}$ , které je řešením soustavy rovnic

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{\alpha=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_\alpha | \boldsymbol{\theta}) = \mathbf{0}, \quad (2.37)$$

vezmeme střední hodnotu této soustavy, tak získáme

$$E_{G(\mathbf{x}_n)} \left[ \sum_{\alpha=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_\alpha | \boldsymbol{\theta}) \right] = n E_{G(z)} \left[ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z | \boldsymbol{\theta}) \right]. \quad (2.38)$$

Proto pro náhodné veličiny  $X$  spojitého typu platí, že pokud je  $\boldsymbol{\theta}_0$  řešením rovnice

$$E_{G(z)} \left[ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z | \boldsymbol{\theta}) \right] = \int_{-\infty}^{\infty} g(z) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(z | \boldsymbol{\theta}) dz = \mathbf{0}, \quad (2.39)$$

tak maximálně věrohodný odhad  $\hat{\boldsymbol{\theta}}$  konverguje v pravděpodobnosti k  $\boldsymbol{\theta}_0$  pro  $n \rightarrow \infty$ . Pro diskrétní modely se postupuje analogicky. Za použití předchozích výsledků můžeme odhadnout vychýlení

$$b(G) = E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n | \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) - n E_{G(z)} \left[ \log f(Z | \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) \right] \right]. \quad (2.40)$$

Práce bude jednodušší, pokud si 2.40 rozdělíme do tří částí, které postupně vypočítáme, tj.

$$\begin{aligned} b(G) &= E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n | \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) - \log f(\mathbf{X}_n | \boldsymbol{\theta}_0) \right] \\ &\quad + E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n | \boldsymbol{\theta}_0) - n E_{G(z)} \left[ \log f(Z | \boldsymbol{\theta}_0) \right] \right] \\ &\quad + E_{G(\mathbf{x}_n)} \left[ n E_{G(z)} \left[ \log f(Z | \boldsymbol{\theta}_0) \right] - n E_{G(z)} \left[ \log f(Z | \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) \right] \right] \\ &= D_1 + D_2 + D_3. \end{aligned}$$



Začněme s výpočtem  $D_2$ . Vzhledem k tomu, že  $D_2$  neobsahuje žádné odhady, tak je jeho výpočet nejjednodušší, tj.

$$\begin{aligned} D_2 &= E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n | \boldsymbol{\theta}_0) - n E_{G(z)} [\log f(Z | \boldsymbol{\theta}_0)] \right] \\ &= E_{G(\mathbf{x}_n)} \left[ \sum_{\alpha=1}^n \log f(X_\alpha | \boldsymbol{\theta}_0) - n E_{G(z)} [\log f(Z | \boldsymbol{\theta}_0)] \right] \\ &= 0, \end{aligned} \quad (2.41)$$

kde jsme použili vztah 2.38.

Výpočet složky  $D_3$  už je o něco náročnější. Nejprve si zavedem pomocnou funkci

$$\eta(\hat{\boldsymbol{\theta}}) := E_{G(z)} [\log f(Z | \hat{\boldsymbol{\theta}})]. \quad (2.42)$$

Podíváme-li se na Taylorův rozvoj funkce  $\eta(\hat{\boldsymbol{\theta}})$  v okolí  $\boldsymbol{\theta}_0$  (které je řešením 2.39), tak dostaneme

$$\begin{aligned} \eta(\hat{\boldsymbol{\theta}}) &= \eta(\boldsymbol{\theta}_0) + \sum_{i=1}^p (\hat{\theta}_i - \theta_i^{(0)}) \frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} \\ &\quad + \frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p (\hat{\theta}_i - \theta_i^{(0)}) (\hat{\theta}_j - \theta_j^{(0)}) \frac{\partial^2 \eta(\boldsymbol{\theta}_0)}{\partial \theta_i \partial \theta_j} + \dots, \end{aligned} \quad (2.43)$$

kde  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)^T$  a  $\boldsymbol{\theta}_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})^T$ . Díky tomu, že  $\boldsymbol{\theta}_0$  je řešením rovnice 2.39, tak platí že

$$\frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} = E_{G(z)} \left[ \frac{\partial}{\partial \theta_i} \log f(Z | \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} \right] = 0, \quad i = 1, 2, \dots, p, \quad (2.44)$$

kdy  $\frac{\partial}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_0}$  značí parciální derivaci vyhodnocenou v bodě  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . Můžeme aproximovat 2.43 jako

$$\eta(\hat{\boldsymbol{\theta}}) = \eta(\boldsymbol{\theta}_0) - \frac{1}{2} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (2.45)$$

kde matice  $J(\boldsymbol{\theta}_0)$  je  $p \times p$  matice daná vztahem

$$J(\boldsymbol{\theta}_0) = -E_{G(z)} \left[ \frac{\partial^2 \log f(Z | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}_0} \right] = - \int_{-\infty}^{\infty} g(z) \frac{\partial^2 \log f(z | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}_0} dz, \quad (2.46)$$

jejíž  $(i, j)$  prvek je

$$j_{ij} = -E_{G(z)} \left[ \frac{\partial^2 \log f(Z | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \Big|_{\boldsymbol{\theta}_0} \right] = - \int_{-\infty}^{\infty} g(z) \frac{\partial^2 \log f(z | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \Big|_{\boldsymbol{\theta}_0} dz.$$

Jelikož  $D_3$  je vlastně střední hodnota rozdílu  $\eta(\boldsymbol{\theta}_0) - \eta(\hat{\boldsymbol{\theta}})$  s ohledem na  $G(\mathbf{x}_n)$ ,

získáváme

$$\begin{aligned}
 D_3 &= E_{G(\mathbf{x}_n)} \left[ nE_{G(z)} [\log f(Z|\boldsymbol{\theta}_0)] - nE_{G(z)} [\log f(Z|\hat{\boldsymbol{\theta}})] \right] \\
 &= \frac{n}{2} E_{G(\mathbf{x}_n)} \left[ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \right] \\
 &= \frac{n}{2} E_{G(\mathbf{x}_n)} \left[ \text{tr} \left\{ J(\boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \right\} \right] \\
 &= \frac{n}{2} \text{tr} \left\{ J(\boldsymbol{\theta}_0) E_{G(\mathbf{x}_n)} \left[ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \right] \right\}. \tag{2.47}
 \end{aligned}$$

Dosazením (asymptotické) kovarianční matice (viz Věta 2.2.2, bod (ii)) maximálně věrohodného odhadu  $\hat{\boldsymbol{\theta}}$ ,

$$E_{G(\mathbf{x}_n)} [(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T] = \frac{1}{n} J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}, \tag{2.48}$$

do rovnice 2.47, dostaneme

$$D_3 = \frac{1}{2} \text{tr} \{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \}, \tag{2.49}$$

kde matici  $J(\boldsymbol{\theta}_0)$  známe z 2.46 a matice  $I(\boldsymbol{\theta}_0)$  je  $p \times p$  matice

$$I(\boldsymbol{\theta}_0) = E_{G(z)} \left[ \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \bigg|_{\boldsymbol{\theta}_0} \right].$$

Nyní už nám zbývá vyjádřit člen  $D_1$ , u kterého se bude postupovat podobně jako u  $D_3$ . Označíme-li si  $l(\boldsymbol{\theta}) = \log f(\mathbf{X}_n|\boldsymbol{\theta})$  a aplikujeme-li Taylorův rozvoj v okolí maximálně věrohodného odhadu  $\hat{\boldsymbol{\theta}}$ , tak dostaneme

$$l(\boldsymbol{\theta}) = l(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + \frac{1}{2} \frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \dots \tag{2.50}$$

$\hat{\boldsymbol{\theta}}$  je pochopitelně řešením rovnice  $\frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$  z důvodu, že jako maximálně věrohodný odhad je řešením  $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ .

Můžeme si všimnout, že hodnota

$$\frac{1}{n} \frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{1}{n} \frac{\partial^2 \log f(\mathbf{X}_n|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{2.51}$$

konverguje v pravděpodobnosti k  $J(\boldsymbol{\theta}_0)$  v 2.46 pro  $n$  jdoucí do nekonečna. Lehce se to dá odůvodnit tím, že maximálně věrohodný odhad  $\hat{\boldsymbol{\theta}}$  konverguje k  $\boldsymbol{\theta}_0$  a pomocí bodu (3.63) v [17]. Kombinací těchto výsledků máme aproximaci

$$l(\boldsymbol{\theta}_0) - l(\hat{\boldsymbol{\theta}}) \approx -\frac{n}{2} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0) (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \tag{2.52}$$

pro vztah 2.50. Spolu s kovarianční maticí z 2.48 a tímto výsledkem můžeme konečně vyjádřit  $D_1$  jako

$$\begin{aligned}
 D_1 &= E_{G(\mathbf{x}_n)} \left[ \log f(\mathbf{X}_n | \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) - \log f(\mathbf{X}_n | \boldsymbol{\theta}_0) \right] \\
 &= \frac{n}{2} E_{G(\mathbf{x}_n)} \left[ (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0) (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \right] \\
 &= \frac{n}{2} E_{G(\mathbf{x}_n)} \left[ \text{tr} \left\{ J(\boldsymbol{\theta}_0) (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \right\} \right] \\
 &= \frac{n}{2} \text{tr} \left\{ J(\boldsymbol{\theta}_0) E_{G(\mathbf{x}_n)} \left[ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \right] \right\} \\
 &= \frac{1}{2} \text{tr} \{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \}, \tag{2.53}
 \end{aligned}$$

kde symbol  $\text{tr}$  označuje stopu matice. Sečtením výsledků 2.41, 2.49 a 2.53 dostáváme asymptotický vzorec pro vychýlení způsobené odhadováním střední hodnoty logaritmu věrohodnostní funkce pomocí logaritmu věrohodnostní funkce statistického modelu. Dohromady tedy

$$\begin{aligned}
 b(G) &= D_1 + D_2 + D_3 \\
 &= \frac{1}{2} \text{tr} \{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \} + 0 + \frac{1}{2} \text{tr} \{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \} \\
 &= \text{tr} \{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \}. \tag{2.54}
 \end{aligned}$$

Vzhledem k tomu, že tento vztah je opět závislý na skutečné distribuci  $G$ , která generuje naše data, tak musíme opět použít odhad. Pokud bychom měli konzistentní odhady  $\hat{I}$  a  $\hat{J}$ , mohli bychom odhadnout vychýlení  $b(G)$  jako

$$\hat{b} = \text{tr}(\hat{I}\hat{J}^{-1}). \tag{2.55}$$

Přirozeného odhadu těchto dvou matic dosáhneme, nahradíme-li neznámé pravděpodobnostní rozdělení  $G(z)$  pomocí empirického rozdělení  $\hat{G}(z)$  a tedy

$$I(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{\alpha=1}^n \frac{\partial \log f(x_\alpha | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x_\alpha | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \Big|_{\hat{\boldsymbol{\theta}}}, \tag{2.56}$$

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{\alpha=1}^n \frac{\partial^2 \log f(x_\alpha | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\hat{\boldsymbol{\theta}}}. \tag{2.57}$$

## 2.5 AIC

V této sekci odvodíme tvar Akaikeho informačního kritéria (AIC). K tomu budeme chtít vyjádřit tvar vychýlení ve formě, se kterou se výpočetně dá jednoduše manipulovat. Předpokládejme tedy, že skutečná distribuce  $g(x)$  je součástí množiny parametrické rodiny  $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ , tj. existuje takový parametr  $\boldsymbol{\theta}_0 \in \Theta$ , že  $g(x) = f(x|\boldsymbol{\theta}_0)$ .

Čtenáři připomínáme, že podobnou podmínku jsme použili i ve Větě 2.2.1, kdy jsme si poté ověřili, že za této podmínky platí rovnost matic  $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ . Tedy dohromady máme, že

$$E_{G(\mathbf{x}_n)} \left[ \sum_{\alpha=1}^n \log f(X_\alpha | \hat{\boldsymbol{\theta}}) - n E_{G(z)} \log f(Z | \hat{\boldsymbol{\theta}}) \right] = \text{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\} = \text{tr}(I_p) = p, \quad (2.58)$$

kde  $I_p$  je jednotková  $p \times p$  matice. AIC se definuje jako

$$\text{AIC} = -2 \sum_{\alpha=1}^n \log f(X_\alpha | \hat{\boldsymbol{\theta}}) + 2p. \quad (2.59)$$

AIC je tedy odhadem střední hodnoty K-L informace mezi naším modelem a skutečným modelem. Parametr  $p$  lze pochopit jako rozměr parametrického vektoru  $\boldsymbol{\theta}$ , a tedy jako počet proměnných daného statistického modelu.

*Poznámka.* Akaike [2] uvádí, že maximum logaritmu věrohodnostní funkce a konstanta se násobí číslem  $(-2)$  z „historických důvodů“. Lze to například vysvětlit pomocí Wilksovy věty, která říká, že za určitých podmínek a předpokladů má takovýto násobek logaritmu podílu dvou maximálně věrohodných hodnot asymptotické  $\chi^2$  rozdělení, viz [30].

Tato definice je praktická, co se výpočtů týče, zvláště když máme dostatečně velké množství dat. V případě malého počtu dat se používá korekce AIC, kterou značíme  $\text{AIC}_c$  a definujeme jako

$$\text{AIC}_c = \text{AIC} + \frac{2k(k+1)}{n-k-1}, \quad (2.60)$$

kde  $k$  je počet parametrů a  $n$  je počet dat, které máme k dispozici. Je zřejmé, že pro  $n \rightarrow \infty$  bude druhý člen limitně roven 0, a tedy se bude rovnat předchozí definici AIC. Tato korekce je vhodná v případě, kdy podíl  $n/k$  je malý (například  $< 40$ , viz [8]).

**Příklad 2.5.1.** Předvedeme si, jak vypadá explicitní výpočet AIC pro model s normálním rozdělením. Předpokládejme tedy, že

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad (2.61)$$

je hustota normálního rozdělení s parametry  $\boldsymbol{\theta} = (\mu, \sigma^2)$ . Nejprve si odvodíme případ, kdy skutečná hustota  $g(x)$  může pocházet z jakéhokoliv rozdělení. Mějme tedy  $n$  dat  $\{x_1, x_2, \dots, x_n\}$  pocházející ze skutečné hustoty  $g(x)$  a dále mějme náš statistický model daný pomocí

$$f(x|\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left\{ -\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2} \right\} \quad (2.62)$$

s maximálně věrohodnými odhady  $\hat{\mu} = n^{-1} \sum_{\alpha=1}^n x_\alpha$  a  $\hat{\sigma}^2 = n^{-1} \sum_{\alpha=1}^n (x_\alpha - \hat{\mu})^2$ . Vychýlení v tomto případě,

$$E_{P_X} \left[ \frac{1}{n} \sum_{\alpha=1}^n \log f(X_\alpha | \hat{\mu}, \hat{\sigma}^2) - \int_{-\infty}^{\infty} g(z) \log f(z | \hat{\mu}, \hat{\sigma}^2) dz \right], \quad (2.63)$$

lze vypočítat pomocí matic  $I(\boldsymbol{\theta})$  a  $J(\boldsymbol{\theta})$ . Začneme s logaritmem věrohodnostní funkce

$$\log f(x|\boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}.$$

Jeho střední hodnota pak je

$$E_{P_X}[\log f(x|\boldsymbol{\theta})] = -\frac{1}{2} \log(2\pi\sigma^2) - \sigma^2(P_X) + \frac{(\mu - \mu(P_X))^2}{2\sigma^2}, \quad (2.64)$$

kde  $\mu(P_X)$  a  $\sigma^2(P_X)$  je střední hodnota a rozptyl skutečného pravděpodobnostního rozdělení  $P_X$  s hustotou  $g(x)$ . Tedy „pravé“ parametry modelu jsou  $\boldsymbol{\theta}_0 = (\mu(P_X), \sigma^2(P_X))$ . Nyní je třeba vypočítat parciální derivace

$$(i) \quad \frac{\partial}{\partial \mu} \log f(x|\boldsymbol{\theta}) = \frac{x-\mu}{\sigma^2},$$

$$(ii) \quad \frac{\partial}{\partial \sigma^2} \log f(x|\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4},$$

$$(iii) \quad \frac{\partial^2}{\partial \mu^2} \log f(x|\boldsymbol{\theta}) = -\frac{1}{\sigma^2},$$

$$(iv) \quad \frac{\partial^2}{\partial \mu \partial \sigma^2} \log f(x|\boldsymbol{\theta}) = -\frac{x-\mu}{\sigma^4},$$

$$(v) \quad \frac{\partial^2}{(\partial \sigma^2)^2} \log f(x|\boldsymbol{\theta}) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

Pak  $2 \times 2$  matice  $J(\boldsymbol{\theta}_0)$  je ve tvaru

$$J(\boldsymbol{\theta}) = - \begin{bmatrix} E_{P_X} \left[ \frac{\partial^2}{\partial \mu^2} \log f(X|\boldsymbol{\theta}) \right] & E_{P_X} \left[ \frac{\partial^2}{\partial \sigma^2 \partial \mu} \log f(X|\boldsymbol{\theta}) \right] \\ E_{P_X} \left[ \frac{\partial^2}{\partial \mu \partial \sigma^2} \log f(X|\boldsymbol{\theta}) \right] & E_{P_X} \left[ \frac{\partial^2}{(\partial \sigma^2)^2} \log f(X|\boldsymbol{\theta}) \right] \end{bmatrix},$$

po dosazení

$$J(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{E_{P_X}[X-\mu] E_{P_X}[(X-\mu)^2]}{\sigma^4 \sigma^6} \\ \frac{E_{P_X}[X-\mu]}{\sigma^4} & \frac{E_{P_X}[(X-\mu)^2]}{\sigma^6} - \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

Podobně pro matici  $I(\boldsymbol{\theta})$ .

$$I(\boldsymbol{\theta}) = E_{P_X} \begin{bmatrix} \frac{(X-\mu)^2}{\sigma^4} & -\frac{X-\mu}{2\sigma^4} + \frac{(X-\mu)^3}{2\sigma^6} \\ -\frac{X-\mu}{2\sigma^4} + \frac{(X-\mu)^3}{2\sigma^6} & \frac{1}{4\sigma^4} - \frac{(X-\mu)^2}{4\sigma^6} + \frac{(X-\mu)^4}{4\sigma^8} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{\mu_3}{2\sigma^6} \\ \frac{\mu_3}{2\sigma^6} & \frac{\mu_4}{4\sigma^8} - \frac{1}{4\sigma^4} \end{bmatrix},$$

kde  $\mu_j = E_{P_X}[(X-\mu)^j]$  je  $j$ -tý centrální moment. Vidíme, že obecně  $I(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$ . Jednoduchým vynásobením ověříme, že

$$I(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1} = \begin{bmatrix} 1 & \frac{\mu_3}{\sigma^2} \\ \frac{\mu_3}{2\sigma^4} & \frac{\mu_4}{2\sigma^4} - \frac{1}{2} \end{bmatrix},$$

a tedy

$$\text{tr}\{I(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}\} = 1 + \frac{\mu_4}{2\sigma^4} - \frac{1}{2} = \frac{1}{2} \left( 1 + \frac{\mu_4}{\sigma^4} \right).$$

Lze si všimnout toho, že pravá strana není rovna dvojnásobku počtu parametrů (v tomto případě  $2 \times 2$ ). Ovšem budeme-li předpokládat, že existuje hodnota parametru  $\boldsymbol{\theta}_0$  taková, že  $f(x|\boldsymbol{\theta}_0) = g(x)$ , pak  $g(x)$  má normální rozdělení, a tedy  $\mu_3 = 0$  a  $\mu_4 = 3\sigma^4$ . Z toho pak plyne, že

$$\frac{1}{2} + \frac{\mu_4}{2\sigma^4} = \frac{1}{2} + \frac{3\sigma^4}{2\sigma^4} = 2.$$

Ve výsledku dostáváme, že AIC pro model normálního rozdělení je

$$\text{AIC} = -2 \sum_{\alpha=1}^n \log f(x_\alpha|\hat{\mu}, \hat{\sigma}^2) + 2 \times 2, \quad (2.65)$$

kde maximum logaritmu věrohodnostní funkce je

$$\sum_{\alpha=1}^n \log f(x_\alpha|\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{n}{2}. \quad (2.66)$$

Na závěr je třeba upozornit na to, že AIC nelze použít samo o sobě při testování kvality modelu, tj. je třeba mít minimálně dva modely, které můžeme vzájemně srovnávat. To nás vede k hlavní slabině AIC – jedná se pouze o relativní kvalitu modelu vzhledem k ostatním. Pokud budeme pracovat s množinou modelů, které budou všechny nekvalitní analýzou našich dat, tak AIC vybere ten nejméně nekvalitní model z nich.

## Kapitola 3

# Modelování pomocí AIC

### 3.1 Lineární a polynomiální regrese

Předpokládejme, že máme závislou proměnnou  $y$  a  $m$  nezávislých proměnných  $x_1, \dots, x_m$ . Uvažujme lineární regresní model ve tvaru

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.1)$$

kde  $\varepsilon_i$  je náhodná chyba, o které předpokládáme, že  $\varepsilon_i \sim N(0, \sigma^2)$ . Hustota tohoto modelu je

$$f(y|x_1, \dots, x_m) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \left( y - \beta_0 - \sum_{j=1}^m \beta_j x_j \right)^2 \right\}. \quad (3.2)$$

Pokud máme množinu  $n$  nezávislých pozorování  $\{y_i, x_{i1}, \dots, x_{im}\}$  pro  $i = 1, \dots, n$ , její logaritmus věrohodnostní funkce můžeme vyjádřit jako

$$l(\beta_0, \dots, \beta_m, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^m \beta_j x_{ij} \right)^2 \quad (3.3)$$

a maximálně věrohodné odhady  $\hat{\beta}_0, \dots, \hat{\beta}_m$  můžeme získat (ať už metodou ortogonální projekce, nejmenších čtverců atd.) jako řešení následující soustavy rovnic

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}, \quad (3.4)$$

kde  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_m)^T$  a  $n \times (m+1)$  matice  $\mathbf{X}$  a  $n$ -rozměrný vektor  $\mathbf{y}$  jsou definované jako

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \quad (3.5)$$

Maximálně věrohodný odhad  $\hat{\sigma}^2$ , který ovšem není nestranný, je dán

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im})\}^2. \quad (3.6)$$

Dosadíme-li rovnici 3.6 do rovnice 3.3, tak dostáváme

$$l(\hat{\beta}_0, \dots, \hat{\beta}_m, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log d(x_1, \dots, x_m) - \frac{n}{2}, \quad (3.7)$$

kde  $d(x_1, \dots, x_m)$  je odhad pro rozptyl reziduí modelu, který vidíme v (3.6). Jelikož počet parametrů v našem regresním modelu je  $m+2$  ( $m+1$  hodnot  $\beta_i$  a  $\sigma^2$ ), tak výsledný vzorec pro AIC je

$$\text{AIC} = n(\log(2\pi) + 1) + n \log d(x_1, \dots, x_m) + 2(m+2). \quad (3.8)$$

**Příklad 3.1.1 (Swiss package data).** V Tabulce 3.1 vidíme prvních několik řádků, které obsahují data z balíčku **Swiss**, který je volně dostupný v rámci knihovny **datasets** jazyka R [22]. Je to soubor dat popisující socio-ekonomické faktory 47 francouzsky mluvících provincií ve Švýcarsku z roku 1888. **Fertility** je standardizovaná míra plodnosti, **Agriculture** je procento lidí pracujících v zemědělství, **Examination** je procento vojáků s nejvyšší známkou z vojenské zkoušky, **Education** je procento vojáků se vzděláním přesahující základní školu, **Catholic** je procento katolické populace a **Infant Mortality** je procento dětí, které nežilo déle jak 1 rok. Pro další informace [26].

	<b>Fertility</b>	<b>Agriculture</b>	<b>Examination</b>	<b>Education</b>	<b>Catholic</b>	<b>Infant.Mortality</b>
Courtellary	80,2	17,0	15	12	9,96	22,2
Delemont	83,1	45,1	6	9	84,84	22,2
Franches-Mnt	92,5	39,7	5	5	93,40	20,2
Moutier	85,8	36,5	12	7	33,77	20,3
Neuveville	76,9	43,5	17	15	5,16	20,6
Porrentruy	76,1	35,3	9	7	90,57	26,6

Tabulka 3.1: Swiss data

Naším cílem bude nalezení nejvhodnějšího modelu pomocí AIC kritéria, kdy budeme proměnnou **Fertility** budeme považovat za závislou proměnnou a zbývající za naše nezávislé. Nejprve porovnáme dva jednoduché modely a následně předvedeme, jak můžeme vyřešit tento problém obecně. První model bude mít následující podobu

$$\text{Fertility}_i = \beta_0 + \beta_1 \times \text{Agriculture} + \varepsilon_i. \quad (3.9)$$

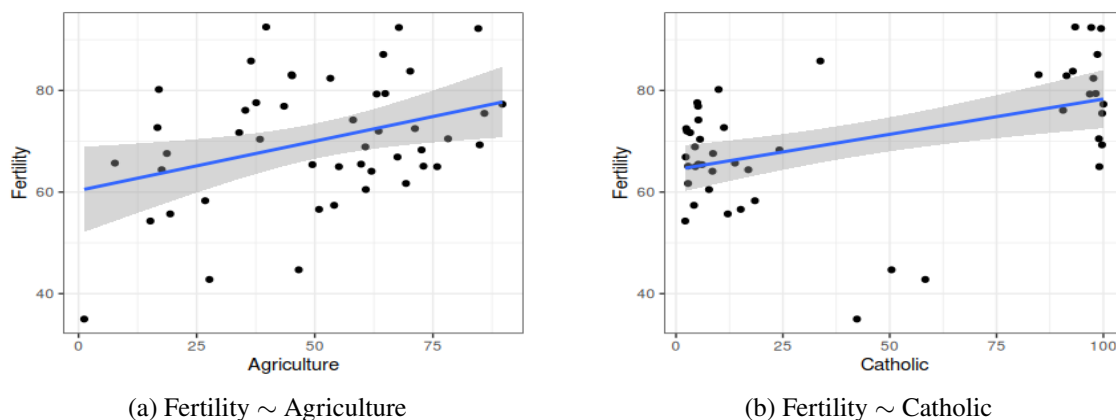
Zatímco model druhý bude vypadat takto,

$$\text{Fertility}_i = \beta_0 + \beta_1 \times \text{Catholic} + \varepsilon_i. \quad (3.10)$$

Jazyk R má ve svém balíčku **stats** implementovanou funkci **AIC()**, jejíž základním nastavením je právě Akaikeho informační kritérium (lze například zvolit Bayesovo informační kritérium místo toho). Pak stačí vytvořit naše modely a vyhodnotit je. Dostáváme hodnoty 369,4675 pro model s proměnnou **Agriculture** a hodnotu 364,3479 pro model s **Catholic**. AIC nám tedy říká, že procentuální zastoupení katolické populace je vhodnější parametr na modelování výsledné plodnosti než procentuální zastoupení populace pracujících v zemědělství. Když se ovšem podíváme na příslušné grafy (viz na Obrázku 3.1), tak vidíme, že ani jeden z modelů není příliš kvalitním popisem situace.

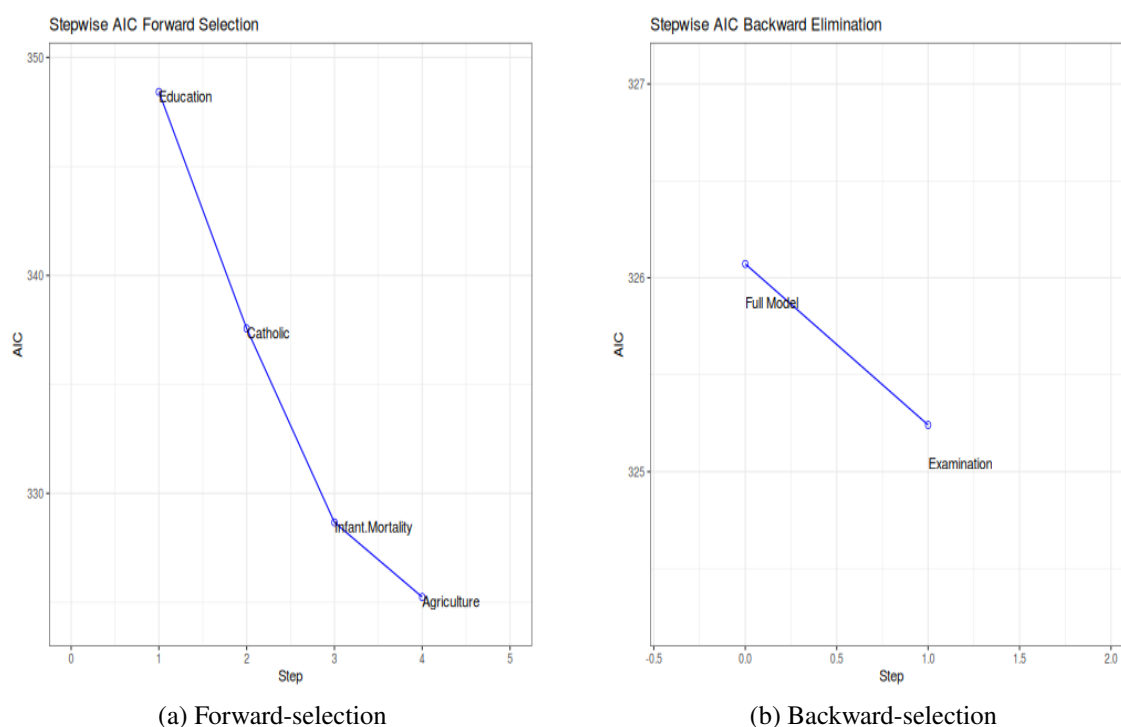
Je patrné, že bude třeba rozšířit model o další proměnné. Tento proces lze automatizovat a máme tu dva přístupy k tomu, konkrétně *forward-selection* a *backward-selection*.





Obrázek 3.1: Srovnání regresních modelů 3.9 a 3.10 pro data Swiss z R

U forward-selection se jedná o proces, kdy začínáme s prázdným modelem a postupně přidáváme proměnné do té doby, než narazíme na minimum AIC. Backward-selection postupuje opačným směrem, tedy z plného modelu postupně odebíráme proměnné za účelem minimalizace AIC modelu. U obou variant lze použít funkce z knihovny **olsrr**. Pro bližší seznámení doporučujeme [14].



Obrázek 3.2: Forward a Backward selection pro regresní modely 3.9 a 3.10 pro data Swiss z R

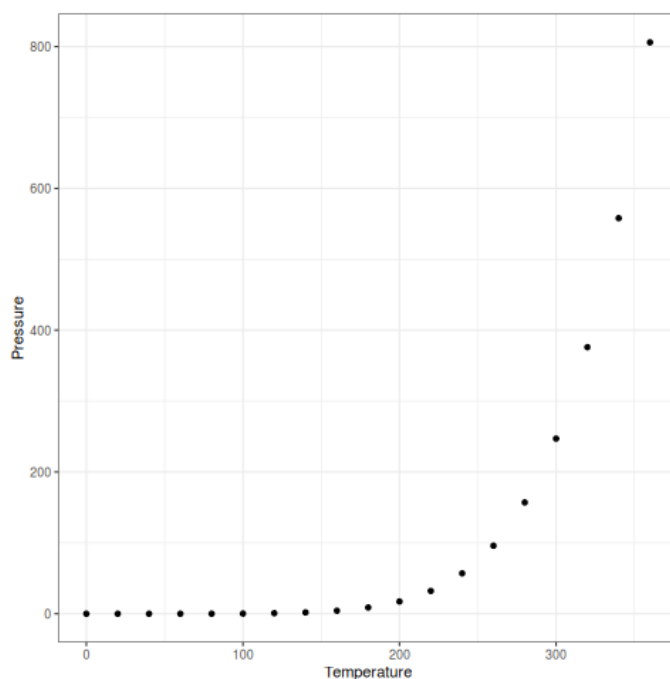
Vidíme, že v obou případech (na obrázku 3.2) jsme dostali ten samý výsledný model

ve tvaru

$$\text{Fertility}_i = \beta_0 + \beta_1 \times \text{Catholic}_i + \beta_2 \times \text{Agriculture}_i + \beta_3 \times \text{Education}_i + \beta_4 \times \text{Infant.Mortality}_i + \varepsilon_i \quad (3.11)$$

s hodnotou AIC rovnou 325,2408. Pochopitelně zkoušet všech  $2^p$  kombinací (bez toho, aniž bychom řešili jejich mocniny a interakce) proměnných je časově a výpočetně náročné, máme-li více proměnných. Také samotný proces automatizace výběru proměnných je silně kritizován, a proto se nedoporučuje používat jako hlavní nástroj pro vybudování modelu.

**Příklad 3.1.2 (Pressure data).** Následující příklad ukazuje, že stejným přístupem lze vybírat modely u polynomiální regrese. Na obrázku 3.3 vidíme, jak vypadají data z balíčku **Pressure**, který obsahuje měření závislosti tlaku na teplotě u výparů z rtuti (více informací na [28]). Na první pohled je jasné, že ta závislost není lineární, a proto to prokládat přímkou nepřipadá v úvahu. Nabízí se vyzkoušet polynomiální regresi. Provedeme lehkou změnu značení a Temperature nahradíme za  $x$  pro jednodušší zápis.



Obrázek 3.3: Grafické znázornění  $\text{Pressure} \sim \text{Temperature}$

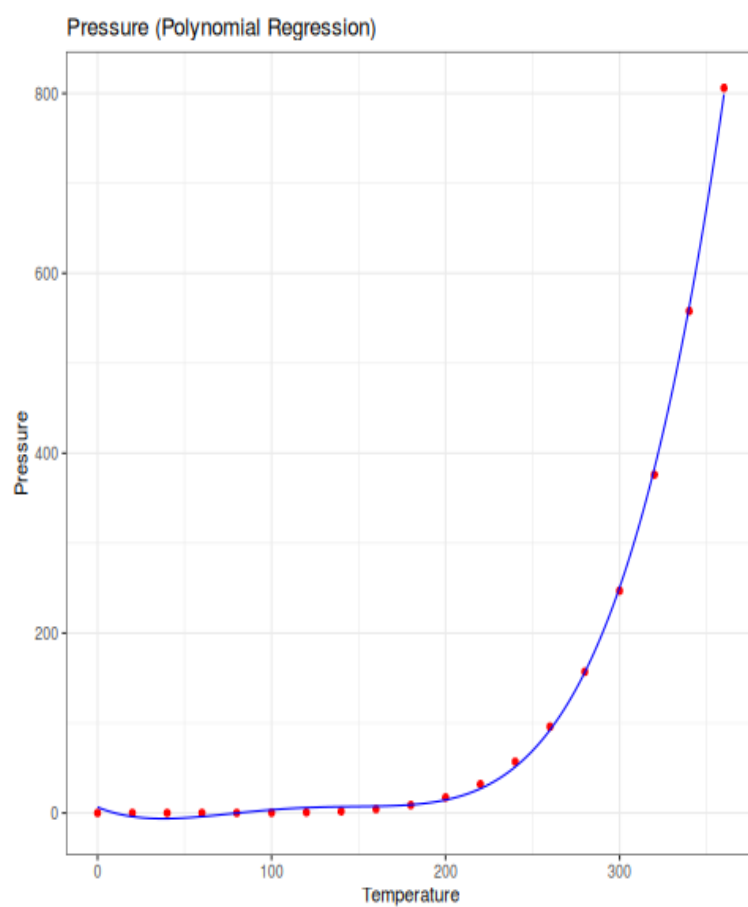
Tabulka 3.2 nám ukazuje postupně všechny kombinace mocnin až do 4. řádu a hodnotu AIC příslušného modelu k tomu. Vidíme tedy, že minimální hodnota AIC je daná pro model

$$\text{pressure}_i = \beta_0 + \beta_1 \times \text{temperature}_i + \beta_2 \times \text{temperature}_i^2 + \beta_3 \times \text{temperature}_i^3 + \beta_4 \times \text{temperature}_i^4 + \varepsilon_i. \quad (3.12)$$

Můžeme si také graficky zobrazit náš model a posoudit jeho kvalitu, viz Obrázek 3.4.

Model	AIC hodnota
$x$	248,4163
$x^2$	236,1611
$x^3$	222,6131
$x^4$	206,2072
$x + x^2$	222,4226
$x + x^3$	208,6165
$x + x^4$	192,0272
$x^2 + x^3$	196,0159
$x^2 + x^4$	179,3811
$x^3 + x^4$	165,4423
$x + x^2 + x^3$	182,6668
$x + x^2 + x^4$	165,6688
$x + x^3 + x^4$	151,2014
$x^2 + x^3 + x^4$	137,6644
$x + x^2 + x^3 + x^4$	124,0586

Tabulka 3.2: Hodnoty AIC regresních podmodelů pro Pressure data v R



Obrázek 3.4: Grafické znázornění regresního modelu 3.12

## 3.2 Histogramy

Histogram obecně je grafickým znázorněním odhadu hustoty pravděpodobnosti, která generuje data se kterými pracujeme. Histogram se vytváří tak, že rozdělíme naše data do tříd (anglicky *bin* nebo *bucket*) pokrývajících rozsah hodnot dat a následně spočítáme kolik z nich se nachází v jaké třídě. Tedy máme-li množinu dat  $M = \{x_1, x_2, \dots, x_n\}$ , je přirozené použít interval  $(a, b)$ , kde  $a$  je minimální hodnotou množiny  $M$  a  $b$  hodnotou maximální. Předpokládejme, že budeme chtít, aby šířky našich tříd byly všechny stejné. Konstrukci histogramu můžeme specifikovat buď množstvím tříd nebo šířkou individuální třídy, neboť mezi nimi existuje vztah

$$h = \frac{b - a}{k}, \quad (3.13)$$

kde  $h$  je šířkou třídy a  $k$  je množství těchto tříd. Klasická funkce **hist**( ) pro ustanovení počtu tříd používá ve svém základním nastavení Sturgesovo pravidlo, které říká, že odhad počtu tříd  $k$  je ve tvaru

$$\hat{k} = 1 + \log_2(n). \quad (3.14)$$

Toto pravidlo je přesné pro data pocházející z normálního rozdělení a je velice jednoduché na implementaci ve výpočetním softwaru. Odvození tohoto pravidla je také poměrně jednoduché. Vzhledem ke konstrukci histogramu se nabízí použít binomické rozdělení k aproximaci normálního rozdělení za použití centrální limitní věty. Mějme tedy náhodnou veličinu  $Y \sim Bi(m, p)$ . Abychom mohli aproximovat normální rozdělení, musí platit, že  $mp > 5$  a nejlepší aproximace dosáhneme pro  $p = 1/2$ . Tedy

$$f(y) = P(Y = y) = \binom{m}{y} p^y (1 - p)^{m-y}, \quad y = 0, 1, \dots, m, \quad (3.15)$$

pak pro  $p = 1/2$  vypadá následovně,

$$f(y) = \binom{m}{y} 2^{-m}. \quad (3.16)$$

Hlavní myšlenkou Sturgesova pravidla je, že u „ideálního“ souboru dat, která jsou z normálního rozdělení by počet bodů v jednotlivých třídách, značeno tady jako  $n_i$ , právě odpovídal binomickým koeficientům  $\binom{m}{i}$  pro  $i = 0, \dots, m$ . Máme tedy histogram s  $m + 1$  třídami. Sečtením absolutních četností bodů v jednotlivých třídách z množiny dat dostaneme

$$n = \sum_{i=0}^m n_i = \sum_{i=0}^m \binom{m}{i} = (1 + 1)^m = 2^m, \quad (3.17)$$

kde jsme v třetí rovnosti použili binomickou větu. Řešením této rovnice je  $m = \log_2(n)$ . Zároveň platí, že počet tříd  $k$  je roven  $m + 1$  a máme

$$k = 1 + m = 1 + \log_2(n). \quad (3.18)$$

Odvození spolu s kritikou Sturgesova pravidla lze najít například v [23]. Naším cílem je použít AIC k výběru počtu tříd v histogramu, respektive k porovnání hodnot AIC pro histogramy s různými počty tříd a následnému výběru histogramu s nejnižší hodnotou AIC.

Histogram s  $k$  třídami můžeme totiž považovat za model s multinomickým rozdělením s  $k$  parametry ve tvaru

$$P(n_1 + \dots + n_k | p_1, \dots, p_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}, \quad (3.19)$$

kde  $n_1 + \dots + n_k = n$  a  $p_1 + \dots + p_k = 1$ ,  $n$  je celkový počet dat a  $n_i$  je počet bodů v  $i$ -té třídě. Logaritmus věrohodnostní funkce je pak rovný

$$l(p_1, \dots, p_k) = C + \sum_{j=1}^k n_j \log(p_j), \quad (3.20)$$

kde  $C = \log(n!) - \sum_{j=1}^k \log(n_j!)$  je člen nezávislý na hodnotách  $p_j$ . Maximálně věrohodným odhadem parametru  $p_j$  je tedy

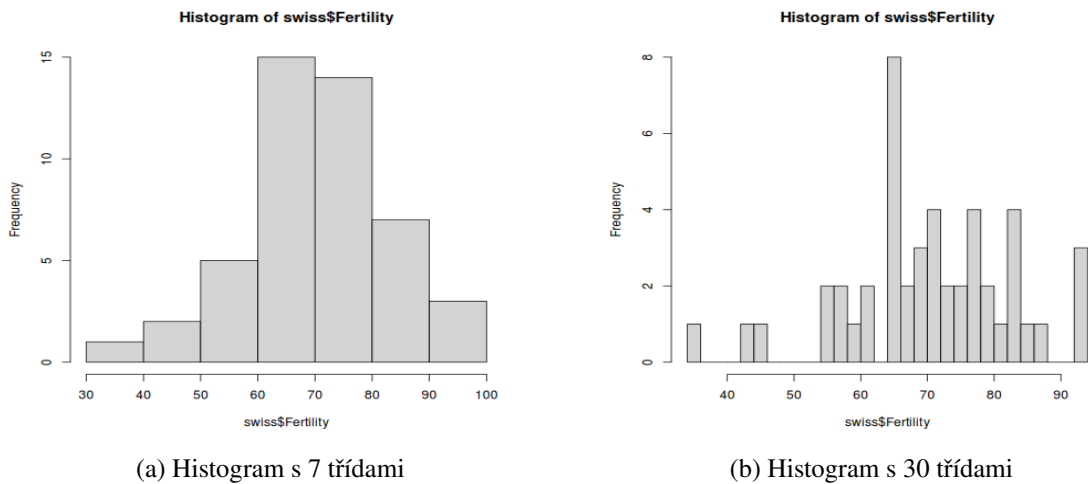
$$\hat{p}_j = \frac{n_j}{n}. \quad (3.21)$$

Dohromady je AIC histogramu ve tvaru

$$AIC = (-2) \left\{ C + \sum_{j=1}^k n_j \log\left(\frac{n_j}{n}\right) \right\} + 2(k-1). \quad (3.22)$$

Může se stát, že v jedné či vícero třídách se nenachází žádný bod, a tedy  $n_j = 0$ . V takovém případě v rámci korektnosti položíme  $0 \log\left(\frac{0}{n}\right) = 0$ , kde můžeme stejně jako u definice entropie argumentovat pomocí limity  $\lim_{x \rightarrow 0} x \log(x) = 0$ .

**Příklad 3.2.1.** Podíváme-li se zpět na balíček **Swiss**, tak nás v případě modelování proměnné **Fertility** může zajímat její rozdělení, které lze aproximovat právě histogramem.



Obrázek 3.5: Srovnání 7 a 30 tříd v histogramu pro proměnnou **Fertility** z dat Swiss v R

Na obrázku 3.5 vidíme rozdíl mezi volbou 7 a 30 tříd pro histogram proměnné **Fertility**. Výsledky výpočtu AIC pro histogram s 7, 12 a 30 třídami lze vidět v Tabulce 3.3.

Vidíme tedy, že histogram se 7 třídami je nejlepší volbou v tomto případě. Je vidět, že s 30 třídami je histogram příliš citlivý a nezachycuje dobře hustotu rozdělení našich dat.

Počet tříd	AIC
7	18,24804
12	28,60021
30	46,32346

Tabulka 3.3: Tabulka hodnot AIC histogramů pro různé třídy

Za zmínku stojí zmínit funkci **histogram()** z knihovny **histogram**, která automaticky vybírá množství tříd na základě penalizací v nabídce funkce (více informací na [20]). Mezi ty patří právě i AIC v následujícím tvaru. Nechť  $k$  označuje opět počet tříd,  $n$  je celkový počet bodů,  $n_i$  je počet bodů v dané třídě pro  $i = 1, \dots, k$  a  $w_i$  je šířka třídy  $i$ ,  $i = 1, \dots, k$ . Pak logaritmus věrohodnostní funkce s penalizací na základě AIC je ve tvaru

$$\sum_{i=1}^k n_i \log \left( \frac{n_i}{nw_i} \right) - \alpha k, \quad (3.23)$$

kde  $\alpha$  je parametr, který lze měnit dle potřeby se základní hodnotou  $\alpha = 1$ . Funkce **histogram()** následně hledá takovou hodnotu  $k$ , která maximalizuje tento výraz.

### 3.3 Rovnost dvou diskretních distribucí

AIC lze také použít k určení zdali dvě množiny dat pocházejí ze stejné diskretní distribuce, či naopak z různých. Předpokládejme tedy, že máme dvě různé množiny dat v následující podobě,

Kategorie	1	2	...	$k$
Množina dat 1	$n_1$	$n_2$	...	$n_k$
Množina dat 2	$m_1$	$m_2$	...	$m_k$

Počty pozorování se zároveň rovnají  $n_1 + n_2 + \dots + n_k = n$  a  $m_1 + m_2 + \dots + m_k = m$  respektive. Dále předpokládáme, že data z obou množin se řídí multinomickým rozdělením s  $k$  kategoriemi, tedy

$$P(n_1, n_2, \dots, n_k | p_1, p_2, \dots, p_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}, \quad (3.24)$$

$$P(m_1, m_2, \dots, m_k | q_1, q_2, \dots, q_k) = \frac{m!}{m_1! m_2! \dots m_k!} q_1^{m_1} q_2^{m_2} \dots q_k^{m_k}, \quad (3.25)$$

kde  $p_i$  a  $q_i$  označují pravděpodobnost jevu, že každý jev v množině dat 1 a 2 patří do kategorie  $i$  a zároveň  $\mathbf{p} = (p_1, p_2, \dots, p_k)$  a  $\mathbf{q} = (q_1, q_2, \dots, q_k)$  splňují  $p_i > 0$  a  $q_i > 0$ . Zaprvé uvažujme model, kdy množina dat 1 a 2 pocházejí z různých distribucí. Pak

logaritmus věrohodnostní funkce v tomto případě je ve tvaru

$$l_1(p_1, \dots, p_k, q_1, \dots, q_k) = \log(n!) - \sum_{j=1}^k \log(n_j!) + \sum_{j=1}^k n_j \log(p_j) \\ + \log(m!) - \sum_{j=1}^k \log(m_j!) + \sum_{j=1}^k m_j \log(q_j). \quad (3.26)$$

Maximálně věrohodné odhady parametrů  $p_j$  a  $q_j$  jsou pak dány výrazy

$$\hat{p}_j = \frac{n_j}{n}, \quad \hat{q}_j = \frac{m_j}{m} \quad (3.27)$$

a maximum logaritmu věrohodnostní funkce je rovné

$$l_1(\hat{p}_1, \dots, \hat{p}_k, \hat{q}_1, \dots, \hat{q}_k) = C + \sum_{j=1}^k n_j \log\left(\frac{n_j}{n}\right) + \sum_{j=1}^k m_j \log\left(\frac{m_j}{m}\right), \quad (3.28)$$

kde  $C = \log(n!) + \log(m!) + \sum_{j=1}^k (\log(n_j!) + \log(m_j!))$  je konstanta nezávislá na parametrech. Počet parametrů je zde  $2(k-1)$  a tedy AIC je ve tvaru

$$\text{AIC} = -2l_1(\hat{p}_1, \dots, \hat{p}_k, \hat{q}_1, \dots, \hat{q}_k) + 2 \times 2(k-1) \\ = -2 \left\{ C + \sum_{j=1}^k n_j \log\left(\frac{n_j}{n}\right) + \sum_{j=1}^k m_j \log\left(\frac{m_j}{m}\right) \right\} + 4(k-1). \quad (3.29)$$

Naopak předpokládáme-li, že množina dat 1 a 2 pocházejí z té samé distribuce, tak platí rovnost  $p_j = q_j \equiv r_j$ . Pak logaritmus věrohodnostní funkce je ve tvaru

$$l_2(r_1, \dots, r_k) = C + \sum_{j=1}^k (n_j + m_j) \log(r_j). \quad (3.30)$$

Maximálně věrohodný odhad  $r_j$  je v tomto případě

$$\hat{r}_j = \frac{n_j + m_j}{n + m} \quad (3.31)$$

a maximum logaritmu věrohodnostní funkce je ve tvaru

$$l_2(\hat{r}_1, \dots, \hat{r}_k) = C + \sum_{j=1}^k (n_j + m_j) \log\left(\frac{n_j + m_j}{n + m}\right). \quad (3.32)$$

Tentokrát máme  $k-1$  parametrů a tedy AIC je nakonec

$$\text{AIC} = -2l_2(\hat{r}_1, \dots, \hat{r}_k) + 2(k-1) \\ = -2 \left\{ C + \sum_{j=1}^k (n_j + m_j) \log\left(\frac{n_j + m_j}{n + m}\right) \right\} + 2(k-1). \quad (3.33)$$

Systém/Distribuce	Windows	MacOS	Ubuntu	Gentoo	FreeBSD
Průzkum 1	1600	250	100	40	10
Průzkum 2	1450	300	125	100	25

Tabulka 3.4: Výsledky průzkumu operačních systémů pro příklad 3.3.1

**Příklad 3.3.1.** Mějme průzkum týkající se využívání různých operačních systémů či jejich distribucí. Ptali jsme se 2000 lidí, který z následujících systémů používají denně na osobním počítači/notebooku – Windows, MacOS, Ubuntu, Gentoo nebo FreeBSD. Tento pokus opakujeme dohromady dvakrát s rozstupem 2 let mezi sebou. Výsledky průzkumu jsou shrnuty v Tabulce 3.4.

V Tabulce 3.5 jsme spočítali odhady pro  $p_j$  i  $q_j$ . Zároveň jsou zde spočítané odhady pro  $r_j$ , tj. pro model, kdy jsou distribuce obou množin stejné.

Systém/Distribuce	Windows	MacOS	Ubuntu	Gentoo	FreeBSD
$\hat{p}_j$	0,8	0,125	0,05	0,02	0,005
$\hat{q}_j$	0,725	0,15	0,0625	0,05	0,0125
$\hat{r}_j$	0,7625	0,1375	0,05625	0,035	0,00875

Tabulka 3.5: Tabulka odhadů  $p_j, q_j$  a  $r_j$  v příkladu 3.3.1

Výpočet AIC je pak už jen dosazením do vztahů 3.29 a 3.33 s tím, že  $2(k-1) = 8$  a  $k-1 = 4$ . Tedy

$$\text{Model 1 : } AIC = 6290,882,$$

$$\text{Model 2 : } AIC = 6410,0366,$$

V obou případech jsme ignorovali konstantu  $C$ , která ovšem nemění závěr tohoto výpočtu. AIC naznačuje, že ty dvě množiny dat pocházejí z různých distribucí.

### 3.4 Rovnost středních hodnot a rozptylů

AIC lze také použít, máme-li dvě různé množiny dat pocházející ze stejného rozdělení a zajímá nás, jestli parametr rozdělení je identický u obou množin, či jestli je rozdílný.

**Příklad 3.4.1** (Poissonovo rozdělení). Mějme dvě různé množiny dat z Poissonova rozdělení o velikostech  $n$  a  $m$  respektive, tj.  $\{x_1, \dots, x_n\}$  a  $\{x_{n+1}, \dots, x_{n+m}\}$ . Dále předpokládejme, že  $x_1, \dots, x_n \sim Po(\lambda_1)$  a  $x_{n+1}, \dots, x_{n+m} \sim Po(\lambda_2)$ . Z toho plyne, že

$$p(x_i|\lambda_1) = \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}, \quad i = 1, \dots, n,$$

$$p(x_i|\lambda_2) = \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}, \quad i = n+1, \dots, n+m. \quad (3.34)$$



Můžeme nyní analyzovat dvě situace. Budeme předpokládat prvně, že  $\lambda_1 = \lambda_2 = \lambda$ . Věrohodnostní funkce bude ve tvaru

$$L(\lambda; x_1, \dots, x_{n+m}) = \prod_{i=1}^{n+m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \quad (3.35)$$

a tedy logaritmus věrohodnostní funkce vypadá jako

$$l(\lambda; x_1, \dots, x_{n+m}) = -(n+m)\lambda + \ln(\lambda) \sum_{i=1}^{n+m} x_i - \sum_{i=1}^{n+m} \ln(x_i!). \quad (3.36)$$

Odhad parametru  $\lambda$  získáme ve tvaru

$$\hat{\lambda} = \frac{1}{n+m} \sum_{i=1}^{n+m} x_i, \quad (3.37)$$

tedy průměr naměřených či získaných hodnot  $x_i$ . Pak maximální hodnota logaritmu věrohodnostní funkce je

$$l(\hat{\lambda}; x_1, \dots, x_{n+m}) = - \sum_{i=1}^{n+m} x_i + \ln\left(\frac{1}{n+m} \sum_{i=1}^{n+m} x_i\right) \sum_{i=1}^{n+m} x_i - \sum_{i=1}^{n+m} \ln(x_i!). \quad (3.38)$$

Tedy AIC pro model, kdy  $\lambda_1 = \lambda_2 = \lambda$  je rovno

$$\text{AIC} = -2l(\hat{\lambda}; x_1, \dots, x_{n+m}) + 2, \quad (3.39)$$

kde počet parametrů  $p = 1$ . Uvažujeme-li nyní případ, kdy  $\lambda_1 \neq \lambda_2$ , tak věrohodnostní funkce tohoto modelu je tentokrát

$$L(\lambda_1, \lambda_2; x_1, \dots, x_{n+m}) = \prod_{i=1}^n \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} \prod_{i=n+1}^{n+m} \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}. \quad (3.40)$$

Logaritmus věrohodnostní funkce je v tomto případě

$$\begin{aligned} l(\lambda_1, \lambda_2; x_1, \dots, x_{n+m}) = & -n\lambda_1 + \ln(\lambda_1) \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!) \\ & - m\lambda_2 + \ln(\lambda_2) \sum_{i=n+1}^{n+m} x_i - \sum_{i=n+1}^{n+m} \ln(x_i!). \end{aligned} \quad (3.41)$$

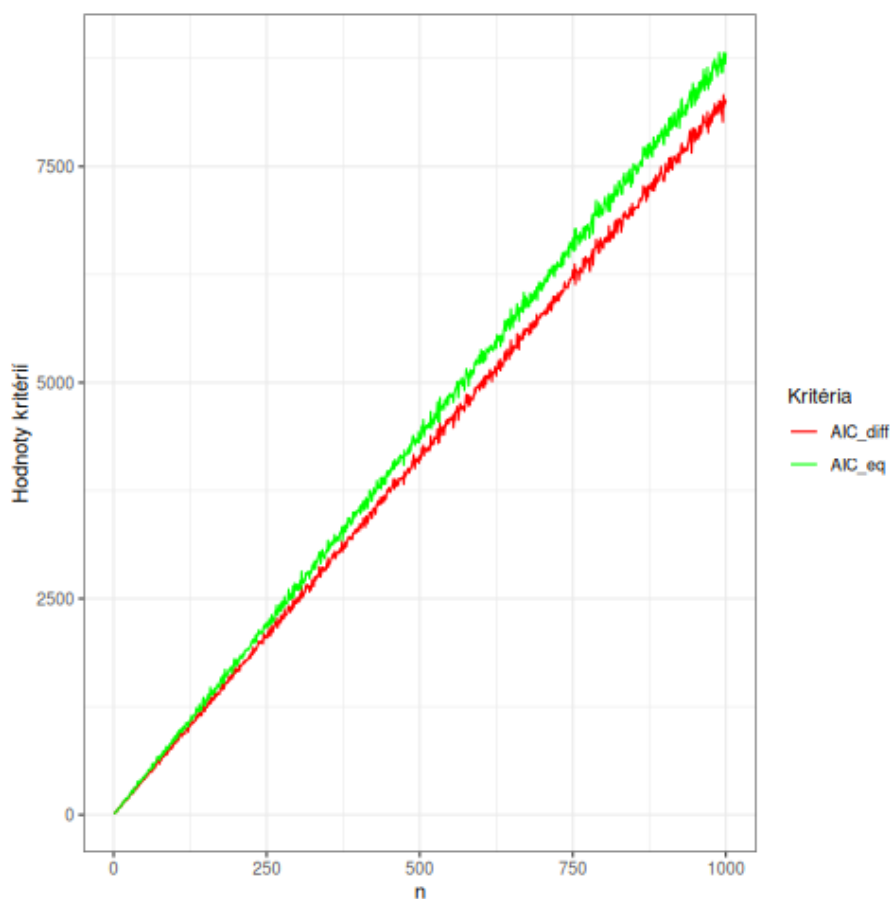
Odhady pro parametry  $\lambda_1$  a  $\lambda_2$  získáme jako řešení soustavy rovnic  $\frac{\partial l}{\partial \lambda_1} = \frac{\partial l}{\partial \lambda_2} = 0$ . Maximální hodnota logaritmu věrohodnostní funkce pak vypadá následovně

$$\begin{aligned} l(\hat{\lambda}_1, \hat{\lambda}_2; x_1, \dots, x_{n+m}) = & - \sum_{i=1}^n x_i + \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!) \\ & - \sum_{i=n+1}^{n+m} x_i + \ln\left(\frac{1}{m} \sum_{i=n+1}^{n+m} x_i\right) \sum_{i=n+1}^{n+m} x_i - \sum_{i=n+1}^{n+m} \ln(x_i!), \end{aligned} \quad (3.42)$$

kdy výsledné AIC s počtem parametrů  $p = 2$  je

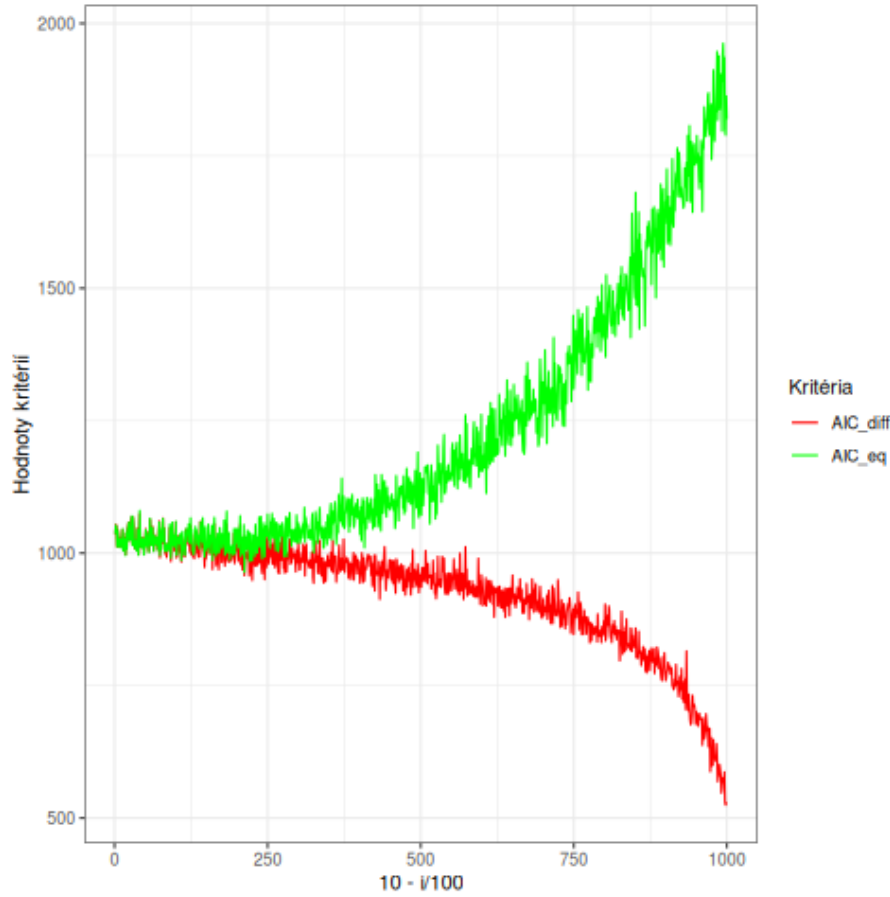
$$\text{AIC} = -2l(\hat{\lambda}_1, \hat{\lambda}_2; x_1, \dots, x_{n+m}) + 2 \times 2. \quad (3.43)$$

Ověříme numerickou přesnost a stabilitu teoretického vzorce. Na Obrázku 3.6 demonstrujeme vývoj hodnot AIC kritéria pro oba modely  $\lambda_1 = \lambda_2$  (označeno jako AIC\_eq) a  $\lambda_1 \neq \lambda_2$  (označeno jako AIC\_diff), kde nás zajímá jejich závislost na počtu pozorování. Vzali jsme tedy náhodně vygenerovaná pozorování z Poissonova rozdělení o  $\lambda_1 = 3$  a  $\lambda_2 = 5$  a postupně jsme vypočítali hodnoty AIC pro počet pozorování  $n = 1, \dots, 1000$ . Je důležité si všimnout toho, že AIC neroste striktně lineárně, ale dochází tam k viditelným oscilacím. Pro  $n = 1, \dots, 5$  nám vychází, že  $\lambda_1$  by se mělo rovnat  $\lambda_2$ . Na Obrázku 3.7 jsme naopak zjišťovali závislost na vzdálenosti  $\lambda_1$  od  $\lambda_2$ . Uvažovali jsme naopak fixní  $n = 100$ ,  $\lambda_1 = 10$  a  $\lambda_i = 10 - \frac{i}{100}$  pro  $i = 1, \dots, 1000$ . Tady už vidíme podstatně silnější oscilace ve vývoji obou hodnot AIC. Také vychází, že pro  $i = 1, \dots, 100$  AIC preferuje model  $\lambda_1 = \lambda_2$ . Shrňme-li oba naše výsledky, tak dojdeme k závěru, že AIC má tendenci vybrat špatný model pro malý počet pozorování a velice blízké hodnoty parametrů.



Obrázek 3.6: Grafický vývoj hodnot AIC pro  $\lambda_1 = \lambda_2$  a  $\lambda_1 \neq \lambda_2$  pro  $n$  od 1 do 1000 v Příkladu 3.4.1

**Příklad 3.4.2** (Normální rozdělení). Mějme pro změnu dvě množiny dat  $\{y_1, \dots, y_n\}$  a  $\{y_{n+1}, \dots, y_{n+m}\}$  pocházející z normálního rozdělení, kdy platí, že  $y_1, \dots, y_n \sim N(\mu_1, \sigma_1^2)$



Obrázek 3.7: Grafické vývoje AIC pro  $\lambda_1 = 10$  a  $\lambda_2 = 10 - \frac{i}{100}$  pro  $i$  od 1 do 1000 v Příkladu 3.4.1

a  $y_{n+1}, \dots, y_{n+m} \sim N(\mu_2, \sigma_2^2)$ . Hustota rozdělení jednotlivých množin je

$$f(y_i | \mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left\{ -\frac{(y_i - \mu_1)^2}{2\sigma_1^2} \right\}, \quad i = 1, \dots, n,$$

$$f(y_i | \mu_2, \sigma_2^2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp \left\{ -\frac{(y_i - \mu_2)^2}{2\sigma_2^2} \right\}, \quad i = n+1, \dots, n+m.$$

Nejprve budeme uvažovat situaci, kdy  $\mu_1 \neq \mu_2$  a  $\sigma_1^2 \neq \sigma_2^2$ . V tom případě je logaritmus věrohodnostní funkce dán jako

$$l(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = -\frac{n}{2} \ln(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{j=1}^n (y_j - \mu_1)^2 - \frac{m}{2} \ln(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2} \sum_{j=n+1}^{n+m} (y_j - \mu_2)^2. \quad (3.44)$$

Odtud lze spočítat maximálně věrohodné odhady parametrů,

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n} \sum_{j=1}^n y_j, & \hat{\sigma}_1^2 &= \frac{1}{n} \sum_{j=1}^n (y_j - \hat{\mu}_1)^2, \\ \hat{\mu}_2 &= \frac{1}{m} \sum_{j=n+1}^{n+m} y_j, & \hat{\sigma}_2^2 &= \frac{1}{m} \sum_{j=n+1}^{n+m} (y_j - \hat{\mu}_2)^2.\end{aligned}$$

Maximum logaritmu věrohodnostní funkce je

$$l(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2) = -\frac{n}{2} \ln(2\pi\hat{\sigma}_1^2) - \frac{m}{2} \ln(2\pi\hat{\sigma}_2^2) - \frac{n+m}{2} \quad (3.45)$$

a jelikož počet parametrů tohoto modelu  $p = 4$ , výsledné AIC je rovné

$$\text{AIC} = (n+m)(\ln(2\pi) + 1) + n \ln(\hat{\sigma}_1^2) + m \ln(\hat{\sigma}_2^2) + 2 \times 4. \quad (3.46)$$

Tento model můžeme srovnat s třemi následujícími. První model předpokládá, že  $\mu_1 = \mu_2 = \mu$  a  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , tj. máme  $n+m$  pozorování  $y_1, \dots, y_{n+m}$  pocházející ze stejného normálního rozdělení. AIC takového modelu se zjednoduší z 3.46 na

$$\text{AIC} = (n+m)\{\ln(2\pi\hat{\sigma}^2) + 1\} + 2 \times 2, \quad (3.47)$$

kdy odhady parametrů  $\hat{\mu}$  a  $\hat{\sigma}^2$  jsou

$$\hat{\mu} = \frac{1}{n+m} \sum_{j=1}^{n+m} y_j, \quad \hat{\sigma}^2 = \frac{1}{n+m} \sum_{j=1}^{n+m} (y_j - \hat{\mu})^2.$$

Druhý model předpokládá pouze rovnost rozptylů, tj.  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . V takovém to případě vypadá logaritmus věrohodnostní funkce jako

$$l_2(\mu_1, \mu_2, \sigma^2) = -\frac{n+m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - \mu_1)^2 - \frac{1}{2\sigma^2} \sum_{j=n+1}^{n+m} (y_j - \mu_2)^2. \quad (3.48)$$

Maximálně věrohodné odhady parametrů dostaneme ve tvaru

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n} \sum_{j=1}^n y_j, & \hat{\mu}_2 &= \frac{1}{m} \sum_{j=n+1}^{n+m} y_j, \\ \hat{\sigma}^2 &= \frac{1}{n+m} \left\{ \sum_{j=1}^n (y_j - \hat{\mu}_1)^2 + \sum_{j=n+1}^{n+m} (y_j - \hat{\mu}_2)^2 \right\}.\end{aligned} \quad (3.49)$$

Maximum této funkce je

$$l_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) = -\frac{n+m}{2} \ln(2\pi\hat{\sigma}^2) - \frac{n+m}{2} \quad (3.50)$$

a vzhledem k tomu, že máme tři parametry, tak AIC v tomto případě je

$$\text{AIC} = (n+m)\{\ln(2\pi\hat{\sigma}^2) + 1\} + 2 \times 3. \quad (3.51)$$

Stejně tak můžeme uvažovat třetí model, kdy se naopak předpokládá, že  $\mu_1 = \mu_2 = \mu$ . Logaritmus věrohodnostní funkce je pak ve tvaru

$$l_3(\mu, \sigma_1^2, \sigma_2^2) = -\frac{n}{2} \ln(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{j=1}^n (y_j - \mu)^2 - \frac{m}{2} \ln(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2} \sum_{j=n+1}^{n+m} (y_j - \mu)^2. \quad (3.52)$$

Bohužel, získat explicitní tvar odhadu pro  $\mu$  není jednoduché. Jako vždy začínáme se soustavou rovnic

$$\frac{\partial l_3}{\partial \mu} = 0, \quad \frac{\partial l_3}{\partial \sigma_1^2} = 0, \quad \frac{\partial l_3}{\partial \sigma_2^2} = 0. \quad (3.53)$$

Začneme s maximálně věrohodnými odhady pro  $\sigma_1^2$  a  $\sigma_2^2$ , tj.

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_{j=1}^n (y_j - \mu)^2, \quad \hat{\sigma}_2^2 = \frac{1}{m} \sum_{j=n+1}^{n+m} (y_j - \mu)^2. \quad (3.54)$$

Dosadíme do rovnice pro odhad  $\mu$  3.53 a získáme

$$\frac{\partial l_3}{\partial \mu} = \frac{1}{\hat{\sigma}_1^2} \sum_{j=1}^n (y_j - \mu)^2 + \frac{1}{\hat{\sigma}_2^2} \sum_{j=n+1}^{n+m} (y_j - \mu)^2 = 0. \quad (3.55)$$

To lze upravit jako

$$n \sum_{j=1}^n (y_j - \mu) \sum_{j=n+1}^{n+m} (y_j - \mu)^2 + m \sum_{j=n+1}^{n+m} (y_j - \mu) \sum_{j=1}^n (y_j - \mu)^2 = 0. \quad (3.56)$$

Předchozí výraz můžeme přepsat jako kubickou rovnici

$$\mu^3 + A\mu^2 + B\mu + C = 0. \quad (3.57)$$

Koeficienty  $A$ ,  $B$  a  $C$  získáme pomocí následujících substitucí

$$\begin{aligned} A &= -\{(1+w_2)\hat{\mu}_1 + (1+w_1)\hat{\mu}_2\}, \\ B &= 2\hat{\mu}_1\hat{\mu}_2 + w_2s_1^2 + w_1s_2^2, \\ C &= -(w_1\hat{\mu}_1s_2^2 + w_2\hat{\mu}_2s_1^2), \end{aligned} \quad (3.58)$$

kde  $w_1 = n/(n+m)$ ,  $w_2 = m/(n+m)$ ,  $\hat{\mu}_1 = (1/n) \sum_{j=1}^n y_j$ ,  $\hat{\mu}_2 = (1/m) \sum_{j=n+1}^{n+m} y_j$  a

$$s_1^2 = \frac{1}{n} \sum_{j=1}^n y_j^2, \quad s_2^2 = \frac{1}{m} \sum_{j=n+1}^{n+m} y_j^2. \quad (3.59)$$

Rovnici 3.57 lze dále redukovat na tvar

$$\lambda^3 + 3p\lambda + q = 0, \quad (3.60)$$

kde  $\lambda = \mu + A/3$ ,  $p = (3B - C^2)/9$  a  $q = (2A^3 - 9AB + 27C)/27$ . Tato rovnice má následující trojici řešení,

$$\lambda = \sqrt[3]{\alpha} + \sqrt[3]{\beta}, \quad \omega \sqrt[3]{\alpha} + \omega^2 \sqrt[3]{\beta}, \quad \omega^2 \sqrt[3]{\alpha} + \omega \sqrt[3]{\beta}, \quad (3.61)$$

kdy  $\alpha, \beta$  a  $\omega$  jsou

$$\omega = \frac{-1 + \sqrt{3}i}{2}, \quad (3.62)$$

$$\alpha, \beta = \frac{-q \pm \sqrt{q^2 + 4p^3}}{2}. \quad (3.63)$$

Pro ukázkou můžeme vzít dvě množiny pocházející z normálního rozdělení, každou o deseti pozorování. První má parametry  $\mu = 10$  a  $\sigma^2 = 1$ , zatímco druhá  $\mu = 3$  a  $\sigma^2 = 4$ . Předem lze předpokládat, že odhady střední hodnoty a rozptylu nebudou příliš přesné. V Tabulce 3.6 vidíme numerické výsledky výpočtů. Není příliš překvapivé, že minimální

Model	AIC	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$\mu_1 = \mu_2 \quad \sigma_1^2 = \sigma_2^2$	104,372	6,786	6,786	8,853	8,853
$\mu_1 \neq \mu_2 \quad \sigma_1^2 = \sigma_2^2$	66,785	9,548	4,024	1,223	1,223
$\mu_1 = \mu_2 \quad \sigma_1^2 \neq \sigma_2^2$	391,409	5,782	5,782	0,451	1,994
$\mu_1 \neq \mu_2 \quad \sigma_1^2 \neq \sigma_2^2$	63,709	9,548	6,073	0,451	1,994

Tabulka 3.6: Výsledky modelů z příkladu 3.4.2

hodnota AIC se nachází u modelu, kde všechny 4 parametry jsou od sebe rozdílné. Je dobré si také všimnout toho, jak se odhady jednotlivých parametrů od sebe mění, jakmile na ně položíme další podmínky.

### 3.5 Mallowsovo $C_p$

Můžeme představit další kritérium na posouzení kvality modelu, které jak zjistíme, je spojené s Akaikeho informačním kritériem. Mallowsovo  $C_p$  kritérium se stejně jako AIC snaží zabránit situaci známé jako *overfitting*, ke které dochází pokud používáme  $R^2$  koeficient jako jediné kritérium kvality. Předpokládejme, že máme  $n$  množin pozorování  $\{(y_\alpha, \mathbf{x}_\alpha); \alpha = 1, \dots, n\}$  náhodné proměnné  $Y$  a  $p$  vysvětlujících proměnných  $x_1, \dots, x_p$ . Dále předpokládejme, že střední hodnota a kovarianční matice  $n$ -rozměrného vektoru pozorování  $\mathbf{y} = (y_1, \dots, y_n)^T$  je

$$E[\mathbf{y}] = \boldsymbol{\mu}, \quad D(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T] = \omega^2 I_n, \quad (3.64)$$

kde  $I_n$  je  $n \times n$  jednotková matice. Střední hodnotu  $\boldsymbol{\mu}$  lze odhadnout pomocí lineárního regresního modelu

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad D(\boldsymbol{\varepsilon}) = \sigma^2 I_n, \quad (3.65)$$

kde  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$  a  $X$  je  $n \times (p+1)$  matice plánu. Pak pomocí odhadu metody nejmenších čtverců  $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$  vektoru  $\boldsymbol{\beta}$  lze odhadnout  $\boldsymbol{\mu}$  jako

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y}, \quad (3.66)$$

kdy  $H$  značí projekční matici (anglicky také známou jako *hat matrix*). Pro určení efektivnosti odhadu můžeme použít střední kvadratickou chybu

$$\Delta_p = E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})]. \quad (3.67)$$

Vzhledem k tomu že střední hodnota odhadu  $\hat{\boldsymbol{\mu}}$  je

$$E[\hat{\boldsymbol{\mu}}] = X(X^T X)^{-1} X^T E[\mathbf{y}] = H\boldsymbol{\mu}, \quad (3.68)$$

můžeme vyjádřit střední kvadratickou chybu 3.67 ve tvaru

$$\begin{aligned} \Delta_p &= E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})] \\ &= E\left[\{H\mathbf{y} - H\boldsymbol{\mu} - (I_n - H)\boldsymbol{\mu}\}^T \{H\mathbf{y} - H\boldsymbol{\mu} - (I_n - H)\boldsymbol{\mu}\}\right] \\ &= E\left[(\mathbf{y} - \boldsymbol{\mu})^T H(\mathbf{y} - \boldsymbol{\mu})\right] + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu} \\ &= \text{tr}\{HD(\mathbf{y})\} + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu} \\ &= (p+1)\omega^2 + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu}, \end{aligned} \quad (3.69)$$

kde jsme využili skutečnosti, že  $H$  a  $I_n - H$  jsou idempotentní matice a tedy platí, že  $H^2 = H$ ,  $(I_n - H)^2 = I_n - H$ ,  $H(I_n - H) = 0$  a  $\text{tr}(H) = \text{tr}\{X(X^T X)^{-1} X^T\} = \text{tr}(I_{p+1}) = p+1$ ,  $\text{tr}(I_n - H) = n - p - 1$ . Můžeme si všimnout, že v poslední rovnosti 3.69 bude první člen,  $(p+1)\omega^2$ , růst s počtem proměnných v našem modelu. Naopak člen druhý,  $\boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu}$ , který můžeme interpretovat jako sumu čtverců vychýlení odhadu  $\hat{\boldsymbol{\mu}}$ , bude s rostoucím počtem parametrů klesat. Tedy pokud jsme schopni odhadnout  $\Delta_p$ , tak to lze použít jako kritérium pro posouzení kvality modelu. Dále spočítáme střední hodnotu sumy čtverců reziduí

$$\begin{aligned} E[SSE_p] &= E[(\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})] \\ &= E[(\mathbf{y} - H\mathbf{y})^T (\mathbf{y} - H\mathbf{y})] \\ &= E[\{(I_n - H)(\mathbf{y} - \boldsymbol{\mu}) + (I_n - H)\boldsymbol{\mu}\}^T \{(I_n - H)(\mathbf{y} - \boldsymbol{\mu}) + (I_n - H)\boldsymbol{\mu}\}] \\ &= E[(\mathbf{y} - \boldsymbol{\mu})^T (I_n - H)(\mathbf{y} - \boldsymbol{\mu})] + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu} \\ &= \text{tr}\{(I_n - H)V(\mathbf{y})\} + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu} \\ &= (n - p - 1)\omega^2 + \boldsymbol{\mu}^T (I_n - H)\boldsymbol{\mu}. \end{aligned} \quad (3.70)$$

Srovnáme-li 3.70 s 3.69 a předpokládáme, že  $\omega^2$  je známé, dostaneme nestranný odhad  $\Delta_p$  jako

$$\hat{\Delta}_p = SSE_p + \{2(p+1) - n\}\omega^2. \quad (3.71)$$

Vydělíme-li nyní obě strany rovnice odhadem parametru  $\omega^2$ , dostaneme Mallowsovo  $C_p$  kritérium ve tvaru

$$C_p = \frac{SSE_p}{\hat{\omega}^2} + \{2(p+1) - n\}. \quad (3.72)$$

Podobně jako u AIC platí, že čím menší hodnota  $C_p$ , tím lepší je náš model. Jako odhad parametru  $\omega^2$  se často bere nestranný odhad chyby rozptylu plného modelu.

**Příklad 3.5.1.** Opět použijeme sadu dat **Swiss** na demonstrování Mallowsova  $C_p$  kritéria a zároveň srovnáme výsledek i s AIC. Pro tuto demonstraci použijeme funkci z knihovny **olsrr** (viz R skript v Příloze), která testuje všechny možné lineární kombinace parametrů a vypočítá příslušné kritérium – mezi nimi AIC, BIC,  $R^2$ , Mallowsovo  $C_p$  a mnoho dalších. Tabulka 3.7 ukazuje numerické výsledky, kdy sloupec **mindex** je pořadové číslo modelu a **n** je počet proměnných v modelu. Uvažujeme zde pouze modely s **n** větší jak 3.

mindex	n	predictors	cp	aic
16	3	Ed+Cat+Inf	8,18	328,67
17	3	Agr+Ed+Cat	11,01	331,41
18	3	Ex+Ed+Inf	14,25	334,36
19	3	Ex+Ed+Cat	20,44	339,53
20	3	Agr+Ed+Inf	21,66	340,49
21	3	Agr+Ex+Ed	22,95	341,47
22	3	Ex+Cat+Inf	25,18	343,13
23	3	Agr+Ex+Inf	25,34	343,25
24	3	Agr+Ex+Cat	38,44	351,96
25	3	Agr+Cat+Inf	46,86	356,81
26	4	Agr+Ed+Cat+Inf	5,03	325,24
27	4	Ex+Ed+Cat+Inf	9,99	330,48
28	4	Agr+Ex+Ed+Cat	11,96	332,41
29	4	Agr+Ex+Ed+Inf	12,72	333,13
30	4	Agr+Ex+Cat+Inf	26,64	344,74
31	5	Agr+Ex+Ed+Cat+Inf	6,00	326,07

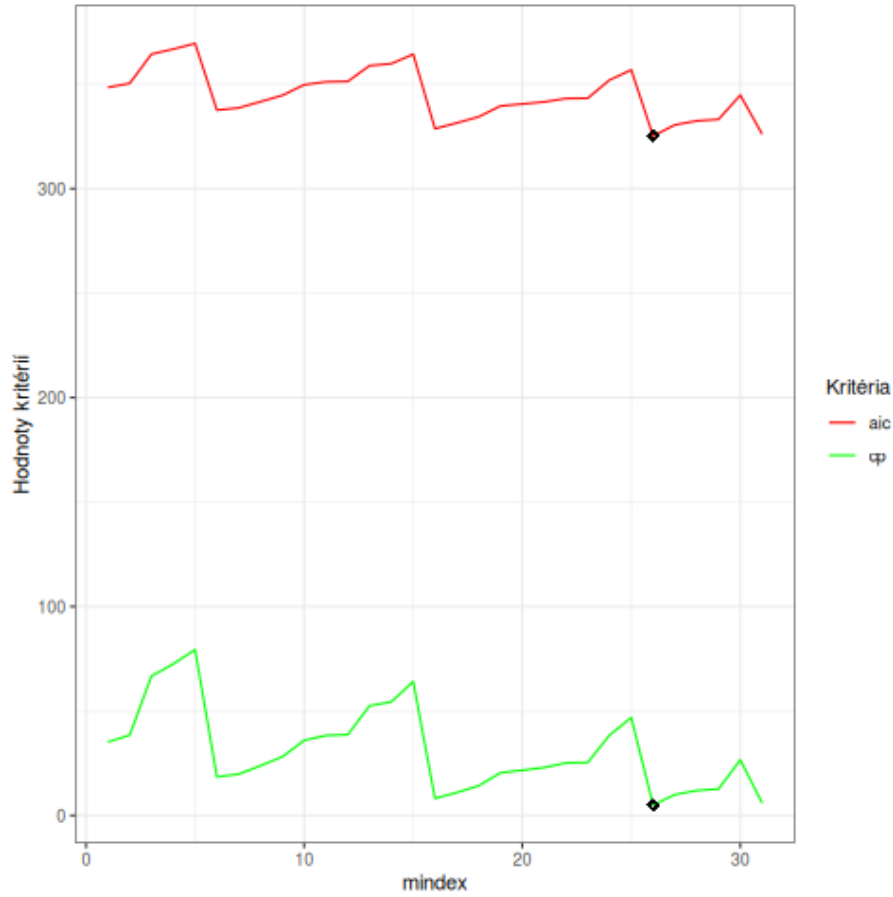
Tabulka 3.7: AIC vs  $C_p$  v Příkladu 3.5.1

Jako jména proměnných jsou zde uvedena první 2–3 písmena názvu. Je důležité si všimnout toho, že AIC i  $C_p$  mají obě minimální hodnotu u stejného modelu. Grafické znázornění změny hodnot jednotlivých kritérií najdeme na Obrázku 3.8, kde je vykreslena i minimální hodnota obou metod, opět v tom samém bodě. Obecně platí, že Mallowsovo  $C_p$  je ekvivalentní AIC v rámci normálních lineárních regresních modelů. Důkaz tohoto tvrzení lze najít například v [7].

## 3.6 Analýza hlavních komponent

Častokrát je pro nás výhodné zredukovat počet proměnných vektorů pozorování. K tomu nás mohou vést záležitosti čistě technické, jako výpočetní doba a složitost algoritmů pro výpočet odhadů, intervalů spolehlivosti, anebo nás zajímá pouze znát kovariance (a rozptyly) mezi danými proměnnými. V takovém případě je jednodušší pracovat s nižším počtem proměnných. Dochází tedy k redukci dimenze našich pozorování, což automaticky s sebou nosí ztrátu určitého množství informace v původním vektoru a je pochopitelně naším cílem tuto ztrátu minimalizovat. Jednou z takových metod je analýza hlavních komponent (anglicky *principal component analysis* – PCA). Ukážeme si základní odvození PCA a





Obrázek 3.8: Grafické vývoje AIC a  $C_p$  kritéria u jednotlivých modelů v Příkladu 3.5.1. Jsou zde zvýrazněné i minimální hodnoty obou kritérií

poté se na ní zaměříme jak z úhlu výběru modelu pomocí AIC, tak i z pohledu čistě informačně-teoretického.

Mějme tedy náhodný vektor  $\mathbf{x} = (x_1, \dots, x_p)^T$  s nulovou střední hodnotou ( $E\mathbf{x} = 0$ ) a s kovarianční maticí  $D\mathbf{x} = \mathbf{\Sigma}$ . Budou nás zajímat lineární transformace vektoru  $\mathbf{x}$  ve tvaru  $Y_i = c_i^T \mathbf{x}$ , kdy  $Y_i$  nazveme  $i$ -tou hlavní komponentou vektoru  $\mathbf{x}$ . Chceme aby tyto nové proměnné byly nekorelované a zároveň zachovávaly celkovou variabilitu vektoru  $\mathbf{x}$ . Taktéž chceme aby první komponenta v sobě obsahovala co nejvíc informace o  $\mathbf{x}$ , tj. aby maximalizovala jeho rozptyl. Tedy maximalizujeme rozptyl

$$D(Y_1) = D(c_1^T \mathbf{x}) = c_1^T \mathbf{\Sigma} c_1 \quad (3.73)$$

přes všechny hodnoty  $c_1$ . Pochopitelně je třeba položit nějakou podmínku na  $c_1$ , aby optimalizační úloha dávala smysl. Typicky se klade podmínka  $c_1^T c_1 = 1$ , tj. aby  $c_1$  mělo jednotkovou normu. Optimalizační úlohu můžeme například řešit pomocí Lagrangeových multiplikátorů, kdy budeme hledat maximum funkce

$$c_1^T \mathbf{\Sigma} c_1 - \lambda (c_1^T c_1 - 1). \quad (3.74)$$

Zderivujeme-li 3.74 podle  $c_1$ , dostaneme

$$\mathbf{\Sigma} c_1 - \lambda c_1 = \mathbf{0}, \quad (3.75)$$

neboli

$$(\mathbf{\Sigma} - \lambda I_p) c_1 = \mathbf{0}, \quad (3.76)$$

kdy  $I_p$  je  $p \times p$  jednotková matice. Z výrazu 3.76 plyne, že  $c_1$  je vlastní vektor matice  $\mathbf{\Sigma}$  s příslušným vlastním číslem  $\lambda$ . Abychom zjistili, které z  $p$  vlastních čísel zde máme, tak pouze dosadíme do vztahu, který maximalizujeme, tj.

$$c_1^T \mathbf{\Sigma} c_1 = c_1^T \lambda c_1 = \lambda c_1^T c_1 = \lambda, \quad (3.77)$$

kdy jsme v poslední rovnosti dosadili normalizační podmínku na vektor  $c_1$ . Tedy  $c_1$  je vlastní vektor největšího vlastního čísla  $\lambda = \lambda_1$  a zároveň platí

$$DY_1 = D(c_1^T \mathbf{x}) = D(c_1^T \mathbf{\Sigma} c_1) = \lambda_1. \quad (3.78)$$

Komponenta  $Y_2$  musí být nekorelovaná s  $Y_1$  a zároveň musí vysvětlovat co nejvíce ze zbývajících variability. Tedy platí že

$$C(Y_1, Y_2) = C(c_1^T \mathbf{X}, c_2^T \mathbf{X}) = c_1^T \mathbf{\Sigma} c_2 = c_2^T \mathbf{\Sigma} c_1 = c_2^T c_1 = 0. \quad (3.79)$$

Podobně pro rozptyl  $Y_2$  platí, že  $DY_2 = D(c_2^T \mathbf{x}) = c_2^T \mathbf{\Sigma} c_2$ . Opět budeme maximalizovat tento rozptyl za podmínky, že  $c_2^T c_2 = 1$ . Stejným postupem dostaneme, že  $c_2$  je vlastní vektor odpovídající druhému největšímu vlastnímu číslu  $\lambda_2$  a že  $DY_2 = \lambda_2$ . Analogicky tento postup aplikujeme na všech  $p$  hlavních komponent. Je třeba zmínit, že pochopitelně matici  $\mathbf{\Sigma}$  neznáme a je třeba jí odhadnout, a to výběrovou kovarianční maticí  $\mathbf{S}$ .

Ovšem účelem této transformace je redukovat dimenzi našich pozorování, a tedy nechceme použít všech  $p$  hlavních komponent. To si klade otázku, dle jakého kritéria určíme, kolik hlavních komponent je třeba pro odpovídající reprezentaci našich dat. Nejčastěji se používá *scree plot*, kdy hledáme bod, kde se láme křivka závislosti vysvětlené variability na počtu hlavních komponent, nebo nás zajímá kumulativní součet vlastních čísel  $\left( \frac{\lambda_1 + \lambda_2 + \dots}{\sum_{i=1}^p \lambda_i} \right)$  a požadujeme součet nad  $\geq 80$  %. Podrobnější analýzu a srovnání těchto kritérií s mnoha dalšími najdeme například v článku [11].

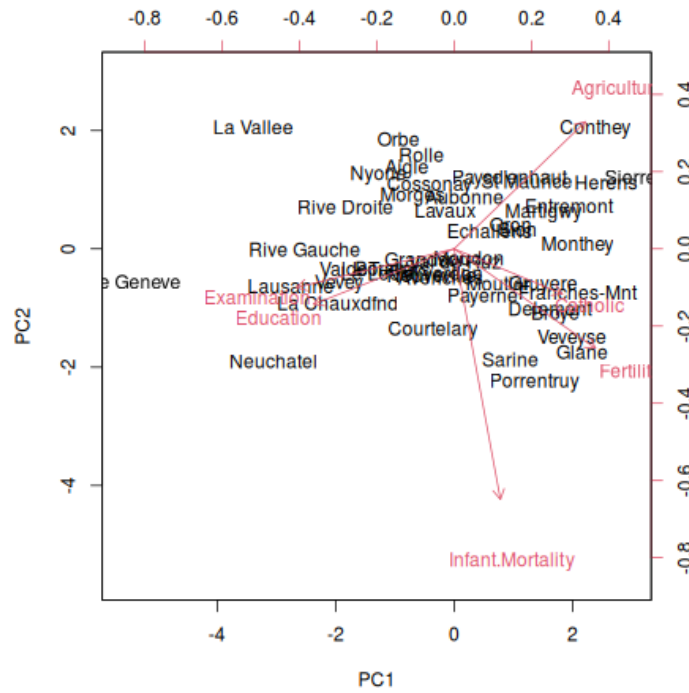
**Příklad 3.6.1.** Použijeme-li opět data **Swiss** a provedeme na nich analýzu hlavních komponent, dostaneme následující výsledky. V Tabulce 3.8 vidíme hodnoty  $c_i$  jednotlivých hlavních komponent pro všechny proměnné v **Swiss**. Pokud bychom se pokusili interpre-

Proměnné	PC1	PC2	PC3	PC4	PC5	PC6
Fertility	0,46	-0,32	0,17	-0,54	-0,38	0,47
Agriculture	0,42	0,41	-0,04	0,64	-0,37	0,31
Examination	-0,51	-0,13	0,09	0,05	-0,81	-0,22
Education	-0,45	-0,18	-0,53	0,10	0,07	0,68
Catholic	0,35	-0,15	-0,81	-0,10	-0,18	-0,40
Infant.Mortality	0,15	-0,81	0,16	0,53	0,10	-0,07

Tabulka 3.8: Hodnoty hlavních komponent pro proměnné v souboru Swiss data v R

tovat první hlavní komponentu, vidíme u ní silnou zápornou hodnotu u **Examination** a

**Education** a naopak silné kladné hodnoty u **Fertility** a **Agriculture**. Mohli bychom říct, že první hlavní komponenta nám rozděluje územní celky na ty zaměřené na zemědělství a mající vysokou porodnost a naopak na ty kde úroveň vzdělání je nízká. Podobně bychom postupovali pro další komponenty. Tento postup pochopitelně není exaktní, naopak je více subjektivní záležitostí a proto je třeba zpětně ověřit, zda tato rozdělení dávají smysl pro naše data. Je vždy vhodné se podívat na grafické znázornění (anglicky *biplot*). Použijeme zde k tomu první dvě hlavní komponenty pro jednoduchost vykreslení, které lze vidět na Obrázku 3.9.



Obrázek 3.9: Biplot prvních dvou komponent souboru Swiss data v R

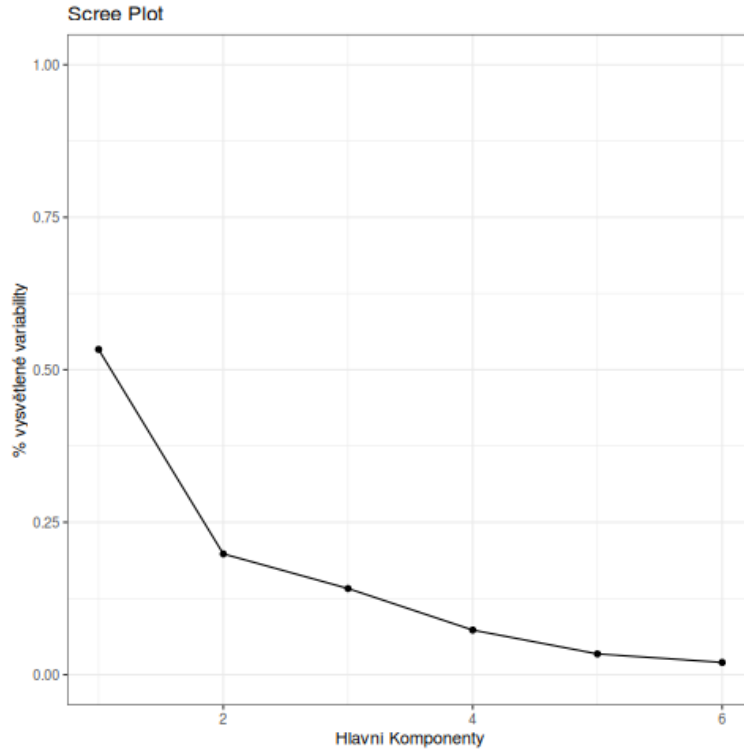
*Scree plot* vidíme na obrázku 3.10.

Je vidět, že pokud bychom chtěli mít aspoň 80 % vysvětlené variability, je třeba použít první 3 hlavní komponenty. Pokud ovšem nám bude stačit pouze 70 % variability, spokojíme se s dvěma komponentami.

V následující sekci odvodíme AIC pro analýzu hlavních komponent v podobě, kterou nalezneme v článku [12]. Nejprve musíme identifikovat, který model budeme vůbec uvažovat a srovnávat s dalšími. Definujme model  $M_j$  jako

$$M_j : \lambda_1 > \lambda_2 > \dots > \lambda_j > \lambda_{j+1} = \dots = \lambda_p, \quad (3.80)$$

kdy  $\lambda_i$  jsou vlastní hodnoty matice  $\Sigma$ . Tedy bude-li model  $M_j$  ten "skutečný", tak počet hlavních komponent v něm použitých bude právě  $j$ . Budeme chtít pomocí AIC vybrat model z množiny  $\{M_0, \dots, M_{p-1}\}$ . Nechtě jsou  $\mathbf{x}_1, \dots, \mathbf{x}_N$  vektory pozorování, kde  $N = n + 1$ , mající  $p$ -rozměrné normální rozdělení  $N_p(\boldsymbol{\mu}, \Sigma)$  a dále ať platí  $p < n$ . Označme  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ .



Obrázek 3.10: Scree plot vysvětlené variability u PCA pro Swiss data

Označme věrohodnostní funkci  $\mathbf{X}$  jako  $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  a maximálně věrohodné odhady těchto parametrů u modelu  $M_j$  jako  $\hat{\boldsymbol{\mu}}_j$  a  $\hat{\boldsymbol{\Sigma}}_j$ . Pak platí, že  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = (1/N) \sum_{i=1}^N \mathbf{x}_i$  a tedy

$$-2L(\hat{\boldsymbol{\mu}}_j, \boldsymbol{\Sigma}) = N \ln |\boldsymbol{\Sigma}| + n \text{tr} \boldsymbol{\Sigma}^{-1} \mathbf{S} + pN \ln(2\pi), \quad (3.81)$$

kde matice  $\mathbf{S}$  je naše výběrová kovarianční matice ve tvaru

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \quad n = N - 1. \quad (3.82)$$

Nechť dále jsou  $l_1, \dots, l_p$  vlastní čísla matice  $\mathbf{S}$  a necht'  $\mathbf{h}_i, i = 1, \dots, p$ , jsou příslušné vlastní vektory. Označme  $\mathbf{L} = \text{diag}(l_1, \dots, l_p)$  a  $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_p)$ . Pak maximálně věrohodný odhad parametru  $\boldsymbol{\Sigma}$  za platnosti modelu  $M_j$  je

$$\hat{\boldsymbol{\Sigma}}_j = \frac{n}{N} \mathbf{S}_j, \quad \mathbf{S}_j = \mathbf{H} \begin{pmatrix} \mathbf{L}_1 & \mathbf{O} \\ \mathbf{O} & \bar{l}_{jp} \mathbf{I}_{p-j} \end{pmatrix} \mathbf{H}^T, \quad (3.83)$$

kde

$$\mathbf{L}_1 = \text{diag}(l_1, \dots, l_j), \quad \bar{l}_{jp} = \frac{1}{p-j} \sum_{i=j+1}^p l_i. \quad (3.84)$$

Odvození tohoto odhadu můžeme najít například v článku [4]. Dohromady jsme získali

$$\begin{aligned} \text{AIC}_j = N \ln(l_1 \cdots l_p) + N(p-j) \ln(\bar{l}_{p-j}) + 2d_j \\ + N \ln \left( \frac{n}{N} \right)^p + Np(\ln(2\pi) + 1). \end{aligned} \quad (3.85)$$

Zde  $d_j$  značí počet nezávislých parametrů, který odvodíme v následující podobě. Nechť  $\boldsymbol{\gamma}_i$  je vlastní vektor matice  $\boldsymbol{\Sigma}$  odpovídající vlastnímu číslu  $\lambda_i$ , tak že  $\boldsymbol{\gamma}_i$  pro  $i = 1, \dots, p$  jsou ortonormální. Nechť

$$\begin{aligned}\boldsymbol{\Lambda} &= \text{diag}(\lambda_1, \dots, \lambda_p), & \boldsymbol{\Lambda}_1 &= \text{diag}(\lambda_1, \dots, \lambda_j), \\ \boldsymbol{\Gamma} &= (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p) = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2), & \boldsymbol{\Gamma}_1 &= (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_j).\end{aligned}$$

Pak lze vyjádřit matici  $\boldsymbol{\Sigma}$  jako

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^T = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2) \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{O} \\ \mathbf{O} & \lambda I_{p-j} \end{pmatrix} (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)^T \quad (3.86)$$

$$= \boldsymbol{\Gamma}_1 (\boldsymbol{\Lambda}_1 - \lambda I_j) \boldsymbol{\Gamma}_1^T + \lambda I_p. \quad (3.87)$$

V poslední rovnosti jsme použili skutečnost, že  $\boldsymbol{\Gamma}_2 \boldsymbol{\Gamma}_2^T = I_p - \boldsymbol{\Gamma}_1 \boldsymbol{\Gamma}_1^T$ . Tedy dimenze  $\{\boldsymbol{\Sigma}, \boldsymbol{\mu}\}$  u modelu  $M_j$  je rovna dimenzi  $\{\boldsymbol{\Gamma}_1, \boldsymbol{\Lambda}_1, \lambda, \boldsymbol{\mu}\}$ , která je dána jako

$$d_j = pj - \frac{1}{2}j(j+1) + j + 1 + p. \quad (3.88)$$

Můžeme si ovšem usnadnit práci tím, že nás ve výsledku zajímá pouze minimální hodnota AIC a nikoliv  $\text{AIC}_j$ , takže lze uvažovat místo toho  $A_j$  ve tvaru

$$A_j = \text{AIC}_j - \text{AIC}_{p-1}, \quad j = 0, \dots, p-1, \quad (3.89)$$

$$A_j = -N \left\{ \sum_{i=j+1}^p \ln(l_i) - (p-1) \ln(\bar{l}_{jp}) \right\} - 2q_j, \quad (3.90)$$

kde  $q_j = \frac{1}{2}(p-j-1)(p-j+2)$  a  $A_{p-1} = 0$ . Poté hledáme takové  $j$ , které minimalizuje jeho hodnotu. Bohužel, kritérium v této podobě není konzistentní, jak je uvedeno v článku [12]. Máme-li fixní  $p$  a necháme  $n \rightarrow \infty$ , tak pravděpodobnost, že  $A_j$  vybere správný model roste, ale nekonverguje k 1 (pro  $n = 5000$  máme pravděpodobnost 84.3 %). Obecně také platí, že  $A_j$  má tendenci vybírat komplikovanější model s pravděpodobností 14.5 % ~ 15.5 %. Situace kdy  $p, n \rightarrow \infty$  a  $p/n \rightarrow c \in (0; 1)$  je více rozebrána do detailu jak v [12], tak i v navazujícím článku [6].

Na závěr si můžeme zmínit zajímavost, které se netýká nutně AIC, ale teorie informace jako takové. Nejprve si definujme pojem vzájemné informace.

**Definice 3.6.1.** Buďte  $X$  a  $Y$  spojité náhodné veličiny. Pak jejich vzájemná informace se definuje jako

$$I(X; Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \log \left( \frac{f(x, y)}{f(x)f(y)} \right) dx dy, \quad (3.91)$$


kde  $f(x, y)$  je sdružená hustota pravděpodobnosti a  $f(x), f(y)$  jsou marginální hustoty pravděpodobnosti. V diskrétním případě to pouze nahradíme sumou.

Vzájemnou informaci můžeme interpretovat tak, že je to veličina, která měří kolik informace  $X$  a  $Y$  sdílí. Tedy pokud jsou tyto náhodné veličiny nezávislé, tak mají nulovou vzájemnou informaci. Budeme-li uvažovat náš vektor pozování jako  $\mathbf{x} = \mathbf{s} + \mathbf{n}$ , kdy  $\mathbf{s}$  je

vektor nosící v sobě žádanou informaci a  $\mathbf{n}$  je vektor šumu či chyb. Pokud dále předpokládáme, že  $\mathbf{s}$  má normální rozdělení a  $\mathbf{n}$  má také normální rozdělení s kovarianční maticí, která je proporcionální jednotkové matici, pak PCA maximalizuje vzájemnou informaci  $I(Y; \mathbf{s})$  mezi vektorem  $\mathbf{s}$  a vektorem  $Y = c^T \mathbf{x}$  poté co proběhl výběr hlavních komponent a došlo k redukci  $Y$  na požadovanou dimenzi. Důkaz tohoto tvrzení najdeme v článku [19].

# Závěr

V rámci této práce jsme nejprve stručně uvedli několik definic z oblasti teorie informací, z nichž nejdůležitější je *Kullback-Leiblerova* divergence (respektive informace), která je stěžejní pro informační kritéria jako taková. V další kapitole jsme rozebrali tematiku metody maximální věrohodnosti a její spojitosti s *Kullback-Leiblerovou* informací, což je základem pro použití teorie informace pro hodnocení statistických modelů, u nichž lze spočítat jejich maximální věrohodnost. Dále jsme provedli důkladné odvození Akaikeho informačního kritéria, kde jsme upozornili na hlavní chyby u jeho využití. Nakonec v třetí kapitole jsme se věnovali aplikacím AIC jak u lineárních a polynomiálních regresních modelů, kde je AIC nejčastěji zmiňováno a použito, ale i též u histogramů, srovnání dvou výběrů z diskrétních rozdělení a posledně výběru počtu hlavních komponent v analýze hlavních komponent.

Za nejdůležitější poznatky považujeme skutečnost, že AIC (a další informační kritéria) nám pomáhají ve výběru modelu na základě jeho věrohodnosti, která je penalizována na základě počtu proměnných, tj. penalizují komplexitu modelu, aby se zabránilo *over-fittingu*. AIC se nejčastěji používá pro hodnocení lineárních regresních modelů, přestože jsme zde ukázali, že ho lze použít na větší třídu statistických modelů. Při zkoumání vlastností AIC jsme zjistili, že AIC je citlivé na počet pozorování a nebylo vždy zaručené, že vybere „správný” model. Veškeré výpočty a grafy v třetí kapitole byly provedeny v statistickém softwaru  [22] a skript je uveden v elektronické příloze jako samostatný soubor.





# Příloha

V elektronické příloze této práce je uložen skript R-code.R, kde jednotlivé sekce jsou odděleny symbolem # s krátkým popisem jejich funkce. Tento skript obsahuje výpočty, které byly použity v práci. Dále se tam nachází funkce pro výpočet AIC histogramu a Obrázky, které se nacházejí v práci.



# Seznam použité literatury

- [1] Aho, K., Derryberry, D., Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3), 631–636.
- [2] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*, 199–213.
- [3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.
- [4] Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1), 122–148.
- [5] Ash, Robert B. *Information theory*. New York: Dover Publications, 1990. xi, 339. ISBN 0486665216.
- [6] Bai, Z., Choi, K. P., Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, 46(3), 1050–1076.
- [7] Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W., Wells, M. T. (2013). AIC, Cp and estimators of loss for elliptically symmetric distributions. *arXiv preprint arXiv:1308.2766*.
- [8] Anderson, D., Burnham, K. (2004). *Model selection and multi-model inference*. Second. NY: Springer-Verlag, 63(2020), 10.
- [9] Burnham, K. P., Anderson, D. R., Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65, 23–35.
- [10] Cover, T. M., Thomas, J. A. (2006). *Elements of information theory* 2nd edn (Hoboken, NJ, John Wiley Sons).
- [11] Ferré, L. (1995). Selection of components in principal component analysis: a comparison of methods. *Computational Statistics Data Analysis*, 19(6), 669–682.
- [12] Fujikoshi, Y., Sakurai, T. (2016). Some properties of estimation criteria for dimensionality in principal component analysis. *American Journal of Mathematical and Management Sciences*, 35(2), 133–142.

- [13] Hartley, R. V. (1928). Transmission of information 1. Bell System technical journal, 7(3), 535–563.
- [14] Hebbali, A. (2020, February 10). Tools for Building OLS Regression Models [R package *olsrr* version 0.5.3]. <https://cran.r-project.org/web/packages/olsrr/index.html>, [cit. 27.04.2023]
- [15] Hurvich, C. M., Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307.
- [16] Ing, C. K., Wei, C. Z. (2005). Order selection for same-realization predictions in autoregressive processes.
- [17] Konishi, S., Kitagawa, G. (2008). Information criteria and statistical modeling.
- [18] Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.
- [19] Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3), 105–117.
- [20] Mildenerger, T., Rozenholc, Y., Zasada, D. (2019, April 26). Construction of Regular and Irregular Histograms with Different Options for Automatic Choice of Bins [R package *histogram* version 0.0–25]. <https://cran.r-project.org/web/packages/histogram/index.html>, [cit. 27.04.2023]
- [21] Nyquist, H. (1924). Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, 43, 412–422.
- [22] R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, [cit. 27.04.2023]
- [23] Scott, D. W. (2009). Sturges’ rule. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3), 303–306.
- [24] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
- [25] Sugiura, N. (1978). Further analysis of the data by Akaike’s information criterion and the finite corrections: Further analysis of the data by akaike’s. *Communications in Statistics-theory and Methods*, 7(1), 13–26.
- [26] Swiss fertility and socioeconomic indicators (1888), R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, [cit. 27.04.2023]
- [27] Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)* 153 12–18.

- [28] Vapor Pressure of Mercury as a Function of Temperature, R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, [cit. 27.04.2023]
- [29] Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2), 228.
- [30] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1), 60–62.



