MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

# Bakalářská práce

BRNO 2013 VERONIKA BENDOVÁ

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

# Regresní diagnostika v chemometrickém modelu

Bakalářská práce

**Veronika Bendová**

**Vedoucí práce: RNDr. Marie Forbelská, Ph.D.** **Brno 2013**

# Bibliografický záznam

| | |
|---|---|
| **Autor:** | Veronika Bendová<br>Přírodovědecká fakulta, Masarykova univerzita<br>Ústav matematiky a statistiky |
| **Název práce:** | Regresní diagnostika v chemometrickém modelu |
| **Studijní program:** | Matematika |
| **Studijní obor:** | Statistika a analýza dat |
| **Vedoucí práce:** | RNDr. Marie Forbelská, Ph.D. |
| **Akademický rok:** | 2012/2013 |
| **Počet stran:** | $10 + 66$ |
| **Klíčová slova:** | lineární regresní model; metoda nejmenších čtverců; metoda maximální věrohodnosti; EM algoritmus; analýza reziduí; odlehlé body; vybočující body; shluková analýza; koeficient determinace; programovací jazyk R; |

# Bibliographic Entry


| | |
|---|---|
| **Author:** | Veronika Bendová<br>Faculty of Science, Masaryk University<br>Department of Mathematics and Statistics |


| | |
|---|---|
| **Title of Thesis:** | Regression diagnostics in chemometric model |
| **Degree Programme:** | Mathematics |
| **Field of Study:** | Statistics and Data Analysis |
| **Supervisor:** | RNDr. Marie Forbelská, Ph.D. |
| **Academic Year:** | 2012/2013 |
| **Number of Pages:** | $10 + 66$ |
| **Keywords:** | linear regression model; ordinary least squared method; maximum likelihood; EM algorithm; analysis of residues; outliers; high leverage points; cluster analysis; coefficient of determination; programming language R; |

# Abstrakt

Cílem této bakalářské práce je zlepšit přesnost predikce dvou lineárních regresních modelů, které předpovídají hodnotu disociační konstanky pKa z hodnot nábojů určitých atomů v molekule. Provedená analýza dat sestává ze dvou hlavních částí: analýzy reziduí a shlukové analýzy, na základě které byly vytvořeny dva nové obecné lineární regresní modely (General Linear Models). Práce sestává ze tří kapitol. V první a druhé kapitole jsou vysvětleny všechny důležité pojmy, které byly při vyhodnocování dat použity. Třetí kapitola podrobně popisuje uskutečněnou datovou analýzu. Data byla vyhodnocována pomocí programovacího jazyka R. Součástí práce jsou mimo jiné popisy použitých příkazů.

# Abstract

The aim of this bachelor thesis is to improve the prediction accuracy of two linear regression models, which determine the acid dissociation constant pKa according to the charges of specific atoms in a molecule. The data analysis comprises of two parts: the analysis of residues and cluster analysis, the latter of which serves as the basis for formulating two new general linear models. The thesis consists of three chapters. The first and the second chapter provide the theoretical background of the notions, which were used to investigate the data. The third chapter thoroughly describes the analyses performed. For calculations with the data R programming language was used. The description of the utilised commands is contained in the thesis.

# Masarykova univerzita

# Přírodovědecká fakulta

## ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: **Veronika Bendová**

Studijní program - obor: **Matematika - Statistika a analýza dat**

Ředitel Ústavu matematiky a statistiky PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s tématem:

### Regresní diagnostika v chemometrickém modelu

**Regression diagnostics in chemometric model**

*Oficiální zadání:* Cílem bakalářské práce je prostudovat a zpracovat tématiku diagnostiky odlehlých či "podezřelých" pozorování v chemometrických regresních modelech a vyložené postupy následně demonstrovat ve vhodném statistickém programovacím prostředí (například v R) na simulovaných či reálných datech.

*Literatura: Doporučená literatura*

*VARMUZA, Kurt a Peter FILZMOSER. Introduction to multivariate statistical analysis in chemometrics. Boca Raton, Fla.: CRC Press, 2009. xiii, 321. ISBN 1420059475.*

*Vedoucí bakalářské práce:* RNDr. Marie Forbelská, Ph.D.

*Datum zadání bakalářské práce:* květen 2012

*Datum odevzdání bakalářské práce:* dle harmonogramu ak. roku 2012/2013

V Brně dne 31.10.2012

v.z. prof. RNDr. Jiří Rosický, DrSc.
Ředitel Ústavu matematiky a statistiky

Zadání bakalářské práce převzal dne: 3. 12. 2012

Podpis studenta

# Poděkování

Touto cestou bych ráda poděkovala RNDr. Marii Forbelské, Ph.D za odborné vedení mé bakalářské práce, za cenné rady a náměty týkající se analýzy dat a za pomoc při tápání v nepřeberném množství funkcí, které programovací jazyk R nabízí. Dále bych chtěla poděkovat RNDr. Radce Svobodové Vařekové, Ph.D. za skvělé téma mé bakalářské práce, za její nadšení pro věc a velmi motivující přístup.

Chtěla bych také poděkovat svým rodičům za korekturu textu, nejvíce však za velikou podporu, ohleduplnost a trpělivost, kterou jsem od nich během svého studia cítila. Moc ráda bych poděkovala svému příteli Peťovi, který mi mnohokrát pomohl, za všechny jeho úsměvy, milá slova a moudré myšlenky, za to, že mě podporoval a dal mi najevo, že je tu pro mě, ve chvílích, kdy jsem to nejvíce potřebovala. Jsem moc ráda, že vás všechny mám.

Na závěr bych ještě ráda poděkovala spolubydlící Zuzce za antistresové debaty a pracovní morálku, kterou jsme dodržovaly při psaní našich výstupních prací, svému notebooku, že mi po celou dobu mého studia připravoval vhodné a spolehlivé softwarové zázemí, knihařství Jihlava za svázání této práce a učebních textů, které jsem ke studiu potřebovala, tvůrcům seriálu Teorie velkého třesku a her Age of Empires a v neposlední řadě i všem mým oblíbeným výrobcům čokolády.

# Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracovala samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 27. května 2013 .....

Veronika Bendová

# Obsah

**Úvod** ..... **viii**

**Kapitola 1. Lineární regrese** ..... **1**
1.1 Lineární regresní model ..... 1
1.1.1 Odhady parametrů v lineárním regresním modelu ..... 3
1.1.2 Koeficient determinace ..... 4
1.2 Regresní diagnostika, rezidua ..... 5
1.2.1 Analýza reziduí ..... 5
1.2.2 Regresní grafy ..... 9
1.3 Rozšířený lineární regresní model a vážená metoda nejmenších čtverců 12

**Kapitola 2. Metoda maximální věrohodnosti, EM algoritmus** ..... **14**
2.1 Metoda maximální věrohodnosti ..... 14
2.2 EM algoritmus ..... 15
2.2.1 EM algoritmus ..... 16
2.2.2 Aplikace EM algoritmu ..... 17

**Kapitola 3. Praktická část - Zpracování dat** ..... **21**
3.1 3d.modely ..... 22
3.1.1 Seznámení s 3d.modelem ..... 22
3.1.2 Analýza reziduí ..... 25
3.1.3 Shluková analýza ..... 28
3.1.4 gls 3d.modely ..... 37
3.2 5d.modely ..... 44
3.2.1 Seznámení s 5d.modelem ..... 44
3.2.2 Shluková analýza ..... 47
3.2.3 gls 5d.model ..... 47

**Závěr** ..... **54**

**Příloha A - Informace o cd** ..... **56**

**Příloha B - Algoritmus generující grafy hustot smíšených rozdělení** ..... **57**

**Příloha C - Přehled R-kových příkazů** ..... **61**

**Příloha D - Seznam obrázků** ..... **62**

**Příloha E - Seznam tabulek** ..... **63**

**Seznam použité literatury** ..... **64**

# Úvod

V posledních letech dochází k velmi významnému rozvoji výpočetní techniky a s tím i k rozšíření možností aplikace matematických a statistických metod do nejrůznějších vědních disciplín. Zejména důležité je uplatnění v oblasti chemie a biologie. S rozvojem informačních technologií a nárůstem množství dat ohledně struktury molekul vzniká také nový vědní obor nazvaný chemoinformatika. Hlavním cílem tohoto oboru je převedení chemických problémů pomocí matematických vztahů na problémy výpočetní a algoritmické. Na základě těchto vztahů potom můžeme například předpovídat některé chemické vlastnosti látek, aniž bychom museli tyto látky složitě a často i velmi draze syntetizovat. Z toho důvodu má chemoinformatika bohaté využití především ve farmaceutickém průmyslu. Zde je však více než kde jinde důležité, aby výsledky, které přináší, byly co nejvíce přesné a spolehlivé.

Základním stavebním kamenem chemické látky jsou molekuly. Ty jsou rovněž nositeli chemických vlastností látky. Jednou ze sledovaných vlastností každé molekuly je hodnota disociační konstanty (značí se pKa) a to proto, že na ní závisí mnoho dalších chemických vlastností (např. rozpustnost, propustnost atp.). Experimentální měření disociační konstanty je však velmi náročné, a tak se jedním z klíčových chemoinformatických témat výzkumu stala právě predikce hodnoty disociační konstanty.

Bylo zjištěno, že hodnota pKa závisí na velikostech parciálních nábojů na určitých atomech molekuly. Na základě této informace byly navrženy dva (lineární) matematické modely, které je možné využít k predikci této konstanty. První (3d) model využívá k predikci parciální náboje na atomech vodíku (qH), kyslíku (qO) a uhlíku (qC1), druhý (5d) model využívá kromě již zmíněných nábojů ještě navíc parciální náboj na uhlíku (qC1D) a kyslíku (qOD) v rámci disociované molekuly (t.j. molekuly s odštěpeným atomem vodíku). Získání hodnot těchto nábojů není výpočetně tak složité, jako získání samotné hodnoty pKa. 3d.model predikuje hodnotu disociační konstanty dané molekuly s přesností (vyjádřenou pomocí koeficientu determinace) od 0.86 do 0.91 (v závislosti na metodě, kterou byly náboje na atomech dopočítány), 5d.model s přesností v rozmezí 0.89–0.92. Cílem této bakalářské práce je zvýšit přesnost predikce obou modelů.

Text práce je rozdělen do tří kapitol. První a druhá kapitola jsou teoretické a poskytují čtenáři shrnutí matematických pojmů využitých při analýze dat. Kroky samotné analýzy jsou potom náplní třetí kapitoly.

V první kapitole se zaměříme na lineární regresi a lineární regresní model. Zavedeme pojem nejlepší nestranný lineární odhad a popíšeme si postup odhadování

parametrů tzv. metodou nejmenších čtverců. Dále si zadefinujeme koeficient determinace, což je hodnota určující kvalitu lineárního regresního modelu. Druhou částí této kapitoly je regresní diagnostika. V ní vysvětlíme rozdíl mezi odlehlými a vybočujícími pozorováními, uvedeme si několik typů reziduí a pojem cookovy vzdálenosti. Nakonec si řekneme pár slov o rozšířeném lineárním regresním modelu a vážené metodě nejmenších čtverců.

Druhá kapitola popisuje další hojně využívanou metodu sloužící k odhadu parametrů a sice metodu maximální věrohodnosti. Také si představíme EM algoritmus a jeho aplikace: odhadování parametrů daného modelu a shlukování dat, které jsou ze smíšeného rozdělení, do skupin.

Třetí kapitola je potom ryze praktická. Provádí nás uskutečněnou analýzou dat, která byla provedena pomocí programovacího jazyka R. Kromě podrobného popisu použitých příkazů a funkcí jsou v ní uvedeny nejdůležitější mezivýsledky a závěry, které z těchto výsledků vyplývají.

# Kapitola 1

# Lineární regrese

Regresní model je jeden z nejdůležitějších statistických modelů. Jeho úkolem je dostatečně přesně popsat reálnou situaci matematickými vztahy. Základní pravidlo pro hledání vhodného modelu říká, že žádný model není správný, ale některé modely jsou užitečné. Při sestavování regresního modelu bychom měli mít na paměti, že užitečný model má splňovat dvě základní podmínky:

- musí dostatečně přesně popisovat reálnou situaci,
- neměl by být přehnaně komplikovaný.

Shrnuto do jedné věty: Čím jednodušší a výstižnější je model, tím je lepší. Naneštěstí pro nás, v praxi není nikdy nic tak snadné, proto se většinou snažíme o dosažení rozumného kompromisu mezi jednoduchostí a výstižností, vždy na úkor druhé vlastnosti.

## 1.1 Lineární regresní model

Předpokládejme nejprve deterministický vztah mezi nenáhodnými veličinami $x_1, \ldots, x_k$, jež si sami volíme, a nenáhodnou veličinou $y$, která na těchto veličinách závisí. Tento vztah můžeme vyjádřit rovnicí:

$$y = \beta_1 x_1 + \cdots + \beta_k x_k, \tag{1.1}$$

kde $\beta_1, \ldots, \beta_k$ jsou neznámé *parametry*. Hodnoty $x_1, \ldots, x_k$ nazýváme *regresory*.

Bližší informace o neznámých parametrech $\beta_1, \ldots, \beta_k$ získáme tak, že si zvolíme hodnoty regresorů $x_1, \ldots, x_k$ a pro ně experimentálně změříme hodnotu $y$. Tento postup opakujeme, dokud nemáme dostatečně velkou sadu $(k+1)$-tic hodnot.

Při praktickém měření veličiny $y$ ale vznikají tzv. *chyby měření*. Značíme je $\varepsilon$ a jejich hodnota není obecně známá. Deterministický vztah obohacený o chybovou složku $\varepsilon$, tedy vztah tvaru

$$y = \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \tag{1.2}$$

budeme nazývat *stochastický vztah*.

Pro jednoduchost si nyní stanovme, že jsme provedli $n$ měření (i s chybami), čili že máme k dispozici celkem $n$ nyní už $(k+2)$-tic:

$$\begin{aligned} Y_1 &= \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ Y_2 &= \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \,. \end{aligned} \tag{1.3}$$

Dostáváme tak *stochastický model*[1]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.4}$$

Je-li soustava rovnic (1.3) lineární v parametrech $\beta_1, \ldots, \beta_k$ (v proměnných $x_1, \ldots, x_k$ lineární být nemusí), potom modelu (1.4) říkáme *lineární stochastický model*. Maticově bychom soustavu (1.3) zapsali takto:

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

kde $\mathbf{X}$ o rozměrech $n \times k$ je tzv. *matice plánu*. Chyby měření $\varepsilon_1, \ldots, \varepsilon_n$ nejsou sice obecně známé, ovšem předpokládejme o nich, že

- jsou náhodné,
- jsou nesystematické (jejich střední hodnota je nulová, tj. $E\varepsilon_i = 0$ pro $i = 1, \ldots, n$),
- mají homogenní rozptyl ($D\varepsilon_i > \sigma^2$ pro $i = 1, \ldots, n$),
- vzájemně jsou nekorelované ($C(\varepsilon_i, \varepsilon_j) = 0$ pro $i, j = 1, \ldots, n.\ i \neq j$)

*Poznámka.* Z předpokladů o chybách měření vyplývají následující vlastnosti:

1. $E\mathbf{Y} = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + \underbrace{E\boldsymbol{\varepsilon}}_{\mathbf{0}} = \mathbf{X}\boldsymbol{\beta}$,
2. $D\mathbf{Y} = D(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = D\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}_n$, kde $\mathbf{I}_n$ je *jednotková matice*.

Lineární stochastický model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

s výše uvedenými vlastnostmi nazýváme *lineární regresní model*. V tomto textu jej budeme značit $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, kde $\mathcal{L}$ je zástupný symbol pro typ rozdělení,

[1] Stochastický model bez vektoru chybových měření, tedy model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, nazýváme *deterministický model*.

$\mathbf{X}\boldsymbol{\beta}$ značí vektor středních hodnot a $\sigma^2\mathbf{I}_n$ varianční matici daného rozdělení. Např. výrazem $\mathbf{Y} \sim N_m(\mathbf{0}, \mathbf{I}_m)$ označujeme lineární regresní model z $m$-rozměrného normálního rozdělení s nulovým vektorem středních hodnot a jednotkovou varianční maticí.

Předpokládejme dále, že $n > k$ a hodnost matice $\mathbf{X}$ (značíme $h(\mathbf{X})$) je rovna $k$, tedy $h(\mathbf{X}) = k$. Lineární regresní model s touto vlastností budeme nazývat *lineární regresní model plné hodnosti*.

### 1.1.1 Odhady parametrů v lineárním regresním modelu

Připomeňme, že máme $n$ $(k+2)$-tic $(Y_1, x_{11}, \ldots, x_{1k}, \varepsilon_1), \ldots, (Y_n, x_{n1}, \ldots, x_{nk}, \varepsilon_n)$ a k nim vytvořený lineární regresní model

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

se (zatím neznámými) parametry $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$.

Naším úkolem je odhadnout hodnoty těchto neznámých parametrů, tedy najít vektor $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$, kde $\hat{\beta}_i$ je odhad parametru $\beta_i$ pro $\forall i = 1 \ldots k$.
Přitom budeme požadovat, aby

- odhad $\hat{\boldsymbol{\beta}}$ byl lineárním odhadem vektoru parametrů $\boldsymbol{\beta}$,
- odhad $\hat{\boldsymbol{\beta}}$ byl nestranným odhadem vektoru parametrů $\boldsymbol{\beta}$.

**Definice 1.1.** Řekneme, že vektor $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ je **lineárním odhadem** vektoru $\boldsymbol{\beta}$, jestliže existuje matice reálných čísel $\mathbf{B}_{k\times n}$ taková, že $\hat{\boldsymbol{\beta}} = \mathbf{B}\mathbf{Y}$.

**Definice 1.2.** Odhad $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nestranným odhadem** vektoru $\boldsymbol{\beta}$, pokud pro každé $\boldsymbol{\beta} \in \mathbb{R}^k$ platí $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

**Definice 1.3.** Odhad $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nejlepší nestranný lineární odhad** vektoru parametrů $\boldsymbol{\beta}$, pokud je lineární, nestranný a platí, že pro každý jiný nestranný lineární odhad $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\mathbf{Y})$ je rozdíl variančních matic $D\tilde{\boldsymbol{\beta}}(\mathbf{Y}) - D\hat{\boldsymbol{\beta}}(\mathbf{Y})$ pozitivně semidefinitní matice.

Nejlepší nestranný lineární odhad budeme hledat takzvanou *metodou nejmenších čtverců*.

**Definice 1.4.** Vektor $\hat{\boldsymbol{\beta}}_{MNČ}$ je odhadem vektoru parametrů $\boldsymbol{\beta}$ **metodou nejmenších čtverců**, pokud platí

$$\hat{\boldsymbol{\beta}}_{MNČ} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

**Věta 1.1.** *Odhad vektoru parametrů $\boldsymbol{\beta}$ v modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ je tvaru*

$$\hat{\boldsymbol{\beta}}_{MNČ} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

*Důkaz.* Viz přednáškový text [6]. □

**Věta 1.2** (Gauss-Markovova)**.** *Mějme model* $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$*. Odhad* $\hat{\boldsymbol{\beta}}_{MNČ}$ *je nejlepším nestranným lineárním odhadem vektoru parametrů* $\boldsymbol{\beta}$*.*

*Důkaz.* Viz přednáškový text [6]. □

Pomocí nalezených odhadů $\hat{\beta}_1, \ldots, \hat{\beta}_k$ můžeme nyní zpětně dopočítat vektor odhadů středních hodnot ($\hat{\mathbf{Y}}$)

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{MNČ}.$$

### 1.1.2 Koeficient determinace

Kvalitu lineárního regresního modelu můžeme zhodnotit pomocí tzv. *koeficientu determinace*. K jeho výpočtu si prvně musíme zadefinovat pojmy *reziduální* a *celkový součet čtverců*.

**Věta 1.3** (O reziduálním součtu čtverců)**.** *Nechť* $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *je lineární regresní model a* $\hat{\boldsymbol{\beta}}_{MNČ}$ *je odhad vektoru parametrů* $\boldsymbol{\beta}$ *metodou nejmenších čtverců. Potom* ***reziduální součet čtverců***

$$S_e = (\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})' = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'_{MNČ}\mathbf{X}'\mathbf{Y}.$$

*Důkaz.* Důkaz věty lze najít například v literatuře [2]. □

**Věta 1.4.** *Odhad* $s^2 = \frac{S_e}{n-k}$ *je nestranným odhadem rozptylu* $\sigma^2$*.*

*Důkaz.* Viz literatura [2]. □

*Celkový součet čtverců* $S_T$ získáme vztahem

$$S_T = (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})' = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2, \tag{1.5}$$

kde $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ je tzv. *výběrový průměr*.

*Koeficient determinace* $R^2$ má potom tvar

$$R^2 = 1 - \frac{S_e}{S_T}. \tag{1.6}$$

Čím je hodnota koeficientu determinace blíže jedné, tím lépe námi vytvořený model popisuje reálnou situaci a odhady středních hodnot $\hat{Y}_i$ se více blíží k původním experimentálně naměřeným hodnotám $Y_i$ ($i = 1, \ldots, n$).

*Poznámka.* Hodnota koeficientu determinace $R^2$ je ovlivňována počtem regresorů vystupujících v lineárním regresním modelu. Z tohoto důvodu se v praxi pro určení přesnosti modelu používá i tzv. *korigovaný (adjustovaný) koeficient determinace*, který se snaží zmíněný vliv regresorů eliminovat. Jeho hodnotu dopočítáme pomocí vzorce

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}\frac{S_e}{S_T}, \tag{1.7}$$

kde $k$ je počet regresorů v modelu a $n$ je počet všech pozorování.

*Poznámka.* Hlubší teorie k tématu koeficientů determinace, která je však nad rámec této práce, je popsána v přednáškovém textu [7].

## 1.2 Regresní diagnostika, rezidua

Najít vyhovující lineární regresní model, který by dával optimální výsledky, není jednoduché. Každý model, který nám připadá vhodný k popisu dané reálné situace, je proto potřeba důkladně ověřit. Takové ověřování je náplní *regresní diagnostiky*. Mezi hlavní úkoly regresní diagnostiky patří

- prozkoumání kvality dat pro navržený model,
- ověření kvality modelu pro daná data,
- splnění základních předpokladů metody nejmenších čtverců.

My se zaměříme na první část, tedy prozkoumání kvality dat. Metody pro ověření kvality modelu a splnění přepokladů metody nejmenších čtverců jsou podrobně rozepsány v literatuře [13].

Kvalita dat významně ovlivňuje vlastnosti regresního modelu. Při bližším zkoumání dat se zaměřujeme přednostně na ta pozorování, která jsou výrazněji odlišná od ostatních. Taková pozorování často mívají významný vliv na okolní data, a proto bývají zdrojem problémů, např. zkreslení odhadů parametrů. Odlišná pozorování nazýváme *vlivnými body* a dělíme je podle polohy výskytu na

- *outliers* - body, které se od ostatních liší svou polohou na ose y,
- *high leverage points* - body, které se odlišují na ose x.

*Poznámka.* Česká terminologie není v názvosloví typů vlivných pozorování ustálená. Např. v publikaci [13] jsou (high) leverage points přeloženy jako *extrémy* a outliers jako *vybočující pozorování*. My budeme v tomto textu používat jinou terminologii a sice outliery budeme nazývat *odlehlá pozorování* a leverage points *vybočující pozorování*. Každopádně je vhodné znát jednoznačné anglické ekvivalenty.

*Poznámka.* Při praktickém zpracování se můžeme setkat i s pozorováními, která jsou zároveň odlehlá i vybočující.

### 1.2.1 Analýza reziduí

Vraťme se nyní k našemu modelu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Už známe hodnoty odhadů ($\hat{\boldsymbol{\beta}}_{MN\check{C}}$) všech parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ a víme, že pro vektor odhadů středních hodnot $\hat{\mathbf{Y}}$ platí

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{MN\check{C}},$$

kde $\hat{\boldsymbol{\beta}}_{MN\check{C}} = (\hat{\beta}_{MN\check{C}_1}, \ldots, \hat{\beta}_{MN\check{C}_k})$ jsou nyní již konkrétní hodnoty dopočítané metodou nejmenších čtverců.

*Poznámka.* Protože v této kapitole budeme používat k odhadu vektoru parametrů $\boldsymbol{\beta}$ výhradně metodu nejmenších čtverců, budeme odteď pro lepší přehlednost vektor

$$\hat{\boldsymbol{\beta}}_{MNČ} = (\hat{\beta}_{MNČ_1}, \ldots, \hat{\beta}_{MNČ_k})$$

značit pouze $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots \hat{\beta}_k)$.

Pro vektor $\hat{\mathbf{Y}}$ tedy platí:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}, \tag{1.8}$$

kde

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{pmatrix} h_{11} & \ldots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \vdots & h_{nn} \end{pmatrix}$$

je tzv. *projekční (hat) matice*.

Odhad ($\hat{Y}_i$) $i$-tého pozorování je potom tvaru

$$\hat{Y}_i = \begin{pmatrix} h_{i1} & \ldots & h_{in} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \sum_{j=1}^{n} h_{ij} Y_j.$$

*Poznámka.* Projekční matice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ je

1. symetrická ($\mathbf{H}' = \mathbf{H}$):

$$\mathbf{H}' = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}.$$

2. idempotentní ($\mathbf{H}^2 = \mathbf{H}$):

$$\begin{aligned} \mathbf{H}^2 = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')^2 &= \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_n}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H} \end{aligned}$$

**Definice 1.5.** Sloupce projekční matice $\mathbf{H}$ nazýváme **vlivové vektory** (značíme je $\mathbf{H}_j, j = 1, \ldots, n$). Symbolem $\|\mathbf{H}_j^2\|$ označujeme **vliv j-tého pozorování** (**j-tý vliv**) na odhad vektoru $\hat{\mathbf{Y}}$.

**Věta 1.5.** *V lineárním regresním modelu* $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti platí*

1. $\|\mathbf{H}_j\|^2 = h_{jj}$,
2. $0 \leq h_{ii} \leq 1$, *pro* $i = 1, \ldots, n$.

*Důkaz.* 1.

$$\|\mathbf{H}_j\|^2 = \mathbf{H}_j'\mathbf{H}_j = \sum_{i=1}^{n} h_{ij}h_{ij} \;\wedge\; \mathbf{H}^2 = \mathbf{H} \Rightarrow h_{jj} = \mathbf{H}_j'\mathbf{H}_j$$

2.

$$\mathbf{H}^2 = \mathbf{H} \Rightarrow h_{jj} > 0 \Rightarrow h_{jj} = \sum_{i=1}^{n} h_{ij}h_{ij} = h_{jj}^2 + \sum_{\substack{i=1 \\ i\neq j}}^{n} h_{ij}^2 \Rightarrow h_{jj} > h_{jj}^2 \Rightarrow h_{jj} < 1.$$

□

Jelikož $\hat{\mathbf{Y}}$ je pouze odhadem střední hodnoty vektoru $\mathbf{Y}$, budou se tyto dva vektory od sebe mírně lišit. Rozdíl těchto dvou vektorů je dobrým ukazatelem kvality námi vytvořeného modelu. Čím menší je rozdíl $\mathbf{Y} - \hat{\mathbf{Y}}$, tím je náš model přesnější.

**Definice 1.6.** Mějme lineární regresní model $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Vektor

$$\mathbf{r} = \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

nazýváme **vektor reziduí**.

**Věta 1.6.** *V lineárním regresním modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ plné hodnosti platí, že*

$$\mathbf{r} = \hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}. \tag{1.9}$$

*Důkaz.*

$$\begin{aligned}\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \\ = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta}) + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} - \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = \\ = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon},\end{aligned}$$

protože $\mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_n}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$. □

**Věta 1.7.** *V lineárním regresním modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ plné hodnosti platí*

1. $D\hat{\mathbf{Y}} = \sigma^2\mathbf{H}$,
2. $D\mathbf{r} = \sigma^2(\mathbf{I} - \mathbf{H})$.

*Důkaz.* 1.

$$D\hat{\mathbf{Y}} = D(\mathbf{H}\mathbf{Y}) = \mathbf{H}\underbrace{(D\mathbf{Y})}_{\sigma^2}\mathbf{H}' = \sigma^2\mathbf{H}\mathbf{H}' = \sigma^2\underbrace{\mathbf{H}^2}_{\mathbf{H}} = \sigma^2\mathbf{H}$$

2.

$$\begin{aligned}D\mathbf{r} = D(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\underbrace{D\boldsymbol{\varepsilon}}_{\sigma^2}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = \sigma^2\underbrace{(\mathbf{I} - \mathbf{H})^2}_{\mathbf{I}-\mathbf{H}} \\ = \sigma^2(\mathbf{I} - \mathbf{H}).\end{aligned}$$

□

Vztah 1.9 si můžeme rozepsat pomocí matic následovně

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} 1-h_{11} & -h_{12} & \dots & -h_{1n} \\ -h_{21} & 1-h_{22} & \dots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \dots & 1-h_{nn} \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Hodnota $i$-tého rezidua má tedy tvar

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j = \varepsilon_i(1-h_{ii}) - \sum_{\substack{j=1 \\ j\neq i}}^{n} h_{ij}\varepsilon_j. \tag{1.10}$$

Ze vztahu 1.10 je dobře viditelné, že pokud se hodnota prvku $h_{ii}$ blíží k 1 (píšeme $h_{ii} \approx 1$), potom $(1-h_{ii}) \approx 0$. Jestliže chyba $i$-tého pozorování $\varepsilon_i$ bude velká a $h_{ii} \approx 1$, neprojeví se tato chyba nijak významně v hodnotě $i$-tého rezidua $r_i$. V takovém případě je dobré znát další typy reziduí, která, i přestože velikost vlivu $h_{ii} \approx 1$, chybu $\varepsilon_i$ odhalí. Než si je přestavíme, zavedeme si *model vynechaného pozorování*.

Zvolme si pevně index $i$ a uvažujme model bez $i$-tého pozorování. Takový model budeme značit

$$\mathbf{Y}_{(i)} \sim \mathcal{L}(\mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)}, \sigma^2\mathbf{I}_{n-1}).$$

Odhad vektoru parametrů $\boldsymbol{\beta}_{(i)}$ metodou nejmenších čtverců ($\hat{\boldsymbol{\beta}}_{(i)_{MNČ}}$, odteď pouze $\hat{\boldsymbol{\beta}}_{(i)}$) je tvaru

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}$$

a odhad střední hodnoty pro $i$-té pozorování $\hat{Y}_{(i)}$ se rovná

$$\hat{Y}_{(i)} = \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(i)},$$

kde $\mathbf{x}_i$ je $i$-tý řádek původní matice plánu $\mathbf{X}$.

**Definice 1.7.** Předpokládejme, že $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ je lineární regresní model. Potom **standardizovaná (normovaná) rezidua** $r_i^*$ jsou rovna

$$r_i^* = \frac{r_i}{s\sqrt{1-h_{ii}}}, \qquad \text{kde } s^2 = \frac{(\mathbf{Y}-\hat{\mathbf{Y}})'(\mathbf{Y}-\hat{\mathbf{Y}})}{n-k} = \frac{S_e}{n-k},$$

**predikovaná rezidua** $r_{(i)}$ můžeme vyjádřit jako

$$r_{(i)} = Y_i - \hat{Y}_{(i)},$$

**studentizovaná (jackknife) rezidua** $r_{(i)}^*$ mají tvar

$$r_{(i)}^* = \frac{r_{(i)}\sqrt{1-h_{ii}}}{s_{(i)}}, \qquad \text{kde } s_{(i)}^2 = \frac{(\mathbf{Y}_{(i)}-\hat{\mathbf{Y}}_{(i)})'(\mathbf{Y}_{(i)}-\hat{\mathbf{Y}}_{(i)})}{n-k-1}$$

a **DFIT rezidua** $d_i$ se definují vztahem

$$d_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{h_{ii}} s_{(i)}}.$$

Nadefinovali jsme si tedy celkem pět typů reziduí. Následující věta nám ukáže, jak vypočítat hodnoty $r_{(i)}$, $s^2_{(i)}$ a $r^*_{(i)}$ pouze pomocí hodnot známých přímo z modelu. To nám umožňuje spočítat tyto statistiky, aniž bychom museli pro každé $i$ zdlouhavě určovat model bez $i$-tého pozorování.

**Věta 1.8.** *Platí*

1. $r_{(i)} = \dfrac{r_i}{1 - h_{ii}}$,
2. $(n-k-1)s^2_i = (n-k)s^2 - \dfrac{r_i^2}{1-h_{ii}}$,
3. $r^*_{(i)} = \dfrac{r_i}{\sqrt{1 - h_{ii} s_{(i)}}}$.

*Důkaz.* Viz přednáškový text [6]. □

Důležitou vlastností každého pozorování je (kromě výše zmíněné polohy) jeho vliv na celkový charakter daného modelu. Z konstrukce lineárního regresního modelu je zřejmé, že jednotlivá pozorování více či méně ovlivňují výsledný odhad vektoru parametrů $\boldsymbol{\beta}$. Odpověď na otázku, jak velká je míra vlivu $i$-tého pozorování na odhad vektoru $\boldsymbol{\beta}$, nám dává tzv. *Cookova vzdálenost*.

**Definice 1.8.** Nechť $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ je lineární regresní model. Statistika $C_i$

$$C_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{k s^2} = \frac{(r^*_{(i)})^2}{k} \frac{h_{ii}}{1 - h_{ii}} \tag{1.11}$$

se nazývá **Cookova vzdálenost**.

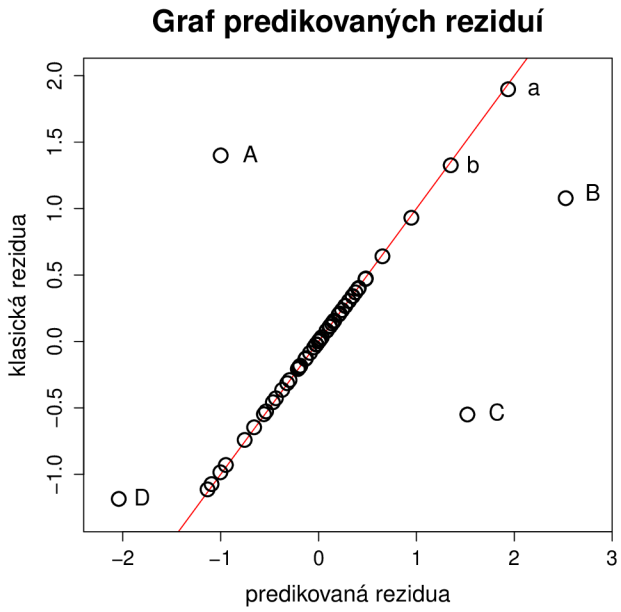
*Poznámka.* Když se dobře podíváme na vzorec 1.11, můžeme si všimnout, že Cookova vzdálenost je vlastně euklidovská vzdálenost mezi vektorem odhadů středních hodnot získaných pomocí metody nejmenších čtverců ($\hat{\mathbf{Y}}$) a vektorem odhadů středních hodnot vypočítaných (též metodou nejmenších čtverců) z modelu bez $i$-tého pozorování ($\hat{\mathbf{Y}}_{(i)}$). Bod považujeme za vlivný, pokud je hodnota $C_i > 1$.

### 1.2.2 Regresní grafy

Při hledání vlivných bodů nám může velmi významně pomoci grafické znázornění. Regresní diagnostika nabízí v tomto ohledu široké spektrum nejrůznějších grafů, které lze využít k odhalení podezřelých pozorování. My si zde uvedeme pouze některé z nich, a to převážně ty, které jsou dále použity v kapitole **3**. Další typy jsou popsané v [13].

### Graf predikovaných reziduí

Při konstrukci tohoto grafu vynášíme na osu $x$ predikovaná rezidua $r_{(i)}$ oproti klasickým reziduím $r_i$ na ose $y$. Graf predikovaných reziduí velmi snadno identifikuje oba typy vlivných pozorování. Běžná pozorování jsou prezentována body rovnoměrně rozprostřenými na přímce $x = y$. Outliery (odlehlá pozorování) sice také leží na přímce $x = y$, ovšem od ostatních bodů jsou výrazně vzdálené. Vybočující pozorování leží zcela mimo přímku $x = y$.

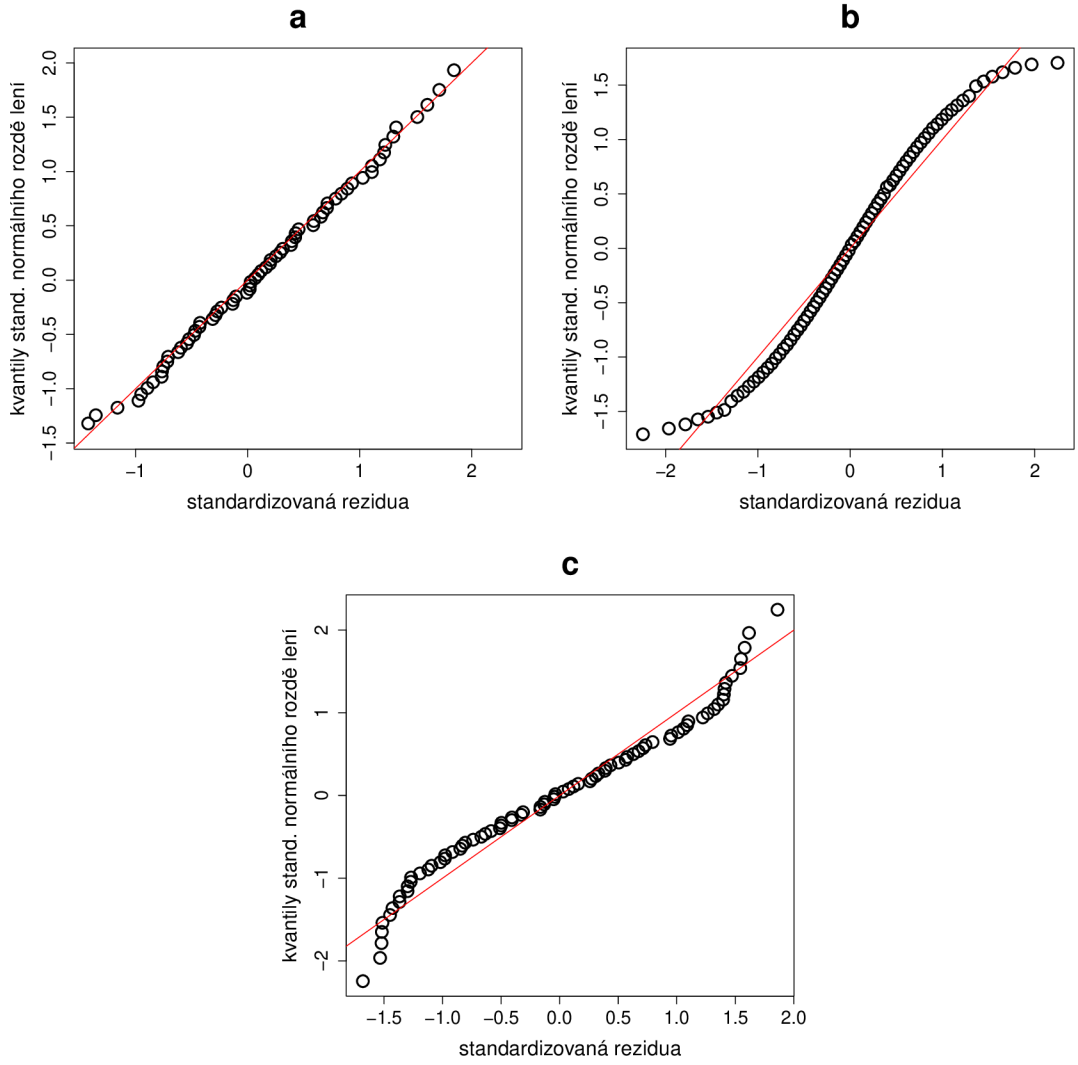



Obrázek 1.1: Graf predikovaných reziduí

Body $A, B, C, D$ vyznačené v obrázku 1.1 jsou vybočující, zatímco body $a, b$ bychom mohli klasifikovat jako odlehlé.

### Rankitové (Q-Q) grafy

Rankitové grafy se využívají k ověření normality rozdělení datového souboru. Na osu $x$ zakreslíme vzestupně setříděné hodnoty standardizovaných reziduí $r_i^*$ a na osu $y$ kvantily standardizovaného normálního rozdělení $u_{P_i}$, přičemž $P_i = \frac{i}{n+1}$, kde $i = 1 \dots n$ a $n$ je počet všech pozorování. V případě, že data pocházejí z normálního rozdělení, vyskytují se body v grafu rovnoměrně okolo přímky $x = y$. Na obrázku 1.2 jsou ilustrované tři Q-Q grafy pro data z různých typů rozdělení.

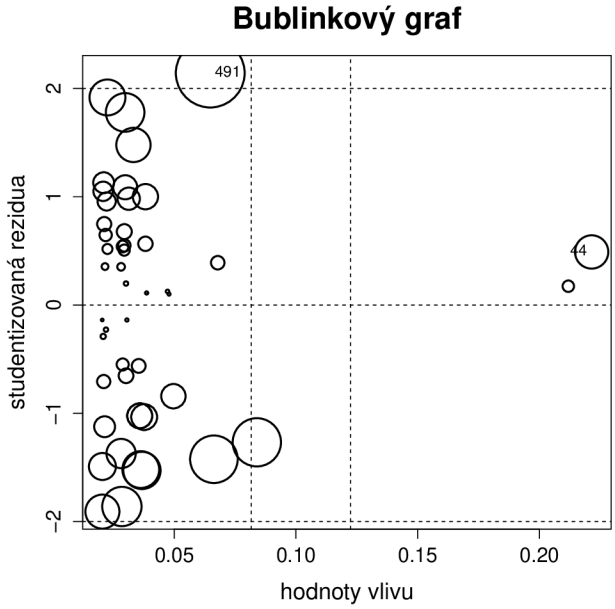



Obrázek 1.2: Rankitový graf pro: a) normální rozdělení, b) rozdělení s dlouhými konci, c) rozdělení s krátkými konci

*Poznámka.* Další grafy můžeme získat vynesením ostatních typů reziduí na osu $x$.

### InfluencePlot (Bublinkový graf)

Bublinkový graf vyvoříme nanesením hodnot vlivů $h_{ii}$ na osu $x$ a hodnot studentizovaných reziduí $r^*_{(i)}$ na osu $y$. Každý bod je zde reprezentován kruhem, jehož obsah je úměrný velikosti Cookovy vzdálenosti $C_i$. Dále do grafu zakreslíme tři horizontální referenční čáry v bodech $-2, 0, 2$ a dvě vertikální čáry jako mezníky vlivnosti jednotlivých pozorování. Vertikální čáry budou vykresleny v hodnotách $2\frac{k}{n}$ a $3\frac{k}{n}$. Ta pozorování, která leží mezi oběma vertikálními přímkami, jsou *vlivná*, ta, která leží napravo od druhé vetrikální přímky, jsou *silně vlivná*, a konečně body, které se nacházejí nad horizontální přímkou v bodě $2$ nebo pod horizontální přímkou v bodě $-2$, klasifikujeme jako *odlehlá*.




Obrázek 1.3: Bublinkový graf

Z obrázku 1.3 dobře vidíme, že bod $44$ je silně vlivný a bod $491$ odlehlý. Všimněme si, že tento bod má i největší Cookovu vzdálenost.

## 1.3 Rozšířený lineární regresní model a vážená metoda nejmenších čtverců

V odstavci 1.1 jsme si představili lineární regresní model $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ se střední hodnotou $E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ a homogenním rozptylem $D\mathbf{Y} = \sigma^2\mathbf{I}_n$. Zaměřme se nyní na případ lineárního regresního modelu, jehož rozptyl není homogenní. S takovou situací se můžeme setkat například tehdy, máme-li smíšený soubor dvou, nebo více typů dat, která jsou z navzájem různých rozdělení. Pracovat v tomto případě s homoskedastickým[2] modelem by nebylo vhodné.

Model s heterogenním rozptylem budeme značit

$$\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}), \qquad \mathbf{W} > 0, \sigma^2 > 0,$$

kde $\mathbf{W}$ je (zpravidla známá) pozitivně definitní matice (značíme $\mathbf{W} > 0$) a $\boldsymbol{\beta}$ a $\sigma^2$ jsou neznámé parametry.

**Věta 1.9** (Aitkenův odhad)**.** *Nechť $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1})$ je lineární regresní model plné hodnosti a $\mathbf{W} > 0$. Potom odhad $\hat{\boldsymbol{\beta}}_{\mathbf{W}}$ vektoru parametrů $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců je ve tvaru*

$$\hat{\boldsymbol{\beta}}_{\mathbf{W}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}.$$

[2]Model s homogenním rozptylem se nazývá homoskedastický.

*Důkaz.* Důkaz této věty je uveden například v [6]. □

Právě matice $\mathbf{W}$ nám svými hodnotami určuje heterogenitu rozptylu, neboť parametr $\sigma^2$ (ač neznámý) je pro všechny složky stejný. $\mathbf{W}$ může být například diagonální matice tvaru

$$\mathbf{W} = \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n_n \end{pmatrix},$$

kde $n_1, \dots n_n$ jsou obecně různé hodnoty. Protože $\mathbf{W}$ je varianční matice, můžeme z ní vyčíst, že jednotlivá pozorování jsou v tomto případě navzájem nekorelovaná ($C(\varepsilon_i, \varepsilon_j) = 0$ pro každé $i \neq j$).

*Poznámka.* Rozptyl $i$-tého pozorování $DY_i$ má potom tvar

$$DY_i = \frac{\sigma^2}{n_i}. \tag{1.12}$$

**Definice 1.9.** Nechť $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1})$ je regresní model plné hodnosti a nechť matice $\mathbf{W} > 0$ je diagonální. Pak model

$$\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1})$$

nazýváme **vážený regresní model** a metodu nejmenších čtverců, pomocí které byly provedeny odhady vektoru parametrů $\boldsymbol{\beta}$, nazýváme **vážená metoda nejmenších čtverců**.

# Kapitola 2

# Metoda maximální věrohodnosti, EM algoritmus

Existuje široké spektrum metod využívaných k získání odhadů parametrů. V kapitole 1 jsme si představili odhad metodou nejmenších čtverců jako nejběžnější způsob pro získávání parametrů v oblasti lineární regrese.

V této kapitole si uvedeme další metodu a to *metodu maximální věrohodnosti*. V druhé části si představíme tzv. *EM algoritmus* - iterační algoritmus, který pracuje na principu metody maximální věrohodnosti, ale je schopný řešit i statistické problémy, které by byly jinak početně nebo časově velmi náročné. Tento algoritmus se používá zejména v případě, kdy je struktura dat neúplná nebo příliš složitá. Další hojně využívanou aplikací je výpočet parametrů smíšeného datového souboru.

## 2.1 Metoda maximální věrohodnosti

Nechť $\mathbf{X} = (X_1, \dots X_n)$ je náhodný výběr a $\boldsymbol{\Theta}$ je množina parametrů. Předpokládejme, že náhodný výběr $\mathbf{X}$ je z rozdělení spojitého typu s hustotou

$$f(\mathbf{x}|\boldsymbol{\theta}).$$

Pro sdruženou hustotu $f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})$ platí vztah

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(x_i|\theta) = L(\boldsymbol{\theta}|x_1, \dots, x_n) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) \\ &= L(\boldsymbol{\theta}|\mathbf{x}), \end{aligned} \tag{2.1}$$

kde vektor $\boldsymbol{\theta} = (\theta_1 \dots \theta_k) \in \boldsymbol{\Theta}$ je vektor neznámých parametrů. Funkci $L(\boldsymbol{\theta}|\mathbf{x})$ nazýváme *věrohodnostní funkcí*.

**Definice 2.1.** Odhad $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$ nazveme **maximálně věrohodným odhadem** vektoru parametrů $\boldsymbol{\theta}$, jestliže pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí

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

Maximálně věrohodný odhad $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$ tedy hledáme přes maximalizaci věrohodnostní funkce $L(\boldsymbol{\theta}|\mathbf{x})$. Pokud jsou prvky výběru nezávislé (a to víme, že jsou, z definice náhodného výběru), můžeme funkci $L(\boldsymbol{\theta}|\mathbf{x})$ převést na logaritmickou. Ta zachovává polohu extrémů původní funkce, výpočty s ní jsou ale většinou jednodušší. Máme tady věrohodnostní logaritmickou funkci tvaru

$$\ln L(\boldsymbol{\theta}|\mathbf{x}) = l(\boldsymbol{\theta}|\mathbf{x}) = l(\theta_1, \ldots, \theta_k|\mathbf{x}).$$

Derivací funkce $l(\boldsymbol{\theta}|\mathbf{x}) = l(\theta_1, \ldots, \theta_k|\mathbf{x})$ dostáváme k určení maximálně věrohodných odhadů soustavu rovnic

$$\begin{aligned} \frac{\partial}{\partial\theta_1} l(\boldsymbol{\theta}|\mathbf{x}) &= \frac{\partial}{\partial\theta_1} \sum_{i=1}^{n} \ln f(x_i|\boldsymbol{\theta}) = 0 \\ &\vdots \\ \frac{\partial}{\partial\theta_k} l(\boldsymbol{\theta}|\mathbf{x}) &= \frac{\partial}{\partial\theta_k} \sum_{i=1}^{n} \ln f(x_i|\boldsymbol{\theta}) = 0. \end{aligned} \tag{2.2}$$

Jejich vyřešením získáme maximálně věrohodné odhady $\hat{\theta}_{MLE_1}, \ldots, \hat{\theta}_{MLE_k}$ jednotlivých parametrů $\theta_1, \ldots, \theta_k$.

*Poznámka.* Fakt, že jsou odhady $\hat{\theta}_{MLE_1}, \ldots, \hat{\theta}_{MLE_k}$ skutečně maximální, můžeme ověřujeme pomocí druhých derivací rovnic 2.2.

## 2.2 EM algoritmus

Základní myšlenkou EM algoritmu (Expectation-Maximization algorithm) je převést vhodným způsobem problém neúplných dat na problém dat úplných, který nabízí jednodušší a efektivnější řešení. Pro oba případy nalezneme věrohodnostní funkce, zjistíme, jaký je mezi nimi vztah, a využijeme jednodušší postup pro další výpočet.

*Poznámka.* V praxi se můžeme setkat také s příkladem úplných dat, k jehož řešení je vhodné použít EM algoritmus, a to proto, že přidáním dodatečných proměnných k proměnným stávajícím, zjednodušíme nebo urychlíme výpočet odhadu parametrů věrohodnostní funkce.

EM algoritmus se skládá ze dvou kroků:

- *krok E (Expectation)*, ve kterém dosazením parametrů za chybějící data vytvoříme odhad nové věrohodnostní funkce pomocí podmíněné střední hodnoty,
- *krok M (Maximization)*, ve kterém získanou věrohodnostní funkci (a s ní i odhad vektoru parametrů $\boldsymbol{\theta}$) maximalizujeme metodou maximální věrohodnosti.

Získaný odhad vektoru $\boldsymbol{\theta}$ poté využijeme v další iteraci k opětovnému odhadnutí parametrů a jejich maximalizaci. Výše uvedený postup opakujeme až do chvíle, kdy je přírůstek věrohodnostní funkce tak malý, že se zastavení algoritmu na výsledku výrazně neprojeví.

*Poznámka.* Vlastnosti EM algoritmu zaručující konvergenci posloupnosti hodnot logaritmické věrohodnostní funkce pro jednotlivé odhady parametrů jsou uvedeny v literatuře [11].

### 2.2.1 EM algoritmus

Předpokládejme, že máme neúplný náhodný výběr $\mathbf{X} = (X_1, \ldots, X_k)$ z rozdělení určeného hustotou $f(\mathbf{x}|\boldsymbol{\theta})$. Podle vztahu 2.1 určuje sdružená hustota $f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})$ věrohodnostní funkci $L(\boldsymbol{\theta}|\mathbf{x})$. V tomto případě nazýváme funkci $L(\boldsymbol{\theta}|\mathbf{x})$ *věrohodnostní funkcí neúplných dat*. Podobně jako v odstavci 2.1 si funkci $L(\boldsymbol{\theta}|\mathbf{x})$ převedeme na logaritmickou věrohodnostní funkci

$$l(\boldsymbol{\theta}|\mathbf{x}) = \ln L(\boldsymbol{\theta}|\mathbf{x}). \tag{2.3}$$

Mějme dále úplný náhodný výběr $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, který vznikl rozšířením náhodného výběru $\mathbf{X}$ novými, ovšem neznámými proměnnými $\mathbf{Y} = (Y_1, \ldots, Y_r)$. Hustota úplného náhodného výběru má tvar

$$g(\mathbf{z}|\boldsymbol{\theta}) = g(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = g_{\mathbf{y}|\mathbf{x},\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) g_{\mathbf{x}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}).$$

Pro sdruženou hustotu $g_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$ platí

$$g_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta}) = g_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = L(\boldsymbol{\theta}|\mathbf{z}).$$

Funkci $L(\boldsymbol{\theta}|\mathbf{z})$, kterou nazýváme *věrohodnostní funkcí úplných dat*, si také převedeme na logaritmickou

$$l(\boldsymbol{\theta}|\mathbf{z}) = \ln L(\boldsymbol{\theta}|\mathbf{z}). \tag{2.4}$$

*Poznámka.* Věrohodnostní funkce $L(\boldsymbol{\theta}|\mathbf{z}) = L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ je vlastně funkcí závislou pouze na neznámých proměnných $\mathbf{y} = (y_1, \ldots, y_r)$, protože hodnoty vektoru $\mathbf{x} = (x_1, \ldots, x_k)$ jsou realizací náhodného výběru $\mathbf{X} = (X_1, \ldots, X_k)$ a hodnoty parametrů $\boldsymbol{\theta}$ jsou konstanty.

**Krok E:** EM algoritmus prvně nalezne hodnotu logaritmické věrohodnostní funkce $l(\boldsymbol{\theta}|\mathbf{z})$ s ohledem na neznámé hodnoty vektoru $\mathbf{Y}$ a odhady parametrů $\boldsymbol{\theta}$ pomocí podmíněné střední hodnoty:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) = E\left[l(\boldsymbol{\theta}|\mathbf{z})|\mathbf{x}, \boldsymbol{\theta}^{(i-1)}\right], \tag{2.5}$$

kde $\boldsymbol{\theta}^{(i-1)}$ je vektor parametrů získaných z předchozí iterace.

**Krok M:** V tomto kroku se EM algoritmus snaží maximalizovat funkci

$$(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}),$$

získanou v kroku E, tedy hledá takovou hodnotu $Q^{(i)}$, aby platilo

$$Q^{(i)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}).$$

*Poznámka.* EM algoritmus může v kroku M postupovat i jiným způsobem. Namísto maximalizování funkce $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)})$ se snaží najít takové $Q^{(i)}$, aby platil následující vztah

$$Q(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i-1)}) > Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}).$$

Tato mírně modifikovaná forma EM algoritmu se nazývá *Generalized EM Algorithm (GEM)*.

*Poznámka.* Konkrétní příklad na použití EM algoritmu lze najít v literatuře [11].

### 2.2.2 Aplikace EM algoritmu

Jak už bylo řečeno výše, EM algoritmus se hojně využívá při odhadování parametrů v neúplných datových souborech. Za speciální případ neúplného datového souboru můžeme uvažovat i situaci, kdy máme smíšený model, ale chybí nám informace o rozdělení dat do jednotlivých shluků a základní charakteristiky těchto shluků[1]. EM algoritmus potom používáme na doplnění těchto informací. Přesněji řečeno stanovíme si počet shluků, do kterých chceme data rozdělit a pomocí EM algoritmu počítáme pravděpodobnosti, že $i$-té pozorování $y_i$ $(i = 1, \ldots, n)$ náleží do $j$-tého shluku $(j = 1, \ldots, c)$. Hodnoty těchto pravděpodobností použijeme dále k výpočtu vybraných charakteristik $j$-tého shluku. Pozorování $y_i$ nakonec zařadíme do toho shluku, do kterého náleží s největší pravděpodobností.

Uvažujme smíšený model s pevně stanoveným počtem ($c$) shluků. Pro takový model si zavedeme pojem *smíšené hustoty*.

**Definice 2.2.** Nechť $\mathbf{Y}_j$ je $k$-rozměrný vektor s hustotou rozdělení $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ a $a_j > 0$ pro všechna $j = 1, \ldots, c$. Dále nechť platí $a_1 + a_2 + \cdots + a_c = 1$. Potom hustotu

$$f(\mathbf{y}) = \sum_{j=1}^{c} a_j f_j(\mathbf{y}|\boldsymbol{\theta}_j) \tag{2.6}$$

nazýváme **smíšenou hustotou**. Rozdělení se smíšenou hustotou názýváme **smíšené rozdělení**.

*Poznámka.* V diplomové práci [8] je výraz smíšená hustota nahrazen pojmem *smíšená distribuce*.

Hodnoty $a_j$ určují váhu $j$-té složky smíšené hustoty, funkce $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ je hustota $j$-té složky s vektorem parametrů $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jp_j})$, $c$ je počet složek, které tvoří smíšenou hustotu a $p_j$ je počet parametrů hustoty $j$-tého shluku.

Průběh EM algoritmu si pro lepší názornost představíme nejprve na případu jednorozměrných dat z normálního rozdělení. Hlubší záběr této tématiky je podrobněji popsán v [3]. Pro naše účely je však toto postačující.

[1] V případě normálního rozdělení bychom základními charakteristikami měli na mysli střední hodnotu a rozptyl. Obecně budeme vektor základních charekteristik značit symbolem $\boldsymbol{\theta}$.

Předpokládejme, že $k = 1$ a že tedy máme náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_n)$, kde $Y_i$ je (jednorozměrná) náhodná veličina (s realizací $y_i$) z normálního rozdělení (pro $\forall i = 1, \ldots, n$). Smíšená hustota má potom následující tvar

$$f(y) = \sum_{j=1}^{c} a_j \phi(y|\mu_j, \sigma_j^2), \tag{2.7}$$

kde $\phi(y|\mu_j, \sigma_j^2)$ je hustota $j$-tého shluku, který má normální rozdělení se střední hodnotou $\mu_j$ a rozptylem $\sigma_j^2$. V tomto případě hledáme tedy celkem $3c - 1$ neznámých parametrů $\boldsymbol{\theta} = (a_1, \ldots, a_{c-1}, \mu_1, \ldots, \mu_c, \sigma_1^2, \ldots, \sigma_c^2)$. Jejich hodnoty budeme odhadovat pomocí EM algoritmu.

Základní myšlenka odhadování parametrů pomocí EM algoritmu:

1. stanovíme počet shluků, do kterých budeme třídit jednotlivá pozorování $y_i$,
2. určíme počáteční odhady parametrů těchto shluků (v případě jednorozměrného normálního rozdělení odhadneme pro každý shluk jeho střední hodnotu, rozptyl a váhu),
3. (krok E) pro každý bod $y_i$ určíme pravděpodobnost, že tento bod náleží do $j$-tého shluku,
4. (krok M) spočítáme střední hodnotu, rozptyl a váhu každého shluku,
5. opakujeme kroky E a M dokud není přírůstek parametrů zanedbatelně malý.

S ohledem na předešlý text předpokládejme rozdělení pozorování do $c$ shluků. Počáteční hodnoty parametrů $a_j^{(0)}, \mu_j^{(0)}, \sigma_j^{2(0)}$ lze určit různými způsoby. My si zde představíme způsob uvedený v literatuře [8]. Data setřídíme vzestupně podle naměřených hodnot, takto setříděné je rozdělíme do $c$ složek. Prvotní odhad vah každého shluku bude $\frac{1}{c}$, střední hodnota $j$-tého shluku $\mu_j^{(0)}$ bude odhadnuta výběrovým průměrem

$$\bar{Y}_j = \frac{1}{n_j} \sum_{l=1}^{n_j} Y_{jl}$$

a rozptyl $j$-tého shluku $\sigma_j^{2(0)}$ empirickým rozptylem

$$S_j^2 = \frac{1}{n_j - 1} \sum_{l=1}^{n_j} (Y_{jl} - \bar{Y}_j)^2,$$

kde $n_j$ je počet pozorování v $j$-tém shluku, $j = 1, \ldots, c$; $n_1 + n_2 + \cdots + n_c = n$.

Smíšenou hustotu $f(y)$ uvedenou ve vztahu 2.7 transformujeme na věrohodnostní funkci náhodného výběru $\mathbf{Y} = (Y_1, \ldots, Y_n)$

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|y_1, \ldots, y_n) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{c} a_j \phi(y_i|\mu_j, \sigma_j^2) \right] \tag{2.8}$$

a tu na logaritmickou věrohodnostní funkci

$$l(\boldsymbol{\theta}|\mathbf{y}) = \ln(L(\boldsymbol{\theta}|\mathbf{y})) = \sum_{i=1}^{n} \ln \left[ \sum_{j=1}^{c} a_j \phi(y_i|\mu_j, \sigma_j^2) \right]. \tag{2.9}$$

Vztah 2.9 použijeme k výpočtu (aposteriorní) pravděpodobnosti, že hodnota $y_i$ náleží do $j$-té složky distribuce (aposteriorní pravděpodobnost).

$$\hat{\alpha}_{ij}(y_i) = \frac{\hat{a}_j \phi(y_i|\hat{\mu}_j, \hat{\sigma}_j^2)}{\hat{f}(y_i|\hat{a}_j, \hat{\mu}_j, \hat{\sigma}_j^2)}, \qquad j = 1, \ldots, c; i = 1, \ldots, n, \tag{2.10}$$

kde

$$\hat{f}(y_i|\hat{a}_j, \hat{\mu}_j, \hat{\sigma}_j^2) = \sum_{j=1}^{c} \hat{a}_j \phi(y_i|\hat{\mu}_j, \hat{\sigma}_j^2).$$

Dalším krokem je výpočet odhadů parametrů $\hat{a}_j$, $\hat{\mu}_j$ a $\hat{\sigma_j^2}$. K tomu je zapotřebí maximalizovat funkci 2.9. Najdeme tedy její parciální derivace podle všech parametrů a položíme je rovny nule. Tím získáme soustavu tří věrohodnostních rovnic, z nichž můžeme vyjádřit vztahy pro odhady hledaných parametrů.

$$\hat{a}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_{ij}, \tag{2.11}$$

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\alpha}_{ij} y_i}{\hat{a}_j}, \tag{2.12}$$

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\alpha}_{ij}(y_i - \hat{\mu}_j)^2}{\hat{a}_j}. \tag{2.13}$$

Dosazení do těchto rovnic získáme konkrétní hodnoty pro odhady parametrů $\hat{a}_j$, $\hat{\mu}_j$ a $\hat{\sigma_j^2}$.

*Poznámka.* Z rovnic 2.11, 2.12 a 2.13 vidíme, že aposteriorní pravděpodobnost se využívá k odhadu všech tří parametrů $\hat{a}_j$, $\hat{\mu}_j$ a $\hat{\sigma_j^2}$.

V případě vícerozměrných dat postupujeme analogicky jako v případě dat jednorozměrných. Mějme náhodný výběr $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, kde $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ jsou $k$-rozměrné vektory. Smíšená distribuce má potom tvar

$$f(\mathbf{y}) = \sum_{j=1}^{c} a_j \phi(\mathbf{y}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j^2). \tag{2.14}$$

Střední hodnoty $\mu_j$ nahradíme vektorem středních hodnot $\boldsymbol{\mu}_j$ a rozptyl $\sigma_j^2$ varianční maticí $\boldsymbol{\Sigma}_j$ pro každé $j = 1, \ldots, c$. Hledáme tedy odhad vektoru parametrů

$$\boldsymbol{\theta} = (a_1, \ldots, a_{c-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_c).$$

Věrohodnostní funkce má tvar

$$L(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_n) = \prod_{i=1}^{n}\left[\sum_{j=1}^{c} a_j \phi(\mathbf{y}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j^2)\right] \tag{2.15}$$

a logaritmická věrohodnostní funkce vypadá následovně

$$l(\boldsymbol{\theta}|\mathbf{y}_1,\ldots\mathbf{y}_n) = \ln(L(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_n)) = \sum_{i=1}^{n}\ln\left[\sum_{j=1}^{c} a_j \phi(\mathbf{y}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j^2)\right]. \tag{2.16}$$

Vztah pro výpočet aposteriorní pravděpodobnosti bude ve tvaru:

$$\hat{\alpha}_{ij}(\mathbf{y}_i) = \frac{\hat{a}_j \phi(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_j,\hat{\boldsymbol{\Sigma}}_j^2)}{\hat{f}(\mathbf{y}_i|\hat{a}_j,\hat{\boldsymbol{\mu}}_j,\hat{\boldsymbol{\Sigma}}_j^2)}, \qquad j=1,\ldots,c; i=1,\ldots,n. \tag{2.17}$$

Logaritmickou věrohodnostní funkci zderivujeme podle všech tří parametrů ($a_j$, $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$), získáme parciální derivace, které položíme rovny nule, a z vytvořených rovnic vyjádříme vztahy pro jednotlivé parametry

$$\hat{a}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}_{ij}, \tag{2.18}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\alpha}_{ij}\mathbf{y}_i}{\hat{a}_j}, \tag{2.19}$$

$$\hat{\boldsymbol{\Sigma}}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\alpha}_{ij}(\mathbf{y}_i-\hat{\boldsymbol{\mu}}_j)(\mathbf{y}_i-\hat{\boldsymbol{\mu}}_j)'}{\hat{a}_j}. \tag{2.20}$$

Dosazením do rovnic 2.18, 2.19 a 2.20 získáme odhady $j$-tých parametrů $\hat{a}_j, \hat{\boldsymbol{\mu}}_j$ a $\hat{\boldsymbol{\Sigma}}_j$.

*Poznámka.* Praktický příklad využití EM algoritmu při rozdělování dat do shluků a podrobnější teorie o smíšené hustotě a smíšeném rozdělení je popsána v diplomové práci [8].

# Kapitola 3

# Praktická část - Zpracování dat

Hlavní náplní této bakalářské práce je vylepšení lineárního regresního modelu popisujícího závislost hodnoty pKa na velikostech nábojů na vybraných atomech v molekule. K dispozici máme 18 datových souborů (každý soubor pro jednu chemickou metodu, kterou byly vypočítány náboje na daných atomech), z nichž každý obsahuje informace o 74 navzájem různých molekulách fenolů. Pro každou molekulu byla experimentálně naměřena hodnota $pKa$ a vypočítány hodnoty $qH$, $qO$, $qC1$, $qOD$, $qC1D$, kde

- $pKa$ ... disociační konstanta,
- $qH$ ... hodnota náboje na atomu vodíku,
- $qO$ ... hodnota náboje na atomu kyslíku,
- $qC1$ ... hodnota náboje na atomu uhlíku,
- $qOD$... hodnota náboje na atomu kyslíku v rámci disociované molekuly,
- $qC1D$ ... hodnota náboje na atomu uhlíku v rámci disociované molekuly.

Dále jsou zadané dva typy lineárních regresních modelů, a sice

- 3d.model ... $pKa \sim qH + qO + qC1$,
- 5d.model ... $pKa \sim qH + qO + qC1 + qOD + qC1D$

a pro ně dopočítaná kritéria kvality modelu

- $R^2$,
- $RMSE$,
- $\Delta$,
- $\sigma^2$.

Úkolem bylo seznámit se s oběma modely, nalézt vlivné body, prozkoumat je, odstranit outliery, případně vybočující pozorování a spočítat parametry nových modelů bez těchto bodů. Dále vypočítat hodnoty všech zmíněných kritérií a porovnat je s hodnotami původními. Stěžejní myšlenkou těchto kroků byla snaha o zlepšení původních dvou modelů s předpokladem, že odstraněním outlierů a vybočujících pozorování se modely zlepší.

K analýze dat byl použit statistický software R. Následující text je pojat jako průvodce uskutečněnou analýzou, proto bude obsahovat i pasáže s R-kovými příkazy a komentáři k nim. Aktuální verze systému R je ke stažení na domovských stránkách **http://cran.r-project.org/**. Doporučuji stáhnout si z uvedené webové adresy i uživatelsky přístupnější prostředí tzv. R-Studio, které usnadňuje a zpřehledňuje práci se systémem R.

## 3.1 3d.modely

### 3.1.1 Seznámení s 3d.modelem

Jelikož by manipulace s osmnácti soubory najednou byla velmi nepřehledná, budeme si kroky analýzy demonstrovat pouze na datech ze souboru Svob2007_chal2. Všechny důležité výstupy (zejména grafy) pro ostatní soubory jsou uvedeny na přiloženém cd.

Datový soubor je nejprve nutné načíst do systému R. To provedeme pomocí příkazu `read.delim`. Tento příkaz má dva povinné parametry a sice `sep` (určuje symbol, kterým jsou od sebe odděleny jednotlivé hodnoty) a `dec` (podává informaci o symbolu, kterým jsou oddělena desetinná místa).

```
dat<-read.delim('Svob2007_chal2.csv',sep=';',dec='.')
```

Načtený soubor jsme uložili příkazem <− do proměnné dat jako datový soubor typu data.frame (datová tabulka). Příkaz `names(dat)` vypíše názvy jednotlivých sloupců data.frame dat.

```
names(dat)
[1]  "N"                "pKa"              "qH"
[4]  "qO"               "qC1"              "qOD"
[7]  "qC1D"             "outlier"          "pKa.3d.model."
[10] "pKa.3d.WO.model." "pKa.5d.model."
```

V případě, že bychom chtěli z datového souboru vypsat pouze jeden ze sloupců, použijeme k tomu příkaz tvaru `nazev_datoveho_souboru$jmeno_sloupce`. Například pokud bychom chtěli z tabulky dat vypsat pouze sloupec $pKa$, příkaz by vypadal následovně

```
dat$pKa
```

Nyní vytvoříme lineární regresní model $pKa \sim qH + qO + qC1$ pomocí funkce `lm` a vložíme jej do proměnné mod.3d. Zavoláním proměnné mod.3d se zobrazí tvar tohoto modelu a odhad vektoru parametrů $\boldsymbol{\beta} = (\beta_{Int}, \beta_{qH}, \beta_{qO}, \beta_{qC1})$.

```
mod.3d <- lm(pKa~qH+qO+qC1,data=dat)
mod.3d

Call:
lm(formula = pKa ~ qH + qO + qC1, data = dat)

Coefficients:
(Intercept)           qH           qO          qC1
     28.100     -109.918       -2.096        6.926
```

Výsledný model má tedy tvar

$$pKa = -109.918\,qH - 2.096\,qO + 6.926\,qC1 + 28.1\,,$$

kde $qH$ je velikost náboje na atomu vodíku, $qO$ velikost náboje na atomu kyslíku a $qC1$ velikost náboje na atomu uhlíku.

Podrobnější informace o modelu získáme zavoláním funkce `summary`.

```
summary(mod.3d)

Call:
lm(formula = pKa ~ qH + qO + qC1, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.71693 -0.28321  0.03858  0.28430  1.49850

Coefficients:
            Estimate Std. Error t value Pr(|>|)
(Intercept)   28.100     21.235   1.323   0.1900
qH          -109.918     50.008  -2.198   0.0313 *
qO            -2.096     38.744  -0.054   0.9570
qC1            6.926     10.376   0.668   0.5066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8139 on 70 degrees of freedom
Multiple R-squared: 0.8737,Adjusted R-squared: 0.8683
F-statistic: 161.4 on 3 and 70 DF,  p-value: < 2.2e-16
```

Vidíme, že tabulka je rozdělená do několika sektorů.

Sektor *Residuals* obsahuje souhrnné statistiky reziduí: minimální hodnotu, první kvartil, medián, třetí kvartil a maximální hodnotu reziduí. Pokud rezidua pochází z normálního rozdělení, bude se hodnota mediánu pohybovat okolo nuly a hodnoty *Min*, *Max*, *1Q* a *3Q* budou kolem mediánu rozloženy zhruba symetricky. Signifikantní je symetrie prvního a třetího kvartilu, protože zajistit symetrii minima a maxima je, vzhledem k jejich vyšší variabilitě, problematičtější. Z tabulky je tedy vidět, že rozdělení reziduí modelu mod.3d je normální.

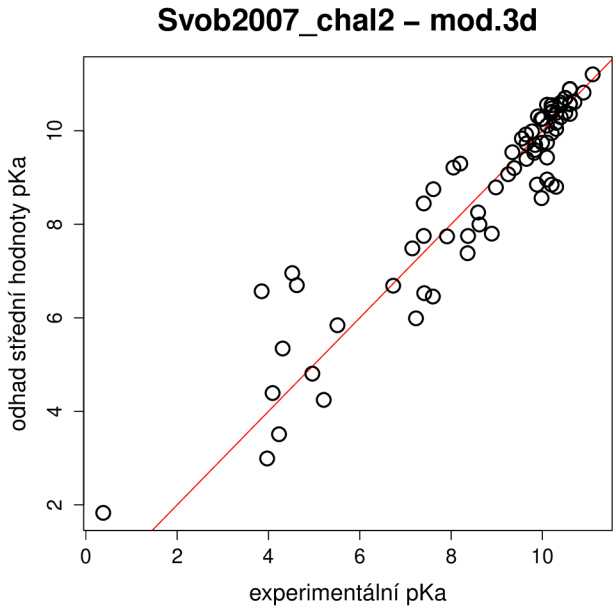
Sektor *Coefficients* obsahuje odhady parametrů (sloupec *Estimate*), jejich střední chyby (*Std. Error*) a nakonec t-hodnoty (*t value*) a p-hodnoty $Pr(|>|)$ testu nulové hypotézy o tom, že skutečná hodnota daného koeficientu je nulová. Hvědičky za sloupcem $Pr(|>|)$ značí hladiny významnosti těchto testů. Např. * značí,

že nulovou hypotézu $H_0$ zamítáme na hladině významnosti $0.05$, ** značí zamítnutí hypotézy na hladině významnosti 0.01 apod.

Poslední sektor obsahuje souhrn vybraných statistik modelu mod.3d. Pro nás budou nejdůležitější *Residual standard error* (odhad $\sigma^2$), *Multiple R-squared* (koeficient determinace $R^2$), případně *Adjusted R-squared* (upravený koeficient determinace $\bar{R}^2$).

Vektor odhadů středních hodnot (ph_mod.3d) $pKa$ získáme z modelu mod.3d pomocí funkce `predict`. Navíc si do vektoru pKa uložíme sloupec hodnot $pKa$ z datové tabulky dat a vytvoříme graf závislosti odhadu střední hodnoty $pKa$ na experimentálně naměřených hodnotách $pKa$. Grafy se vytvářejí příkazem `plot`. Vykreslená data navíc proložíme přímkou $x = y$, abychom lépe viděli, jak dobře se body modelem aproximovaly. K tomu nám poslouží funkce `abline`.

```
ph_mod.3d<-predict(mod.3d)
pKa<-dat$pKa
plot(pKa, ph_mod.3d, ylab=('odhad střední hodnoty pKa'),
     xlab=('experimentální pKa'),
     main='Svob2007_chal2 - 3d',
     abline(a=0,b=1,col='red'))
```




Obrázek 3.1: Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$

*Poznámka.* Grafy pro všech osmnáct souborů jsou uvedeny na přiloženém cd ve složce ***1-3d.model.predict.exp.pKa***.

### 3.1.2 Analýza reziduí

Nyní se zaměříme na vlivná pozorování. K výpočtu různých typů reziduí použijeme následující příkazy.

- `resid(mod.3d)` pro klasická rezidua $r_i$,
- `rstandard(mod.3d)` pro standardizovaná rezidua $r_i^*$,
- `rstudent(mod.3d)` pro studentizovaná rezidua $r_{(i)}^*$,
- `dffits(mod.3d)` pro DFIT rezidua $d_i$.

Cookovu vzdálenost zjistíme pomocí funkce `cooks.distance(mod.3d)`, hodnotu vlivu $i$-tého pozorování příkazem `hatvalues(mod.3d)`. Hodnoty jednotlivých Cookových vzdáleností jsou uvedeny v tabulce 3.1, hodnoty klasických, standardizovaných a studentizovaných reziduí potom v tabulce 3.2.

| Tabulka Cookových vzdáleností pro soubor Svob2007_chal2 | | | | | |
|---|---|---|---|---|---|
| Název | Cooks.dist | Název | Cooks.dist | Název | Cooks.dist |
| NSC_10112 | 0.00169 | NSC_2440 | 0.00039 | NSC_5296 | 0.00008 |
| NSC_11215 | 0.00105 | NSC_249188 | 0.00042 | NSC_5353 | 0.00033 |
| NSC_12987 | 0.00185 | NSC_2599 | 0.00027 | NSC_5387 | 0.02070 |
| NSC_1317 | 0.00391 | NSC_2880 | 1.09309 | NSC_60291 | 0.00150 |
| NSC_14881 | 0.05896 | NSC_3115 | 0.00920 | NSC_60735 | 0.00094 |
| NSC_1532 | 0.00363 | NSC_3142 | 0.01574 | NSC_62011 | 0.00017 |
| NSC_15351 | 0.00145 | NSC_3177 | 0.01155 | NSC_62012 | 0.00062 |
| NSC_1548 | 0.00585 | NSC_33870 | 0.03576 | NSC_6215 | 0.06322 |
| NSC_1549 | 0.00014 | NSC_3504 | 0.00062 | NSC_65646 | 0.00056 |
| NSC_1551 | 0.03590 | NSC_36808 | 0.00061 | NSC_65647 | 0.00007 |
| NSC_1552 | 0.02574 | NSC_36947 | 0.19358 | NSC_6769 | 0.02153 |
| NSC_156083 | 0.00112 | NSC_3696 | 0.00059 | NSC_70955 | 0.00101 |
| NSC_17588 | 0.00099 | NSC_3698 | 0.01561 | NSC_7739 | 0.12747 |
| NSC_1809 | 0.02566 | NSC_3814 | 0.00036 | NSC_82996 | 0.01106 |
| NSC_1858 | 0.00073 | NSC_3815 | 0.03095 | NSC_85475 | 0.00118 |
| NSC_1888 | 0.00043 | NSC_3829 | 0.00109 | NSC_8768 | 0.00000 |
| NSC_202753 | 0.11313 | NSC_38776 | 0.00002 | NSC_881 | 0.00253 |
| NSC_2082 | 0.06972 | NSC_3991 | 0.00045 | NSC_8873 | 0.00207 |
| NSC_2123 | 0.00002 | NSC_41205 | 0.00916 | NSC_8885 | 0.00364 |
| NSC_2127 | 0.01921 | NSC_4875 | 0.00009 | NSC_8895 | 0.03702 |
| NSC_2150 | 0.00120 | NSC_4960 | 0.00399 | NSC_90441 | 0.04060 |
| NSC_21735 | 0.00093 | NSC_4965 | 0.00045 | NSC_9268 | 0.00065 |
| NSC_2209 | 0.00110 | NSC_4969 | 0.04401 | NSC_95810 | 0.00003 |
| NSC_227926 | 0.00079 | NSC_5103 | 0.00058 | NSC_9885 | 0.00100 |
| NSC_23076 | 0.00027 | NSC_5105 | 0.00051 | | |

Tabulka 3.1: Tabulka Cookových vzdáleností

Z hodnot v tabulce vidíme, že Cookova vzdálenost neodhalila žádná vlivná pozorování.

*Poznámka.* Cookova vzdálenost pro bod NSC_2880 $C_{NSC_2880} = 1.09309 > 0$, ale vzhledem k tomu, že tato hodnota přesahuje kritickou hodnotu 1 jen o malinko, nepovažujeme zatím tento bod za významně vlivný. Uvidíme, zda významnou vlivnost prokáže další analýza.

| Tabulka reziduí pro soubor Svob2007_chal | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Název | $r_i$ | $r_i^*$ | $r_{(i)}^*$ | | Název | $r_i$ | $r_i^*$ | $r_{(i)}^*$ |
| NSC_10112 | -0.348 | -0.435 | -0.433 | * | NSC_3698 | -1.162 | -1.449 | -1.461 |
| NSC_11215 | -0.288 | -0.359 | -0.357 | * | NSC_3814 | 0.145 | 0.182 | 0.180 |
| NSC_12987 | 0.337 | 0.423 | 0.420 | * | NSC_3815 | 1.422 | 1.781 | 1.809 |
| NSC_1317 | -0.335 | -0.428 | -0.426 | * | NSC_3829 | 0.240 | 0.302 | 0.300 |
| NSC_14881 | -1.045 | -1.363 | -1.372 | * | NSC_38776 | 0.033 | 0.042 | 0.041 |
| NSC_1532 | -0.301 | -0.387 | -0.384 | * | NSC_3991 | 0.178 | 0.223 | 0.221 |
| NSC_15351 | -0.350 | -0.437 | -0.435 | * | NSC_41205 | 0.625 | 0.790 | 0.788 |
| NSC_1548 | -0.279 | -0.371 | -0.369 | * | NSC_4875 | -0.079 | -0.098 | -0.098 |
| NSC_1549 | 0.111 | 0.138 | 0.137 | * | NSC_4960 | 0.673 | 0.836 | 0.834 |
| NSC_1551 | 0.979 | 1.256 | 1.261 | * | NSC_4965 | -0.201 | -0.251 | -0.249 |
| NSC_1552 | 1.241 | 1.557 | 1.574 | * | NSC_4969 | 1.499 | 1.886 | 1.922 |
| NSC_156083 | 0.154 | 0.200 | 0.199 | * | NSC_5103 | -0.203 | -0.254 | -0.252 |
| NSC_17588 | -0.281 | -0.351 | -0.349 | * | NSC_5105 | -0.107 | -0.138 | -0.137 |
| NSC_1809 | 1.142 | 1.438 | 1.449 | * | NSC_5296 | 0.088 | 0.110 | 0.109 |
| NSC_1858 | -0.284 | -0.353 | -0.351 | * | NSC_5353 | 0.083 | 0.108 | 0.107 |
| NSC_1888 | -0.207 | -0.257 | -0.256 | * | NSC_5387 | 1.145 | 1.435 | 1.446 |
| NSC_202753 | -2.076 | -2.633 | -2.754 | * | NSC_60291 | -0.294 | -0.369 | -0.367 |
| NSC_2082 | -1.033 | -1.361 | -1.370 | * | NSC_60735 | 0.241 | 0.303 | 0.301 |
| NSC_2123 | 0.021 | 0.027 | 0.027 | * | NSC_62011 | 0.116 | 0.145 | 0.144 |
| NSC_2127 | -1.140 | -1.427 | -1.438 | * | NSC_62012 | -0.255 | -0.317 | -0.315 |
| NSC_2150 | 0.170 | 0.219 | 0.217 | * | NSC_6215 | 0.979 | 1.291 | 1.297 |
| NSC_21735 | 0.252 | 0.315 | 0.313 | * | NSC_65646 | -0.198 | -0.247 | -0.246 |
| NSC_2209 | -0.278 | -0.348 | -0.345 | * | NSC_65647 | -0.084 | -0.104 | -0.103 |
| NSC_227926 | 0.292 | 0.363 | 0.361 | * | NSC_6769 | 1.032 | 1.300 | 1.306 |
| NSC_23076 | 0.131 | 0.164 | 0.163 | * | NSC_70955 | -0.201 | -0.255 | -0.253 |
| NSC_2440 | 0.180 | 0.224 | 0.223 | * | NSC_7739 | -2.436 | -3.073 | -3.280 |
| NSC_249188 | -0.217 | -0.270 | -0.268 | * | NSC_82996 | 1.091 | 1.356 | 1.365 |
| NSC_2599 | -0.150 | -0.187 | -0.186 | * | NSC_85475 | 0.250 | 0.315 | 0.313 |
| NSC_2880 | -2.717 | -3.808 | -4.247 | * | NSC_8768 | -0.009 | -0.012 | -0.012 |
| NSC_3115 | 0.621 | 0.785 | 0.783 | * | NSC_881 | -0.331 | -0.418 | -0.416 |
| NSC_3142 | 0.883 | 1.112 | 1.114 | * | NSC_8873 | -0.407 | -0.508 | -0.506 |
| NSC_3177 | -1.097 | -1.365 | -1.373 | * | NSC_8885 | -0.457 | -0.573 | -0.571 |
| NSC_33870 | 0.720 | 0.953 | 0.952 | * | NSC_8895 | 1.353 | 1.704 | 1.728 |
| NSC_3504 | 0.191 | 0.240 | 0.239 | * | NSC_90441 | 0.966 | 1.248 | 1.253 |
| NSC_36808 | 0.247 | 0.307 | 0.305 | * | NSC_9268 | -0.190 | -0.239 | -0.238 |
| NSC_36947 | -1.448 | -1.952 | -1.993 | * | NSC_95810 | 0.044 | 0.055 | 0.055 |
| NSC_3696 | 0.260 | 0.323 | 0.321 | * | NSC_9885 | 0.352 | 0.437 | 0.435 |

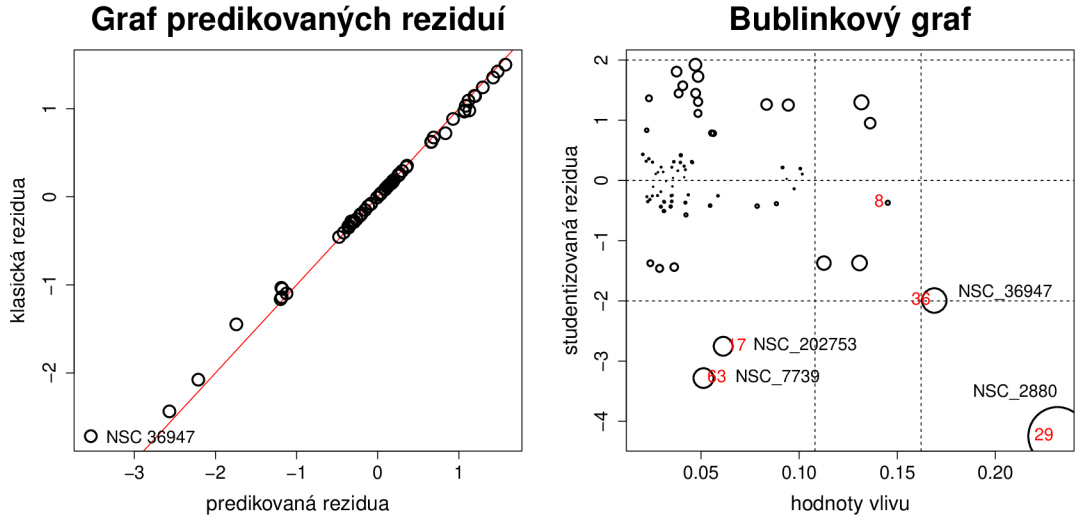
Tabulka 3.2: Hodnoty různých typů reziduí

Vykreslíme si ještě graf predikovaných reziduí a bublinkový graf. Predikovaná rezidua dopočítáme podle vzorce $r_{(i)} = \dfrac{r_i}{1 - h_{ii}}$.

```
res_classic<-resid(mod.3d)
hatval<-hatvalues(mod.3d)
res_predict<-res_classic/(1-hatval)
plot(res_predict, res_classic, xlab='klasická rezidua',
     ylab='predikovaná rezidua',
     main='Graf predikovaných reziduí',
     abline(a=0,b=1,col='red'),cex=1.5)
```

K tomu, abychom mohli vykreslit bublinkový graf, potřebujeme načíst knihovnu **car**. Knihovny se načítají příkazem `library`.

```
library(car)
influencePlot(mod.3d, main='Bublinkový graf',
              xlab='hodnoty vlivu',
              ylab='studentizovaná rezidua',
              id.n=3, id.col='red')
text(x=0.220, y=-3.50, 'NSC_2880')
text(x=0.079, y=-3.26, 'NSC_7739')
text(x=0.093, y=-2.73, 'NSC_202753')
text(x=0.195, y=-1.85, 'NSC_36947')
```




Obrázek 3.2: a) Graf predikovaných reziduí, b) Bublinkový graf

Graf predikovaných reziduí stanovil bod NSC_36947 (č.36) jako vybočující, naopak bublinkový graf označil za odlehlý a silně vlivný bod NSC_2880 (č.29) a body NSC_7739 (č.63) a NSC_202753 (č.17) jako odlehlé. Podíváme se na podrobnější informace o těchto bodech, které nám jako další výstup vypsala funkce `influencePlot`.

```
     StudRes         Hat      CookD
8  -0.3687448 0.14523731 0.07647319
17 -2.7539313 0.06127647 0.33634170
29 -4.2465336 0.23163305 1.04550813
36 -1.9931969 0.16887028 0.43997785
63 -3.2799010 0.05123526 0.35702491
```

Cookovy vzdálenosti vypočítané algoritmem funkce `influencePlot` se trošku liší od vzdáleností vypočítaných funkcí `cooks.distance`. U bodu č. 29 sice cookova vzdálenost přesáhla hodnotu 1, ale ne moc výrazně. Vliv 29 tého pozorování je znatelně vyšší než vliv ostatních pozorování, ale celkově jeho hodnota není tak vysoká, abychom bod museli ze souboru odstranit[1]. Ostatní body mají hodnoty nižší než bod 29 a navíc jejich vliv na okolní pozorování není významný. U zbylých sedmnácti souborů vykazovala analýza reziduí obdobné výsledky. Grafy predikovaných reziduí jsou dostupné na cd ve složce ***2-3d.grafy_predikovanych_rezidui***.

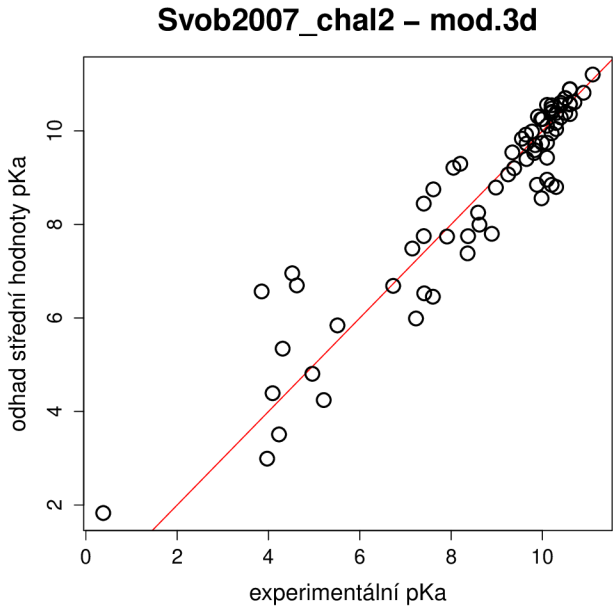
[1]Přes všechny tyto indicie jsme cvičně bod č.29 ze souboru odstranili a přepočítali kritéria kvality modelu. Ke zlepšení predikce však, jak jsme předpokládali, nedošlo, proto jsme bod do datového souboru zase vrátili.

**Závěr analýzy reziduí**

Analýza reziduí tedy prokázala, že žádné pozorování není tak vlivné, odlehlé ani vybočující, aby bylo nutné ho ze souboru odstranit. Tím se náš model stává univerzální pro všechny naměřené molekuly.

### 3.1.3 Shluková analýza

Nicméně zůstává otázka, jak zadané modely $pKa \sim qH + qO + qC1 + qOD + qC1D$ a $pKa \sim qH + qO + qC1$ modifikovat a zvýšit hodnotu statistiky $R^2$. Podívejme se proto ještě jednou na graf závislosti odhadů středních hodnot $pKa$ na experimentálních hodnotách $pKa$.




Obrázek 3.3: Graf závislosti odhadu střední hodnoty $pK$a na experimentální $pKa$

Z grafu je vidět, že datový soubor nemá homogenní rozptyl a že velikost rozptylu klesá s rostoucí hodnotou $pKa$. Zkusme tedy datový soubor rozdělit na několik shluků tak, aby rozptyl v každém z nich byl homogenní. K tomu využijeme funkci `Mclust` z knihovny **mclust**. Podle grafu 3.3 předpokládáme, že by bylo vhodné rozdělit soubor molekul na tři shluky, případně na dva. Pro kontrolu ovšem vyzkoušíme i rozdělení dat na shluky čtyři. Shlukovou analýzu můžeme provést podle kterékoli proměnné ($pKa$, $qH$, $qO$, $qC1$, $qOD$, $qC1D$) nebo podle jejich kombinací. V této práci uskutečníme pouze shlukování podle jednotlivých proměnných. Při hledání optimálního shlukování nás budou nejvíce zajímat tři aspekty

- jednoznačné rozdělení molekul do shluků (ideálně tak, aby se intervaly jednotlivých shluků nepřekrývaly),

- co nejvyšší hodnota koeficientu determinace $R^2$,
- univerzální kritérium dělení pro všech osmnáct chemických metod (například, aby všechny datové soubory byly rozdělitelné podle náboje $qH$).

Funkce `Mclust` má dva povinné parametry. Prvním je výběr dat, která chceme rozdělit do shluků, a druhým je počet shluků, do kterých se mají data rozdělit. Pokud počet shluků neuvedeme, rozdělí `Mclust` výběr dat na tolik skupin, aby hodnota kritéria BIC[2] byla co nejmenší. Kroky shlukové analýzy si demonstrujeme na případu rozdělení dat do tří skupin podle hodnot nábojů $qH$.

```
qH<-dat$qH
library(mclust)
qH.Mclust<-Mclust(qH,G=3)
qH.Mclust

'Mclust' model object:
 best model: univariate, unequal variance (V) with 3 components
```

Výpis funkce `Mclust` je velice stručný. Příkazem `ls(qH.Mclust)` se můžeme podívat, jaké výstupy se nám do proměnné qH.Mclust uložily.

```
 [1] "bic"          "BIC"          "call"     "classification" "d"
 [6] "df"           "G"            "loglik"   "modelName"      "n"
[11] "parameters"   "uncertainty"  "z"
```

Vidíme, že do proměnné Mclust se nám vložila datová struktura, která obsahuje veškeré výsledky týkající se odhadnutých parametrů směsi normálních hustot. Tyto výsledky jsou uloženy pod vypsanými položkami. Nejvíce nás budou zajímat následující tři položky:

- *classification* - vektor přiřazující každému pozorování číslo shluku, do kterého bylo zařazeno,
- *G* - proměnná, ve které je uložen počet shluků, do nichž jsme data rozdělili (v tomto případě G=3),
- *parameters* - sekce obsahující další čtyři 'podsekce' - *mean*, *pro*, *variance* a *Vinv*. Jejich význam si vysvětlíme později.

Nyní si vykreslíme grafy hodnot $qH$ v prostoru a jednotlivé body obarvíme podle skupiny, do které náleží.

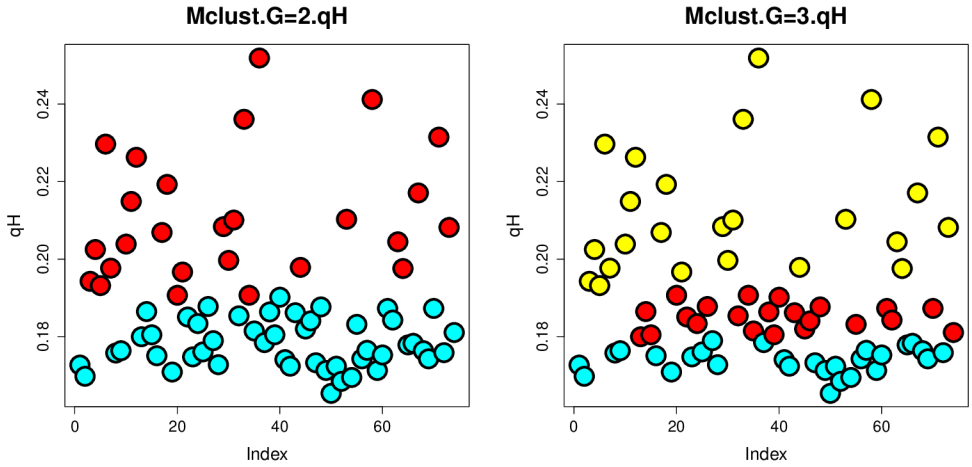
```
qH<-dat$qH
qH.Mclust<-Mclust(qH, G=3)
s3<-qH.Mclust$classification
plot(qH, pch=21, bg=as.numeric(s3*5),
 cex=1.8, main='Mclust.G=3.qH')
```

[2]BIC (Bayesovo informační kritérium) je jedním z kritérií, která posuzují kvalitu regresního modelu. Jeho hodnotu vypočítáme podle vzorce $BIC = -2\ln L + k\ln(n)$, kde $n$ je počet všech pozorování, $L$ je maximální hodnota věrohodnostní funkce, $k$ je počet parametrů v modelu a $n$ je počet všech pozorování). Dalším významným kritériem je například Akaikovo informační kritérium (AIC).

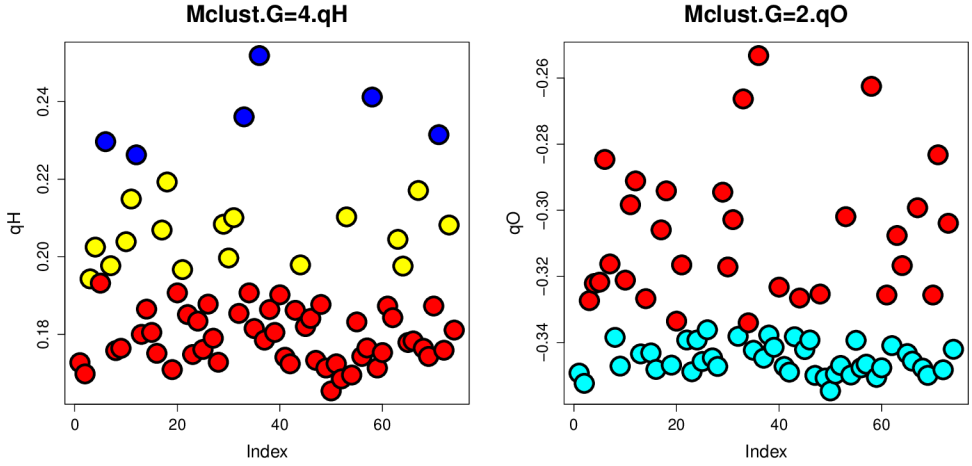
Obdobně vykreslíme graf rozdělení na dva (`G=2`) a čtyři (`G=4`) shluky. Vytvoření všech tří grafů by se dalo zapsat také jedním cyklem.

```
for (k in 2:4){
qH<-dat$qH
qH.Mclust<-Mclust(qH, G=k)
s<-qH.Mclust$classification
plot(qH, pch=21, bg=as.numeric(s*5), cex=1.8,
     main=paste('Mclust.G=', k, '.qH'))}
```

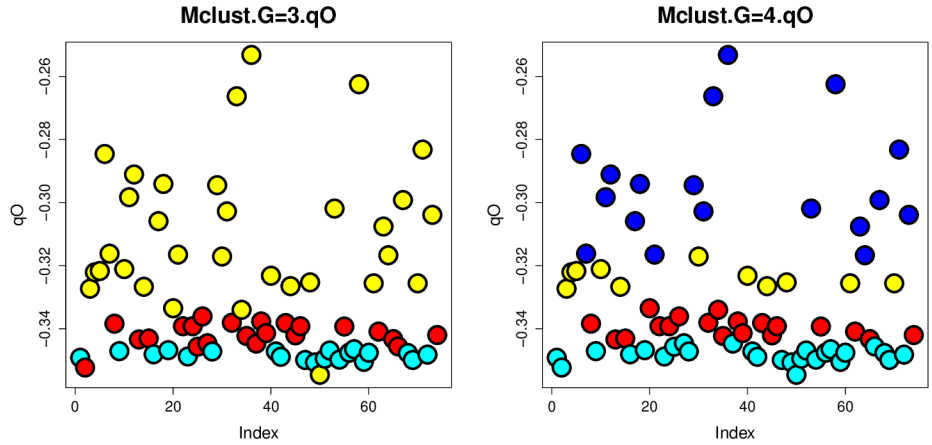
Analogicky vytvoříme grafy rozdělení molekul do shluků podle náboje $qO$.




Obrázek 3.4: Rozdělení dat podle $qH$: na dva shluky (vlevo), na tři shluky (vpravo)

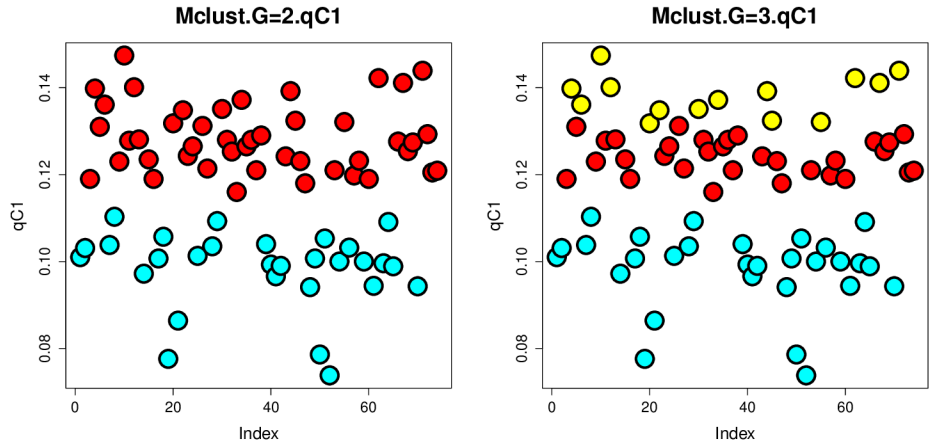



Obrázek 3.5: Rozdělení dat: podle $qH$ na čtyři shluky (vlevo), podle $qO$ na dva shluky (vpravo)



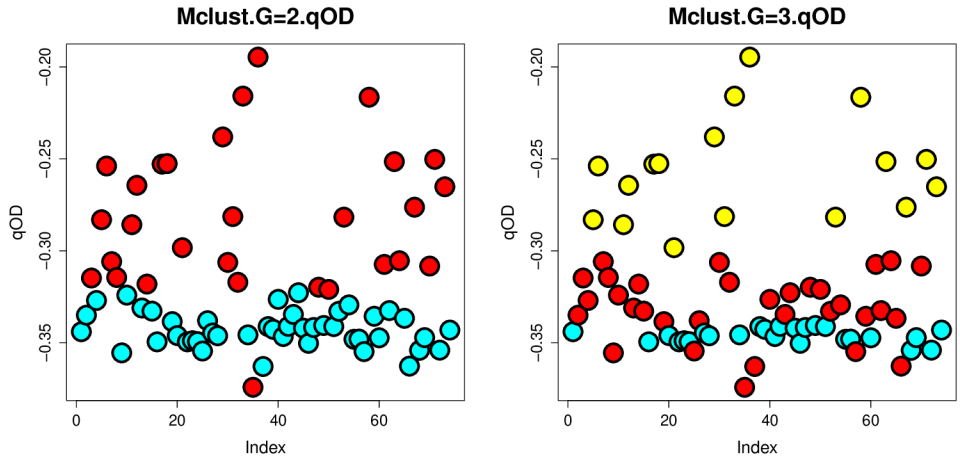

Obrázek 3.6: Rozdělení dat podle $qO$: na tři shluky (vlevo), na čtyři shluky (vpravo)

Z grafů je patrné, že i když má funkce `Mclust` možnost rozdělit data na čtyři skupiny, vystačí si se třemi (viz obrázek 3.5 (vlevo)). U ostatních sedmnácti souborů vypadá situace obdobně a to nejen pro shlukování podle $qH$ a $qO$, ale i podle zbylých typů nábojů. Při zadání parametru `G=4` rozdělí `Mclust` data buď na tři skupiny, nebo na čtyři, ovšem pak je jedna skupina zastoupená nízkým počtem prvků. (Ve více než polovině případů je počet prvků jedné ze skupin $n_s < 10$, $s = 1 \ldots 4$). Rozdělení na čtyři skupiny je vyrovnané pouze při rozdělení dat podle hodnot $pKa$.[3] Pro zbývající typy nábojů zde uvedeme ilustrační grafy rozdělení molekul na dva a tři shluky. Grafy rozdělení dat na dva, tři i čtyři shluky podle jednotlivých nábojů lze najít na cd ve složkách ***3-qH, 4-qO, 5-qC1, 6-qOD, 7-qC1D*** a ***8-pKa***.

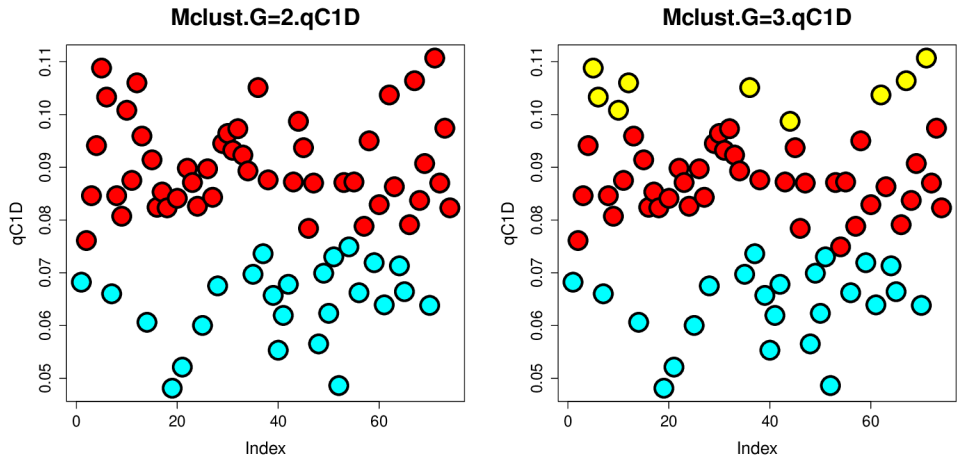



Obrázek 3.7: Rozdělení dat podle $qC1$:na dva shluky (vlevo), na tři shluky (vpravo)

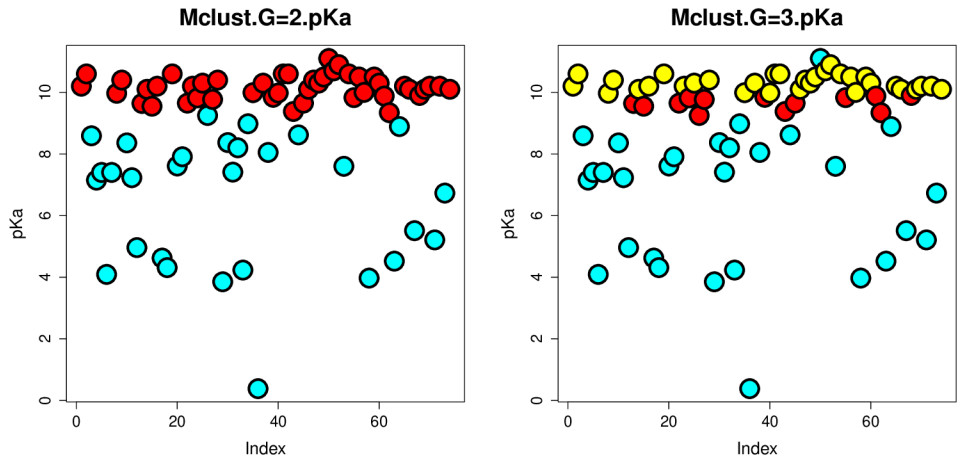
[3]Toto kritérium rozdělení ovšem není moc praktické, protože $pKa$ je ta proměnná, jejíž střední hodnotu se z modelu snažíme odhadnout. Rozdělení molekul podle $pKa$ by mělo smysl pouze v případě, že bychom byli schopni určit (dejme tomu na základě chemických zkušeností), do které skupiny by pravděpodobně molekula mohla patřit. Pomocí modelu bychom pak už jen dopočítali její přesnou hodnotu. To ale není naším záměrem, proto rozdělení dat podle hodnoty $pKa$ nebudeme v dalším textu uvažovat.




Obrázek 3.8: Rozdělení dat podle $qOD$: na dva shluky (vlevo), na tři shluky (vpravo)




Obrázek 3.9: Rozdělení dat podle $qC1D$: na dva shluky (vlevo), na tři shluky (vpravo)

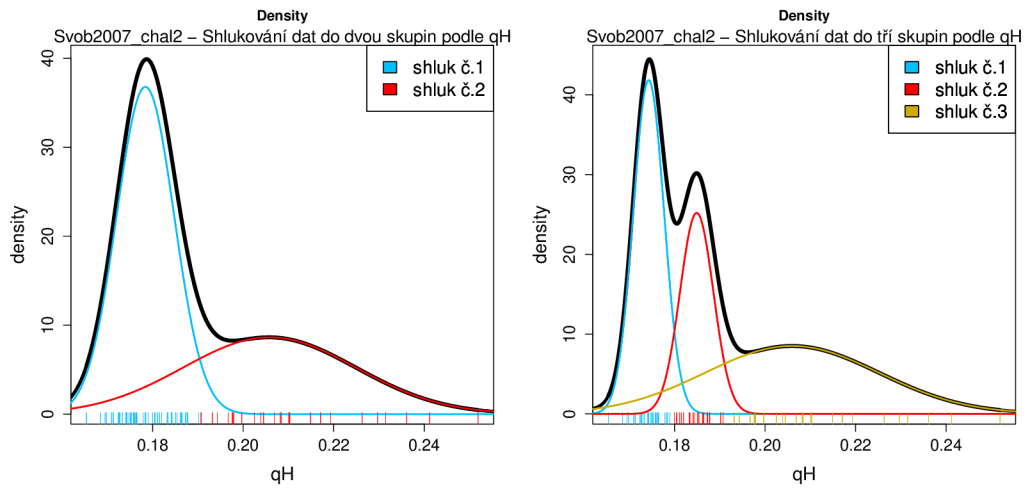



Obrázek 3.10: Rozdělení dat podle $pKa$: na dva shluky (vlevo), na tři shluky (vpravo)

Pro lepší názornost si vykreslíme ještě jeden typ grafů a sice *grafy smíšené hustoty*. Na nich je velmi pěkně vidět, proč byly jednotlivé molekuly zařazeny právě do těch shluků, do kterých byly zařazeny. Algoritmus na vykreslení grafů smíšené hustoty je poněkud složitější, proto ho na tomto místě nebudeme rozebírat, uvedeme si pouze jeho tvar. Podrobnější popis algoritmu je uveden v příloze B.

```
qH<-dat$qH
qH.Mclust<-Mclust(qH,G=3)
pocet_shluku<-qH.Mclust$G
prumery_shluku<-qH.Mclust$parameters$mean
pomery_shluku<-qH.Mclust$parameters$pro
rozptyly_shluku<-qH.Mclust$parameters$variance$sigmasq
barvy<-c("deepskyblue","red","gold3")

USR<-par('usr')
xx<-seq(USR[1],USR[2],length=1000)
fx<-matrix(rep(0,1000*pocet_shluku),ncol=pocet_shluku)
for(k in 1:pocet_shluku){
  fx[,k]<-pomery_shluku[k]*dnorm(xx,mean=prumery_shluku[k],
                                 sd=sqrt(rozptyly_shluku[k]))}
fxx<-apply(fx,1,sum)

plot(qH.Mclust,what="density",ngrid=1000,lwd=5)
mtext('Shlukování dat do tří skupin podle qH')
lines(xx,fxx,col="black",lwd=5)

for(k in 1:pocet_shluku) {
lines(xx,fx[,k],col=barvy[k],lwd=2)
    rug(qH[qH.Mclust$classification==k],col=barvy[k])}
```

Výstupem algoritmu je graf smíšené hustoty pro rozdělení dat do tří shluků podle náboje $qH$, viz obr. 3.11 (vpravo). Obdobně bychom postupovali pro rozdělení dat na shluky dva.

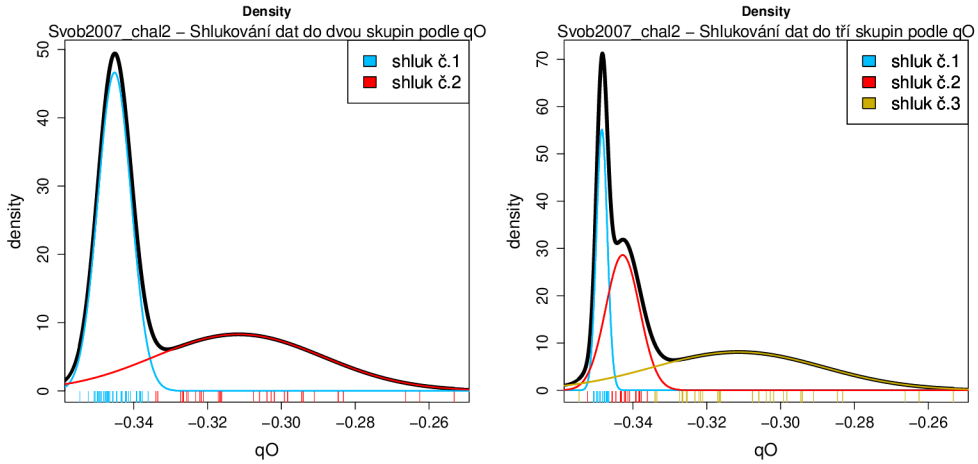



Obrázek 3.11: Graf smíšené hustoty: rozdělení dat podle $qH$: na dva shluky (vlevo), na tři shluky (vpravo)

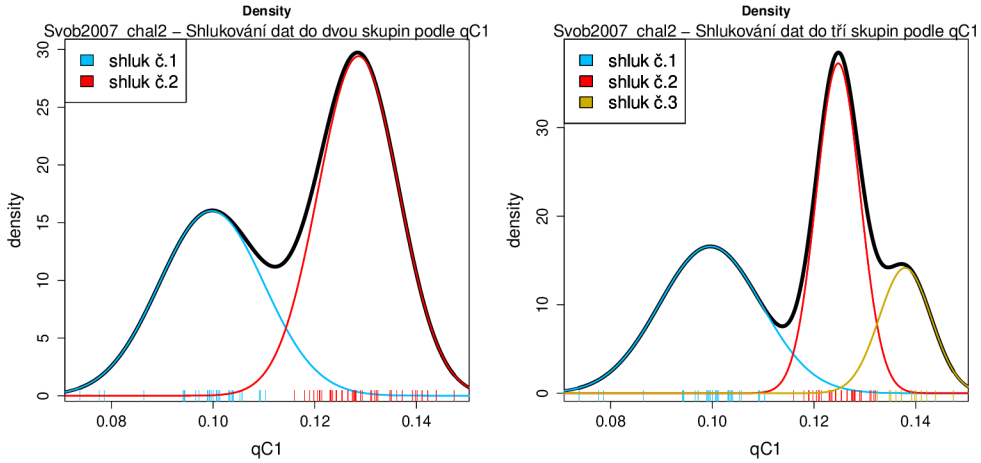
### Zařazování molekul do shluků

U každé molekuly najdeme velikost jejího náboje na ose $x$ a podíváme se na funkční hodnoty hustot jednotlivých shluků v tomto bodě. Molekulu zařadíme do toho shluku, jehož hustota má v daném bodě nejvyšší funkční hodnotu. Např. při rozdělení dat do dvou skupin má molekula s nábojem $qH = 0.18$ při zařazení do první skupiny funkční hodnotu $f_{s1}(0.18) \approx 37$, zatímco při zařazení do druhé skupiny se $f_{s2}(0.18) \approx 6$. Molekulu tedy zařadíme do skupiny č.1.

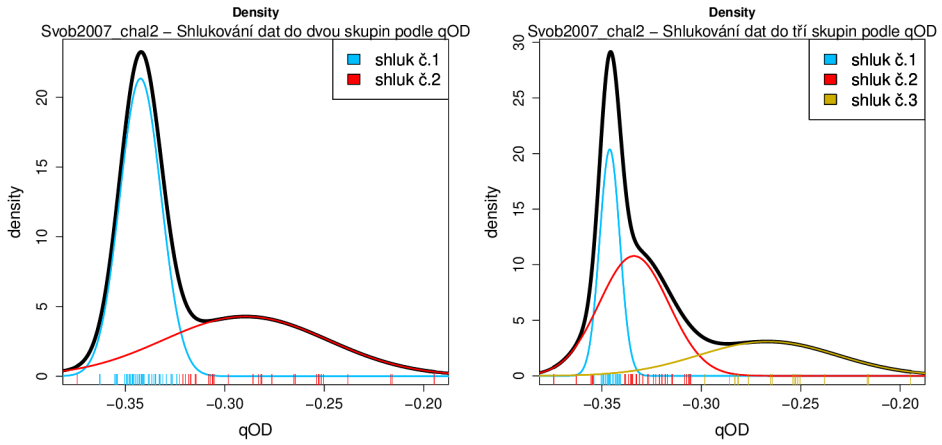
Analogicky potom vykreslíme i grafy smíšené hustoty pro shlukování podle ostatních nábojů.



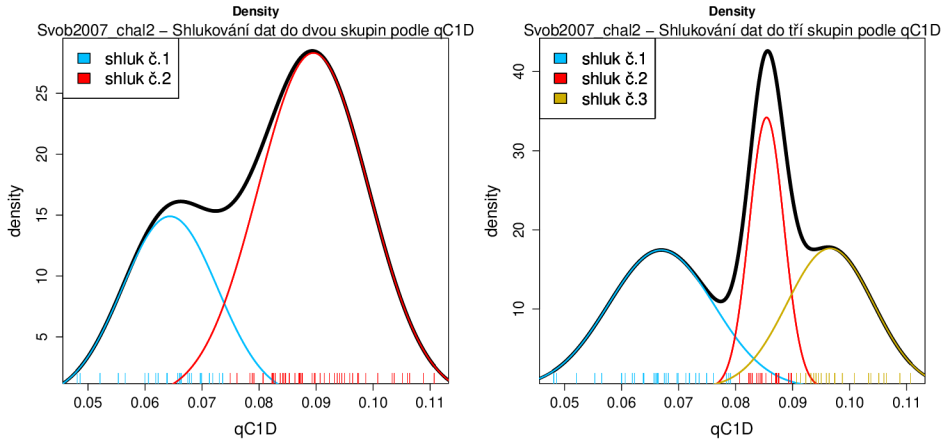

Obrázek 3.12: Grafy smíšené hustoty: rozdělení dat podle $qO$: na dva shluky (vlevo), na tři shluky (vpravo)



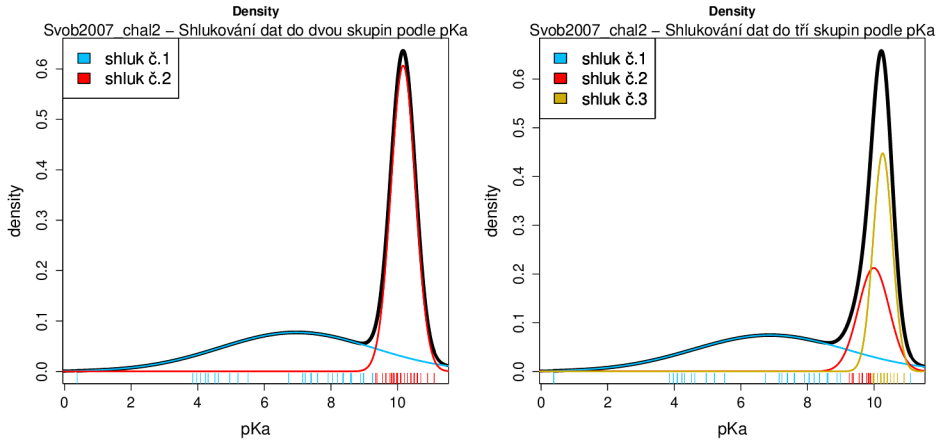

Obrázek 3.13: Grafy smíšené hustoty: rozdělení dat podle $qC1$: na dva shluky (vlevo), na tři shluky (vpravo)




Obrázek 3.14: Grafy smíšené hustoty: rozdělení dat podle $qOD$: na dva shluky (vlevo), na tři shluky (vpravo)




Obrázek 3.15: Grafy smíšené hustoty: rozdělení dat podle $qC1D$: na dva shluky (vlevo), na tři shluky (vpravo)



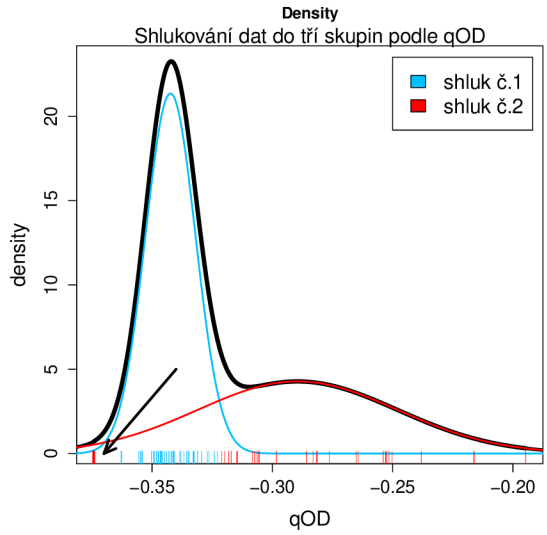

Obrázek 3.16: Grafy smíšené hustoty: rozdělení dat podle $pKa$: na dva shluky (vlevo), na tři shluky (vpravo)

*Poznámka.* Grafy pro ostatních 17 souborů jsou uvedeny na cd v podsložkách složky ***9-grafy_smisene_hustoty***.

Po prozkoumání všech možností shlukování na dvě a tři skupiny jsme došli k závěru, že data je možné rozdělit na dvě skupiny podle kteréhokoli náboje, neboť pro všechny soubory je dělení až na výjimky jednoznačné. Výjimkami máme na mysli soubor Bult2002_hir při dělení podle $qH$, $qO$ a $qC1$, soubor Baek1991 pro shlukování podle $qC1D$ a soubory Baek1991, Svob2007_cbeg2, Svob2007_hm2, Svob2007_cmet2 a Svob2007_chal2 při shlukování podle $qOD$.

V souboru Bult2002_hir funkce `Mclust` při rozdělování zařazuje molekulu č.50 (název NSC_5105) do skupiny č.2 namísto do skupiny č.1 a u ostatních souborů zařazuje molekulu č.35 (název NSC_36808) opět do druhé skupiny místo do první. Tento nedostatek bychom mohli řešit odstraněním těchto molekul z datového souboru, což ale není rozumné, protože analýza reziduí na odlišnosti těchto bodů nepoukázala, nebo jejich přeřazením do skupiny č.1. Při dalších výpočtech je zkusíme přeřadit do první skupiny a porovnat výsledky modelu před a po jejich přemístění.

*Poznámka.* Na souboru Svob2007_chal2 si ukážeme, proč při rozdělení dat podle náboje $qOD$ byla molekula č.35 zařazena do skupiny č.2 namísto do skupiny č.1. K tomu nám poslouží právě graf smíšené hustoty. Hodnota náboje $qOD$ pro molekulu NSC_36808 $qOD_{NSC_36808} = -0.3742$. Na obrázku 3.17 si na x-ové ose najdeme bod $-0.3742$ (ten leží v místě červené čárky, na kterou ukazuje šipka) a podíváme se na funkční hodnoty obou hustot v tomto bodě. Vidíme, že v těchto místech je hodnota hustoty druhé skupiny vyšší než hodnota hustoty první skupiny.Proto byl bod č.35 zařazen do skupiny č.2. Tento výkyv však může být způsoben nedostatečným množstvím dat. Vzhledem k tomu, že funkční hodnoty obou hustot v bodě -0.3742 se nijak výrazně neliší, přikloníme se k možnosti zkusit molekulu přesunout do skupiny č.1. Obdobné analýzy můžeme provést také pro ostatní problémové molekuly.




Obrázek 3.17: Graf smíšené hustoty: rozdělení dat podle $qOD$: na dva shluky

Grafy pro shlukování do tří skupin ukázaly, že data lze rozdělit jednoznačně podle hodnot nábojů $qH$ a $qO$, opět až na nějaké výjimky. Z toho méně výjimek nastává v případě rozdělení molekul podle náboje $qH$, kde optimalitu rozdělení nabourává tatáž molekula č.$50$ ve čtyřech souborech (Baek1991, Bult2002_mul, Chaves2006, Ouy2009) [4].

Po konzultaci s chemickými pracovníky byl podpořen postup rozdělení dat podle náboje na vodíku $qH$. Shlukování by v tomto případě mělo i chemické opodstatnění.

**Závěr shlukové analýzy**

Pro další analýzu jsme vyselektovali modely založené na shlukování do dvou skupin podle všech nábojů a do tří skupin podle náboje $qH$ a $qO$.

### 3.1.4 gls 3d.modely

Modely, které k predikci střední hodnoty disociační konstanty $pKa$ použijeme, se od obyčejných lineárních regresních modelů liší tím, že při prováděné regresi berou ohled na zařazení molekul do příslušných shluků. Jinými slovy v modelu bude vystupovat faktor (v našem případě číslo dané skupiny), podle kterého bude lineární regrese provedena. K tomu využijeme funkci `gls` z knihovny **nlme**. Příkaz zadáváme ve tvaru `gls(tvar.modelu, data=datovy.soubor, weights=vahova.funkce)`.

Další kroky analýzy si demonstrujeme na `gls`-modelu s rozdělením dat na tři shluky podle nábojů $qH$. Nejprve ale musíme z vektoru s3 udělat faktor. Vytvořený model uložíme do proměnné g3.qH.3d.

```
library(nlme)
qH.Mclust<-Mclust(qH,G=3)
s3<-qH.Mclust$classification
s3<-as.factor(s3)
dat3<-cbind(dat, s3)
g3.qH.3d<-gls(pKa~(qH+qO+qC1)*s3, data=dat3,
          weights=varIdent(form=~1|s3))
g3.qH.3d

Generalized least squares fit by REML
  Model: pKa ~ (qH + qO + qC1) * s3
  Data: dat3
  Log-restricted-likelihood: -4.597458

Coefficients:
 (Intercept)           qH           qO          qC1          s32
```

[4] Rozdělení podle ostatních nábojů není tak optimální. Shlukování podle $qOD$ a $qC1D$ není v mnoha souborech jednoznačné a v některých je navíc nevyrovnané. Rozdělení molekul do shluků podle $qC1$ nese lepší výsledky než rozdělení podle $qOD$ a $qC1D$, ovšem v porovnání s dělením podle $qH$ a $qO$ je výrazně horší. Navíc není univerzální pro všechny soubory, protože v případě Svob2007_cmet2 a Svob2007_hm2 rozděluje data pouze do dvou shluků.

```
   5.4519946  -23.3903696  -28.9253131  -10.0005252  118.9380142
         s33        qH:s32       qH:s33       qO:s32       qO:s33
   5.2275113 -296.2010396  -57.1807265  201.0530286   -0.2270246
     qC1:s32      qC1:s33
  27.8271010   40.8324365

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | s3
 Parameter estimates:
       1        3        2
1.000000 7.554013 3.349444
Degrees of freedom: 74 total; 62 residual
Residual standard error: 0.1501403
```

*Poznámka.* Váhová funkce `varIdent(form=~1|s3)` se používá, když rozptyl mezi jednotlivými úrovněmi faktoru je různý, zatímco rozptyl dat v každé úrovni je homogenní.

Z výstupu funkce `gls` vidíme, že algoritmus hledání parametrů je založen na modifikované metodě maximální věrohodnosti (REML)[5]. Důležitými hodnotami pro nás budou odhad $\sigma^2$ (*Residual standard error*) a rozptyly jednotlivých shluků uvedených v sekci *Parameter estimates*. Shluk č.1 je shluk s nejmenším rozptylem, funkce `gls` tedy stanovila jeho rozptyl jako referenční a přiřadila mu hodnotu jedné. Rozptyl druhého shluku je potom $3.35$ krát větší než rozptyl shluku prvního, rozptyl třetího je větší $7.55$ krát.

*Poznámka.* Na základě hodnot uvedených v sekci *Parameter estimates* je potom možné dopočítat váhu každého pozorování podle skupiny, do které toto pozorování náleží.

Z celé tabulky je pro nás nejdůležitější sekce *Coefficients*, která obsahuje odhady parametrů. Koeficienty $qH$, $qO$, $qC1$ a $Intercept$ jsou parametry příslušné k prvkům z prvního shluku. Pro prvky z druhé skupiny bude dopočítání parametrů trochu složitější. K parametru $qH$ přičteme hodnotu parametru $qH : s32$, k parametru $qO$ hodnotu $qO : s32$, k parametru $qC1$ hodnotu $qC1 : s32$ a k parametru $Intercept$ hodnotu $s32$. Obdobně pro prvky ze třetí skupiny. Celkem tedy získáváme tři rovnice (pro každý shluk jednu).

Konkrétně pro model g3.qH.3d:

$$
\begin{aligned}
pKa &= -23.39\,qH - 28.93\,qO - 10.00\,qC1 + 5.45 \\
pKa &= (-23.39 - 296.20)\,qH + (-28.93 + 201.05)\,qO + (-10.00 + 27.83)\,qC1 + \\
&\quad + (5.45 + 118.94) \\
pKa &= (-23.39 - 57.18)\,qH + (-28.93 - 0.23)\,qO + (-10.00 + 40.83)\,qC1 + \\
&\quad + (5.45 + 5.23)
\end{aligned}
$$

[5]Odhady vypočítané pomocí metody REML jsou, narozdíl od odhadů získaných běžnou metodou maximální věrohodnosti, nevychýlené.

Z geometrického hlediska bychom mohli říci, že každá skupina molekul je proložena vlastní nadrovinou. Tyto nadroviny jsou navzájem různoběžné a posunuté o hodnoty $s32$ a $s33$.
Podrobnější informace o `gls`-modelu získáme příkazem `summary(g3.qH.3d)`.

Funkce `gls` nedává jako výstup koeficient determinace (*R-squared)*, proto jej budeme muset dopočítat pomocí vlastního algoritmu podle vzorce

$$R^2 = 1 - \frac{S_e}{S_t},$$

kde $S_e = (Y - \hat{Y})'(Y - \hat{Y})$ je reziduální součet čtverců a $S_t = (Y - \bar{Y})'(Y - \bar{Y})$ je celkový součet čtverců.

```
pKa<-dat$pKa
ph_g3.qH.3d<-predict(g3.qH.3d)
SE_g3.qH.3d<-(pKa-ph_g3.qH.3d)%*%(pKa-ph_g3.qH.3d)
ST<-(pKa-mean(pKa))%*%(pKa-mean(pKa))
R_2_g3.qH.3d<-c(1-SE_g3.qH.3d/ST)
R_2_g3.qH.3d

[1] 0.9160403
```

Vypočítanou hodnotu $R^2_{g3.qH.3d}$ porovnáme s hodnotou koeficientu determinace původního modelu mod.3d.

```
R_2_mod.3d<-summary(mod.3d)$r.squared
R_2_mod.3d
[1] 0.8736875
```

Hodnota $R^2_{g3.qH.3d}$ `gls`-modelu je vyšší než hodnota $R^2_{mod.3d}$ původního modelu. Koeficienty determinace pro ostatní soubory rozdělené do shluků podle náboje $qH$ a dále také podle náboje $qO$ jsou uvedeny v tabulce 3.3.

| | $R^2_{mod.3d}$ | $R^2_{g3.qH.3d}$ | $R^2_{g3.qH.3d.P}$ | $R^2_{g3.qO.3d}$ | $R^2_{g3.qO.3d.P}$ |
|---|---|---|---|---|---|
| Baek1991 | 0.90986 | 0.94542 | 0.94778 | 0.94542 | 0.94778 |
| Bult2002_che | 0.86948 | 0.91350 | 0.91350 | 0.91359 | 0.91359 |
| Bult2002_hir | 0.84149 | 0.88348 | 0.88348 | 0.88289 | 0.88289 |
| Bult2002_mk | 0.86388 | 0.90496 | 0.90496 | 0.90496 | 0.90496 |
| Bult2002_mul | 0.88764 | 0.93362 | 0.93649 | 0.93362 | 0.93649 |
| Bult2002_npa | 0.90440 | 0.94519 | 0.94519 | 0.94519 | 0.94519 |
| Bult2004_aim | 0.86462 | 0.90896 | 0.90896 | 0.90734 | 0.90734 |
| Chaves2006 | 0.89103 | 0.93320 | 0.93534 | 0.93320 | 0.93534 |
| Jir2008_dft | 0.86958 | 0.90462 | 0.90462 | 0.87576 | 0.87696 |
| Jir2008_hf | 0.86121 | 0.89340 | 0.89340 | 0.86709 | 0.86888 |
| Mort1986 | 0.88604 | 0.92568 | 0.92568 | 0.92568 | 0.92568 |
| Ouy2009 | 0.87308 | 0.90993 | 0.91180 | 0.87668 | 0.87668 |
| Ouy2009_elem | 0.87269 | 0.91552 | 0.91552 | 0.88110 | 0.88110 |
| Ouy2009_elemF | 0.88482 | 0.94938 | 0.94938 | 0.89673 | 0.89673 |
| Svob2007_cbeg2 | 0.86710 | 0.91400 | 0.91400 | 0.87590 | 0.87666 |
| Svob2007_chal2 | 0.87369 | 0.91604 | 0.91604 | 0.88070 | 0.88119 |
| Svob2007_cmet2 | 0.86628 | 0.91482 | 0.91482 | 0.87779 | 0.87875 |
| Svob2007_hm2 | 0.86713 | 0.91614 | 0.91614 | 0.87843 | 0.87927 |

Tabulka 3.3: Srovnání koeficientů determinace pro modely g3.qH.3d a g3.qO.3d

V tabulce 3.3 jsou zaznamenány koeficienty determinace

- původního modelu ($R^2_{mod.3d}$),
- `gls`-modelu s rozdělením molekul do shluků podle nábojů $qH$ ($R^2_{g3.qH.3d}$) resp. $qO$ ($R^2_{g3.qO.3d}$),
- `gls`-modelu s rozdělením molekul do shluků podle náboje $qH$ ($R^2_{g3.qH.3d.P}$) případně $qO$ ($R^2_{g3.qO.3d.P}$) s molekulami[6] přeřazenými do skupiny č.1.

Výsledky ukazují, že hodnota $R^2$ vychází u všech souborů lépe (nebo stejně) pro shlukování podle náboje $qH$. Navíc u souborů s přeřazenou molekulou č.50 do první skupiny se hodnota koeficientu determinace nepatrně zlepšila. Z toho důvodu molekulu NSC_5105 trvale přeřadíme do skupiny č.1. Získáváme tak jednoznačné kritérium pro zařazení nové neznámé molekuly do příslušného shluku, a sice hodnotu náboje na vodíku $qH$.

Analogický postup použijeme pro rozdělení molekul do dvou shluků. Výsledky jsou uvedeny v tabulce 3.4.

| | $R^2_{mod.3d}$ | $R^2_{g2.qH.3d}$ | $R^2_{g2.qO.3d}$ | $R^2_{g2.qC1.3d}$ | $R^2_{g2.qOD.3d}$ | $R^2_{g2.qC1D.3d}$ |
|---|---|---|---|---|---|---|
| Baek1991 | 0.90986 | 0.94258 | 0.94258 | 0.94258 | 0.93852 | 0.93852 |
| Bult2002_che | 0.86948 | 0.90592 | 0.90592 | 0.90592 | 0.91451 | 0.91451 |
| Bult2002_hir | 0.84149 | 0.84797 | 0.84797 | 0.84797 | 0.91693 | 0.91693 |
| Bult2002_mk | 0.86388 | 0.90089 | 0.90089 | 0.90089 | 0.90874 | 0.90874 |
| Bult2002_mul | 0.88764 | 0.92499 | 0.92499 | 0.92499 | 0.93420 | 0.93420 |
| Bult2002_npa | 0.90440 | 0.93685 | 0.93685 | 0.93685 | 0.94422 | 0.94422 |
| Bult2004_aim | 0.86462 | 0.90593 | 0.90593 | 0.90593 | 0.90593 | 0.90593 |
| Chaves2006 | 0.89103 | 0.92529 | 0.92529 | 0.93145 | 0.93423 | 0.93423 |
| Jir2008_dft | 0.86958 | 0.88324 | 0.87310 | 0.88672 | 0.87229 | 0.88644 |
| Jir2008_hf | 0.86121 | 0.88846 | 0.86863 | 0.87929 | 0.86745 | 0.87743 |
| Mort1986 | 0.88604 | 0.92467 | 0.92467 | 0.92467 | 0.95409 | 0.95409 |
| Ouy2009 | 0.87308 | 0.88197 | 0.87674 | 0.89306 | 0.87657 | 0.88348 |
| Ouy2009_elem | 0.87269 | 0.89693 | 0.87916 | 0.88817 | 0.87737 | 0.90180 |
| Ouy2009_elemF | 0.88482 | 0.88761 | 0.89686 | 0.88731 | 0.88712 | 0.88767 |
| Svob2007_cbeg2 | 0.86710 | 0.89523 | 0.87640 | 0.87418 | 0.87879 | 0.91792 |
| Svob2007_chal2 | 0.87369 | 0.90999 | 0.88077 | 0.88892 | 0.88289 | 0.91596 |
| Svob2007_cmet2 | 0.86628 | 0.89691 | 0.87800 | 0.86920 | 0.87422 | 0.91708 |
| Svob2007_hm2 | 0.86713 | 0.89787 | 0.87877 | 0.87229 | 0.88068 | 0.91865 |

Tabulka 3.4: Srovnání koeficientů determinace pro modely typu g2.3d

Na základě hodnot $R^2$ vybereme model g2.qC1D.3d jako nejvhodnější. Nakonec srovnáme hodnory koeficientů determinace nejlepšího modelu typu g3.3d a typu g2.3d, viz tabulka 3.5.

[6]Na mysli máme molekulu č.50 pro shlukování dat podle náboje $qH$ a molekuly č.50 a č.2 pro shlukování dat podle náboje $qO$.

| | $R^2_{mod.3d}$ | $R^2_{g3.qH.3d}$ | $R^2_{g2.qC1D.3d}$ |
|---|---|---|---|
| Baek1991 | 0.90986 | 0.94778 | 0.93782 |
| Bult2002_che | 0.86948 | 0.91350 | 0.91451 |
| Bult2002_hir | 0.84149 | 0.88348 | 0.91693 |
| Bult2002_mk | 0.86388 | 0.90496 | 0.90874 |
| Bult2002_mul | 0.88764 | 0.93649 | 0.93420 |
| Bult2002_npa | 0.90440 | 0.94519 | 0.94422 |
| Bult2004_aim | 0.86462 | 0.90896 | 0.90593 |
| Chaves2006 | 0.89103 | 0.93534 | 0.93423 |
| Jir2008_dft | 0.86958 | 0.90462 | 0.88644 |
| Jir2008_hf | 0.86121 | 0.89340 | 0.87743 |
| Mort1986 | 0.88604 | 0.92568 | 0.95409 |
| Ouy2009 | 0.87308 | 0.91180 | 0.88348 |
| Ouy2009_elem | 0.87269 | 0.91552 | 0.90180 |
| Ouy2009_elemF | 0.88482 | 0.94938 | 0.88767 |
| Svob2007_cbeg2 | 0.86710 | 0.91400 | 0.91792 |
| Svob2007_chal2 | 0.87369 | 0.91604 | 0.91596 |
| Svob2007_cmet2 | 0.86628 | 0.91482 | 0.91708 |
| Svob2007_hm2 | 0.86713 | 0.91614 | 0.91865 |

Tabulka 3.5: Srovnání modelů g3.qH.3d a g2.qC1D.3d

Ze srovnání obou modelů vyplývá, že model g3.qH.3d vykazuje lepší výsledky než model g2.qC1D.3d. Ověřme tuto hypotézu funkcí `anova.gls`. Tato funkce z knihovny **nlme** je přízpůsobena k porovnávání `gls` modelů, umí ale srovnávat pouze modely, jejichž parametry byly odhadnuty metodou maximální věrohodnosti (ML), proto před srovnáním musíme oba modely převést z REML na ML. To uděláme pomocí parametru `method='ML'` při vytváření modelu příkazem `gls`. Výstupem funkce `anova` je ANOVA tabulka.

```
qH<-dat$qH
s3.qH<-Mclust(qH,G=3)$classification
s3.qH<-as.factor(s3.qH)
dat3.qH<-cbind(dat,s3.qH)
g3.qH.3d<-gls(pKa~(qH+qO+qC1)*s3.qH,data=dat3.qH,
  weights=varIdent(form=~1|s3.qH),method='ML')

qC1D<-dat$qC1D
s2.qC1D<-Mclust(qC1D,G=2)$classification
s2.qC1D<-as.factor(s2.qC1D)
dat2.qC1D<-cbind(dat,s2.qC1D)
g2.qC1D.3d<-gls(pKa~(qH+qO+qC1)*s2.qC1D, data=dat2.qC1D,
    weights=varIdent(form=~1|s2.qC1D),method='ML')
anova(g3.qH.3d, g2.qC1D.3d)

        Model df       AIC      BIC    logLik   Test
g3.qH.3d       1  15  96.50757 131.0685 -33.25379
g2.qC1D.3d     2  10 152.97059 176.0112 -66.48530 1 vs 2

               L.Ratio p-value
g3.qH.3d
g2.qC1D.3d    66.46302  <.0001
```

Z ANOVA tabulky nás nejvíce zajímá hodnota *p–value*, která nám říká, na jaké hladině významnosti lze zamítnout nulovou hypotézu o rovnocennosti obou mo-

delů. Vzhledem k tomu, že hodnota *p-value* $< 0.0001$, nulovou hypotézu zamítáme na hladině významnosti 0.0001.

Analogicky jsme dopočítali p-hodnoty pro zbývajících sedmnáct souborů. U šestnácti z nich byla zamítnuta nulová hypotéza o rovnocennosti testovaných modelů g3.qH.3d a g2.qC1D.3d na hladině významnosti minimálně 0.01. Pouze u souboru Mort1968 vyšla *p–value* $< 0.1552$. Protože však jedním z našich cílů je snaha o univerzální postup pro všechny soubory a tento soubor je jediný odlišný, a navíc rozdíl koeficientů determinace $|R^2_{g3.qH.3d} - R^2_{g2.qC1D.3d}| \approx 3$, budeme soubor Mort1968 také popisovat modelem g3.qH.3d.

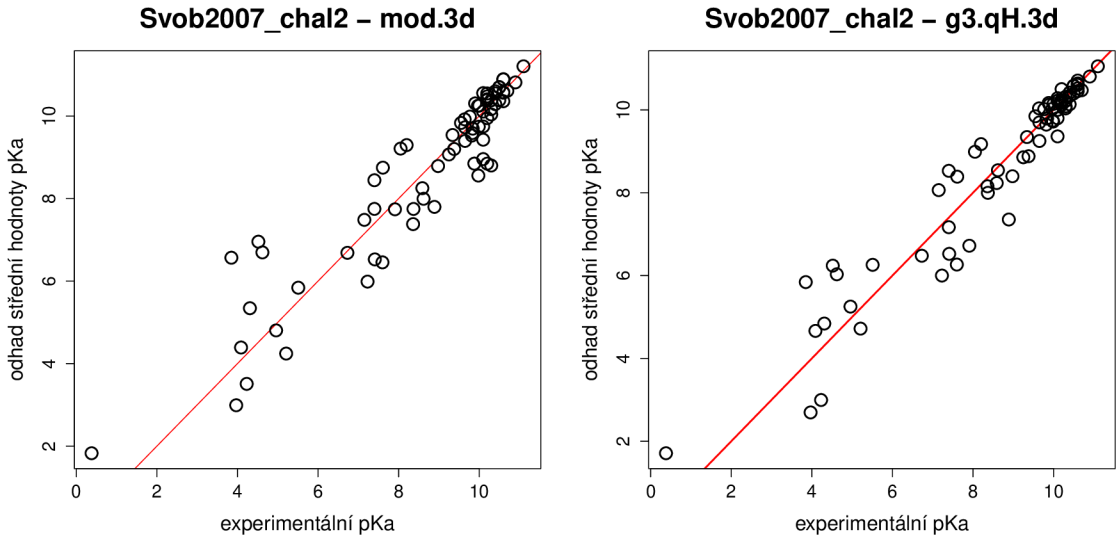
### Závěr pro gls 3d.modely

Ze všech navrhovaných 3d.modelů jsme na základě hodnot koeficientů determinace a p-hodnot vybrali za nejvhodnější `gls`-model

$$pKa \sim (qH + qO + qC1) * s3,$$

kde $s3$ je faktor rozdělující datový soubor do tří skupin podle velikosti nábojů na atomu vodíku $qH$.

Podívejme se nakonec, jak se zlepšil graf odhadů středních hodnot disociační konstanty $pKa$ pomocí modelu g3.qH.3d oproti modelu mod.3d.




Obrázek 3.18: Grafy závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ pro modely: a) mod.3d (vlevo), b) g3.qH.3d (vpravo)

*Poznámka.* Grafy pro ostatní soubory viz složka ***10-g3.qH.3d.predict.exp.pKa***.

### Výpočet kritérií kvality modelu

Pro nový model g3.qH.3d vypočítáme požadovaná kritéria kvality modelu

- $RMSE$ podle vzorce $\sqrt{\frac{S_e}{n}}$,
- $\Delta$ podle vzorce $\sum_{i=1}^{n} |Y - \hat{Y}|$,
- $\sigma^2$.

Hodnotu $\sigma^2$ získáme pomocí příkazu `summary(g3.qH.3d)$sigma`, ostatní hodnoty musíme dopočítat ručně.

```
#RMSE
  g3.qH.3d<-gls(pKa~(qH+qO+qC1)*s3,data=dat3,
    weights=varIdent(form=~1|s3))
  ph_g3.qH.3d<-predict(g3.qH.3d)
  SE_g3.qH.3d<-(pKa-ph_g3.qH.3d)%*%(pKa-ph_g3.qH.3d)
  RMSE_g3.qH.3d<-sqrt((SE_g3.qH.3d)/74)

 [1] 0.6453535

#DELTA
  ph_g3.qH.3d<-predict(g3.qH.3d)
  delta_g3.qH.3d<-sum(abs(pKa-ph_g3.qH.3d))

 [1] 32.22787

#SIGMA
  sigma_g3.qH.3d<-summary(g3.qH.3d)$sigma

 [1] 0.1501403
```

Porovnání výsledků kritérií modelů g3.qH.3d a mod.3d jsou uvedeny v tabulkách 3.6 a 3.7. Všechna kritéria nového modelu vyšla výrazně lépe něž kritéria původního modelu mod.3d.

| | $R^2_{mod.3d}$ | $R^2_{g3.qH.3d}$ | $RMSE_{mod.3d}$ | $RMSE_{g3.qH.3d}$ |
|---|---|---|---|---|
| Baek1991 | 0.90986 | 0.94778 | 0.66869 | 0.50894 |
| Bult2002_che | 0.86948 | 0.91350 | 0.80465 | 0.65506 |
| Bult2002_hir | 0.84149 | 0.88348 | 0.88673 | 0.76026 |
| Bult2002_mk | 0.86388 | 0.90496 | 0.82171 | 0.68661 |
| Bult2002_mul | 0.88764 | 0.93649 | 0.74655 | 0.56130 |
| Bult2002_npa | 0.90440 | 0.94519 | 0.68863 | 0.52143 |
| Bult2004_aim | 0.86462 | 0.90896 | 0.81947 | 0.67200 |
| Chaves2006 | 0.89103 | 0.93534 | 0.73521 | 0.56634 |
| Jir2008_dft | 0.86958 | 0.90462 | 0.80432 | 0.68783 |
| Jir2008_hf | 0.86121 | 0.89340 | 0.82974 | 0.72718 |
| Mort1986 | 0.88604 | 0.92568 | 0.75188 | 0.60716 |
| Ouy2009 | 0.87308 | 0.90993 | 0.79345 | 0.66146 |
| Ouy2009_elem | 0.87269 | 0.91552 | 0.79469 | 0.64733 |
| Ouy2009_elemF | 0.88482 | 0.94938 | 0.75588 | 0.50109 |
| Svob2007_cbeg2 | 0.86710 | 0.91400 | 0.81194 | 0.65314 |
| Svob2007_chal2 | 0.87369 | 0.91604 | 0.79156 | 0.64535 |
| Svob2007_cmet2 | 0.86628 | 0.91482 | 0.81444 | 0.65004 |
| Svob2007_hm2 | 0.86713 | 0.91614 | 0.81185 | 0.64496 |

Tabulka 3.6: Srovnání kritérií kvality modelů mod.3d a g3.qH.3d I

| | $\Delta_{mod.3d}$ | $\Delta_{g3.qH.3d}$ | $\sigma^2_{mod.3d}$ | $\sigma^2_{g3.qH.3d}$ |
|---|---|---|---|---|
| Baek1991 | 39.30965 | 28.51997 | 0.68753 | 0.14503 |
| Bult2002_che | 44.16958 | 35.64147 | 0.82732 | 0.25381 |
| Bult2002_hir | 47.02954 | 39.21315 | 0.91171 | 0.24921 |
| Bult2002_mk | 45.16455 | 36.61013 | 0.84486 | 0.23703 |
| Bult2002_mul | 43.60145 | 31.46375 | 0.76759 | 0.15588 |
| Bult2002_npa | 39.35611 | 27.44795 | 0.70803 | 0.14424 |
| Bult2004_aim | 45.83070 | 36.65289 | 0.84256 | 0.29901 |
| Chaves2006 | 42.18101 | 31.76143 | 0.75593 | 0.14491 |
| Jir2008_dft | 41.10677 | 34.92593 | 0.82699 | 0.17701 |
| Jir2008_hf | 43.05719 | 35.52714 | 0.85312 | 0.15491 |
| Mort1986 | 42.69883 | 32.76005 | 0.77306 | 0.21722 |
| Ouy2009 | 40.04273 | 33.96790 | 0.81581 | 0.14428 |
| Ouy2009_elem | 40.43316 | 32.15449 | 0.81709 | 0.15323 |
| Ouy2009_elemF | 38.40580 | 26.51320 | 0.77717 | 0.21061 |
| Svob2007_cbeg2 | 42.28310 | 32.21599 | 0.83482 | 0.16558 |
| Svob2007_chal2 | 40.97783 | 32.22787 | 0.81386 | 0.15014 |
| Svob2007_cmet2 | 42.73003 | 32.08321 | 0.83738 | 0.15122 |
| Svob2007_hm2 | 42.77684 | 31.92495 | 0.83473 | 0.15024 |

Tabulka 3.7: Srovnání kritérií kvality modelů mod.3d a g3.qH.3d II

## 3.2 5d.modely

Analýza 5d.modelu probíhá obdobně jako analýza 3d.modelu. Z toho důvodu si uvedeme jen R-kové příkazy a výstupy pro soubor Svob2007_chal2.

### 3.2.1 Seznámení s 5d.modelem

```
dat<-read.delim('Svob2007_chal2.csv',sep=';',dec='.')
names(dat)

[1]  "N"                 "pKa"               "qH"
[4]  "qO"                "qC1"               "qOD"
[7]  "qC1D"              "outlier"           "pKa.3d.model."
[10] "pKa.3d.WO.model." "pKa.5d.model."

mod.5d <- lm(pKa~qH+qO+qC1+qOD+qC1D,data=dat)
mod.5d

Call:
lm(formula = pKa ~ qH + qO + qC1 + qOD + qC1D, data = dat)

Coefficients:
(Intercept)           qH           qO          qC1
      53.72      -165.62        37.04        91.68

        qOD         qC1D
      12.49       -98.72
```

Výsledný model má tvar

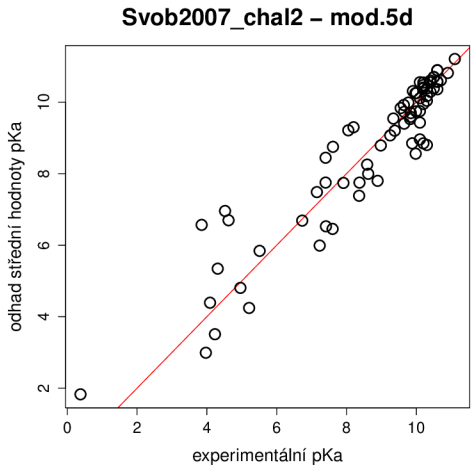
$$pKa = -165.62\,qH + 37.04\,qO + 91.68\,qC1 + 12.49\,qOD - 98.72\,qC1D + 53.72,$$

kde $qH$ je velikost náboje na atomu vodíku, $qO$ na atomu kyslíku, $qC1$ na atomu uhlíku, $qOD$ na atomu kyslíku v rámci disociované molekuly a $qC1D$ na atomu uhlíku v rámci disociované molekuly.

```
summary(mod.5d)

Call:
lm(formula = pKa ~ qH + qO + qC1 + qOD + qC1D, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16132 -0.22291  0.08681  0.40792  1.18556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    53.72      19.89   2.700 0.008733 **
qH           -165.62      41.34  -4.006 0.000155 ***
qO             37.04      72.47   0.511 0.610920
qC1            91.68      79.91   1.147 0.255285
qOD            12.49      46.98   0.266 0.791167
qC1D          -98.72      84.46  -1.169 0.246531
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6561 on 68 degrees of freedom
Multiple R-squared: 0.9203,Adjusted R-squared: 0.9144
F-statistic:   157 on 5 and 68 DF,  p-value: < 2.2e-16
```

Koeficient determinace $R^2_{mod.5d} = 0.9203$. To je mnohem vyšší hodnota ve srovnání s koeficientem determinace původního 3d.modelu ($R^2_{mod.3d} = 0.8737$). Hodnota $\sigma^2 = 0.6561$.

```
ph_mod.5d<-predict(mod.5d)
pKa<-dat$pKa
plot(pKa,ph_mod.5d, ylab=('odhad střední hodnoty pKa'),
     xlab=('experimentální pKa'),
     abline(a=0, b=1, col='red'),cex=1.5,
     main='Svob2007_chal2 - mod.5d')
```

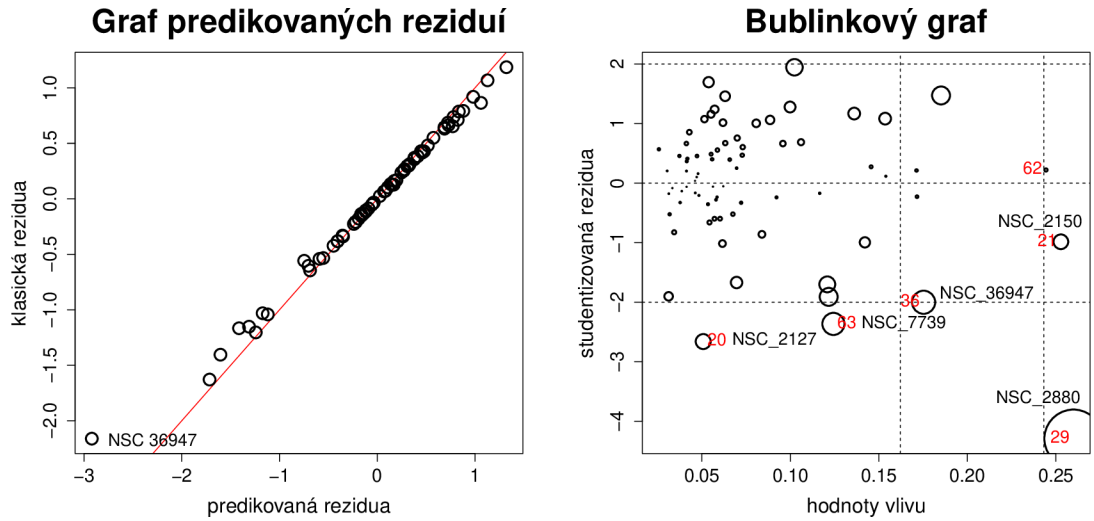



Obrázek 3.19: Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$

*Poznámka.* Grafy pro všech osmnáct souborů lze najít na cd ve složce ***11-5d.model.predict.exp.pKa***.

```
res_classic<-resid(mod.5d)
hatval<-hatvalues(mod.5d)
res_predict<-res_classic/(1-hatval)
plot(res_predict, res_classic, xlab='klasická rezidua',
     ylab='predikovaná rezidua',
     main='Graf predikovaných reziduí',cex=1.5,
     abline(a=0, b=1, col='red'))

library(car)
influencePlot(mod.5d, main='Bublinkový graf',
  xlab='hodnoty vlivu',
              ylab='studentizovaná rezidua',
              id.n=3, id.col='red')
text(x=0.240, y=-3.50, 'NSC_2880')
text(x=0.165, y=-2.34, 'NSC_7739')
text(x=0.093, y=-2.65, 'NSC_2127')
text(x=0.210, y=-0.97, 'NSC_2150')
text(x=0.215, y=-1.85, 'NSC_36947')
```

```
      StudRes         Hat      CookD
20 -2.6597374 0.05083171 0.24075752
21 -0.9849783 0.25284185 0.23397233
29 -4.2923483 0.25997702 0.92667633
36 -2.0008413 0.17529882 0.36854727
62  0.2210674 0.24442011 0.05169352
63 -2.3644048 0.12433459 0.35203669
```




(a) Graf predikovaných reziduí

(b) Bublinkový graf

Obrázek 3.20: Grafy k analýze reziduí 5d.modelu

Bublinkový graf vyznačil molekuly č.20, č.21, č.29, č.36, č.62 a č.63 za podezřelé. Cookova vzdálenost je pro všechny molekuly menší než 1, vliv kterékoli z nich není větší než 0.26. Žádný bod nepůsobí na ostatní natolik, abychom ho museli z modelu odstranit. Bod č.29 vykazuje nějvětší hodnotu studentizovaného rezidua a také nejvyšší vliv na ostatní molekuly. Jeho odstraněním ze souboru ovšem nedošlo k významnému zlepšení modelu, proto ho v datovém souboru ponecháme.

*Poznámka.* Grafy predikovaných reziduí pro model `mod.5d` jsou přístupné na cd ve složce ***12-5d.grafy_predikovanych_rezidui***.

**Závěr analýzy reziduí**

Analýza reziduí nevedla k odstranění žádného bodu z datového souboru.

### 3.2.2 Shluková analýza

Připomeňme si tři důležitá kritéria pro výběr nejvhodnějšího způsobu rozdělení dat do shluků a závěr shlukové analýzy. Ten je shodný jak pro 3d.model, tak pro 5d.model. Kritéria jsou

- jednoznačné rozdělení molekul do shluků,
- co nejvyšší hodnota koeficientu determinace $R^2$,
- stejné kritérium shlukování pro všech osmnáct chemických metod.

**Závěr shlukové analýzy**

Na základě grafů jednotlivých nábojů v závislosti na indexu a grafů smíšené hustoty jsme se rozhodli rozdělit model na tři shluky podle náboje $qH$ případně $qO$ a na dva shluky podle všech nábojů. S ohledem na výsledky 3d.modelu bychom předpokládali, že nejlepších výsledků dosáhneme při rozdělení datového souboru na tři skupiny podle $qH$ a na dvě skupiny podle $qC1D$. Modely vyhodnotíme před a po přesunutí výše zmíněných problémových molekul do skupiny č.1.

### 3.2.3 gls 5d.model

Postup si předvedeme na `gls`-modelu g3.qH.5d pro shlukování do tří skupin podle nábojů $qH$.

```
library(mclust)
library(nlme)
qH<-dat$qH
qH.Mclust<-Mclust(qH, G=3)
s3<-qH.Mclust$classification
s3<-as.factor(s3)
dat3<-cbind(dat, s3)
g3.qH.5d<-gls(pKa~(qH+qO+qC1+qC1D+qOD)*s3, data=dat3,
              weights=varIdent(form=~1|s3))
```

```
g3.qH.5d

Generalized least squares fit by REML
  Model: pKa ~ (qH + qO + qC1 + qC1D + qOD) * s3
  Data: dat3
  Log-restricted-likelihood: 33.55198

Coefficients:
 (Intercept)           qH           qO          qC1         qC1D
  15.6638966  -58.1855389  -16.0178236    4.9272668  -14.8819964
        qOD          s32          s33       qH:s32       qH:s33
   0.5511906  119.7023743   52.9601489 -291.9843045 -137.1067431
      qO:s32       qO:s33       qC1:s32      qC1:s33     qC1D:s32
 216.7584096  110.3481538   23.5931991  119.2422903    1.8200386
    qC1D:s33      qOD:s32      qOD:s33
 -96.1388841  -12.6470294   -9.2311021

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | s3
 Parameter estimates:
       1        3        2
1.000000 5.170001 3.197109
Degrees of freedom: 74 total; 56 residual
Residual standard error: 0.1480267
```

Z výstupní tabulky funkce `gls` vidíme, že shluk č.1 má nejmenší (referenční) rozptyl, druhý shluk má rozptyl $3.19$ krát větší než první shluk a rozptyl třetího je větší $5.17$ krát. Hodnota $\sigma^2 = 0.148$. Sekce *Coefficients* nám dává celkem 18 parametrů pro tři rovnice:

$$\begin{aligned} pKa &= -58.2\,qH - 16.0\,qO + 4.9\,qC1 + 0.6\,qOD - 14.9\,qC1D + 15.7 \\ pKa &= (-58.2 - 292.0)\,qH + (-16.0 + 216.8)\,qO + (4.9 + 23.6)\,qC1 + \\ &\quad + (0.6 - 12.6)\,qOD + (-14.9 + 1.8)\,qC1D + (15.7 + 119.7) \\ pKa &= (-58.2 - 137.1)\,qH + (-16.0 + 110.3)\,qO + (4.9 + 119.2)\,qC1 + \\ &\quad + (0.6 - 9.2)\,qOD + (-14.9 - 96.1)\,qC1D + (15.7 + 53.0) \end{aligned}$$

Koeficient determinace $R^2$ dopočítáme podle vzorce

$$R^2 = 1 - \frac{S_e}{S_t}$$

a porovnáme s koeficientem determinace původního modelu $R^2_{mod.5d}$.

```
pKa<-dat$pKa
ph_g3.qH.5d<-predict(g3.qH.5d)
SE_g3.qH.5d<-(pKa-ph_g3.qH.5d)%*%(pKa-ph_g3.qH.5d)
ST<-(pKa-mean(pKa))%*%(pKa-mean(pKa))
R_2_g3.qH.5d<-c(1-SE_g3.qH.5d/ST)
R_2_g3.qH.5d

[1] 0.9602048
```

```
R_2_mod.5d<-summary(mod.5d)$r.squared
R_2_mod.5d

[1] 0.920269
```

Vidíme, že hodnota $R^2_{g3.qH.5d}$ se zvýšila oproti $R^2_{mod.5d}$ přibližně o 0.04. Analogicky vypočítáme koeficienty determinace pro ostatní soubory a také pro rozdělení dat podle náboje $qO$. V tabulce 3.8 jsou ve sloupcích $R^2_{g3.qH.5d}$ a $R^2_{g3.qO.5d}$ uvedeny hodnoty koeficientů determinace pro modely g3.qH.5d a g3.qO.5d ve sloupcích $R^2_{g3.qH.5d.P}$ a $R^2_{g3.qO.5d.P}$ koeficienty determinace pro modely s přeřazenými molekulami č.50 a č.2 do skupiny č.1.

| | $R^2_{mod.5d}$ | $R^2_{g3.qH.5d}$ | $R^2_{g3.qH.5d.P}$ | $R^2_{g3.qO.5d}$ | $R^2_{g3.qO.5d.P}$ |
|---|---|---|---|---|---|
| Baek1991 | 0.91948 | 0.96771 | 0.96912 | 0.96771 | 0.96912 |
| Bult2002_che | 0.90570 | 0.96208 | 0.96208 | 0.96057 | 0.96057 |
| Bult2002_hir | 0.90499 | 0.96251 | 0.96251 | 0.96241 | 0.96241 |
| Bult2002_mk | 0.91310 | 0.96499 | 0.96499 | 0.96499 | 0.96499 |
| Bult2002_mul | 0.91575 | 0.96825 | 0.96844 | 0.96825 | 0.96844 |
| Bult2002_npa | 0.91804 | 0.95981 | 0.95981 | 0.95981 | 0.95981 |
| Bult2004_aim | 0.90175 | 0.95862 | 0.95862 | 0.95722 | 0.95722 |
| Chaves2006 | 0.91916 | 0.96758 | 0.96848 | 0.96758 | 0.96848 |
| Jir2008_dft | 0.91475 | 0.95350 | 0.95350 | 0.93896 | 0.94672 |
| Jir2008_hf | 0.91543 | 0.95544 | 0.95544 | 0.92926 | 0.93736 |
| Mort1986 | 0.91419 | 0.95519 | 0.95519 | 0.95519 | 0.95519 |
| Ouy2009 | 0.90943 | 0.95102 | 0.95122 | 0.93392 | 0.93392 |
| Ouy2009_elem | 0.91317 | 0.95741 | 0.95741 | 0.93643 | 0.93643 |
| Ouy2009_elemF | 0.88658 | 0.95856 | 0.95856 | 0.90407 | 0.90407 |
| Svob2007_cbeg2 | 0.91792 | 0.95947 | 0.95947 | 0.93144 | 0.93915 |
| Svob2007_chal2 | 0.92027 | 0.96020 | 0.96020 | 0.93228 | 0.93918 |
| Svob2007_cmet2 | 0.91889 | 0.95967 | 0.95967 | 0.93309 | 0.94056 |
| Svob2007_hm2 | 0.91794 | 0.95977 | 0.95977 | 0.93270 | 0.93940 |

Tabulka 3.8: Srovnání koeficientů determinace pro modely g3.qH.5d a g3.qO.5d

Hodnoty $R^2$ ukázaly, že model se shlukováním podle $qH$ vysvětluje větší procento variability v datech, tedy je vhodnějším modelem než model se shlukováním podle nábojů $qO$. Zaměřme se tedy na model g3.qH.5d. Přesun molekuly NSC_5150 ze skupiny č.2 do skupiny č.1 vedl ve všech případech k mírnému zlepšení koeficientu determinace, proto podpoříme myšlenku přesunout v daných souborech tuto molekulu do první skupiny, čímž sjednotíme a navzájem od sebe oddělíme intervaly shlukování dat podle náboje na vodíku. Analogicky provedeme výpočet statistiky $R^2$ pro modely g2.qH.5d, . . . , g2.qC1D.5d. Výsledky jsou uvedeny v tabulce 3.9.

| | $R^2_{mod.5d}$ | $R^2_{g3.qH.5d}$ | $R^2_{g3.qO.5d}$ | $R^2_{g3.qC1.5d}$ | $R^2_{g3.qOD.5d}$ | $R^2_{g3.qC1D.5d}$ |
|---|---|---|---|---|---|---|
| Baek1991 | 0.91948 | 0.95314 | 0.95314 | 0.95314 | 0.95780 | 0.95780 |
| Bult2002_che | 0.90570 | 0.94528 | 0.94528 | 0.94528 | 0.95827 | 0.95827 |
| Bult2002_hir | 0.90499 | 0.92397 | 0.92397 | 0.92397 | 0.95062 | 0.95062 |
| Bult2002_mk | 0.91310 | 0.96219 | 0.96219 | 0.96219 | 0.96519 | 0.96519 |
| Bult2002_mul | 0.91575 | 0.95656 | 0.95656 | 0.95656 | 0.96825 | 0.96825 |
| Bult2002_npa | 0.91804 | 0.95756 | 0.95756 | 0.95756 | 0.95949 | 0.95949 |
| Bult2004_aim | 0.90175 | 0.95055 | 0.95055 | 0.95055 | 0.95055 | 0.95055 |
| Chaves2006 | 0.91916 | 0.95492 | 0.95492 | 0.96579 | 0.96666 | 0.96666 |
| Jir2008_dft | 0.91475 | 0.95043 | 0.93403 | 0.93441 | 0.94811 | 0.93910 |
| Jir2008_hf | 0.91543 | 0.95352 | 0.93649 | 0.93895 | 0.94951 | 0.92868 |
| Mort1986 | 0.91419 | 0.95340 | 0.95340 | 0.95340 | 0.96178 | 0.96178 |
| Ouy2009 | 0.90943 | 0.94636 | 0.93205 | 0.92659 | 0.94819 | 0.92043 |
| Ouy2009_elem | 0.91317 | 0.94211 | 0.93149 | 0.92345 | 0.94690 | 0.92821 |
| Ouy2009_elemF | 0.88658 | 0.89028 | 0.90219 | 0.89087 | 0.89133 | 0.89109 |
| Svob2007_cbeg2 | 0.91792 | 0.94835 | 0.93855 | 0.93428 | 0.95208 | 0.93937 |
| Svob2007_chal2 | 0.92027 | 0.94855 | 0.93850 | 0.94510 | 0.95016 | 0.93669 |
| Svob2007_cmet2 | 0.91889 | 0.95007 | 0.93997 | 0.93103 | 0.94508 | 0.93722 |
| Svob2007_hm2 | 0.91794 | 0.94914 | 0.93883 | 0.93507 | 0.95113 | 0.93952 |

Tabulka 3.9: Srovnání koeficientů determinace pro modely typu g2.5d

V případě rozdělení molekul do dvou shluků vykazuje nejlepší výsledky shlukování podle náboje $qOD$. Druhým nejlepším kritériem je dělení podle $qC1D$. Shluková analýza dat podle $qH$ vykazuje třetí nejlepší výsledky. Nakonec si uvedeme tabulku srovnávající výsledky $R^2$ pro nejlepší model typu g3.5d a model typu g2.5d. Viz tabulka 3.10.

| | $R^2_{mod.5d}$ | $R^2_{g3.qH.5d}$ | $R^2_{g2.qOD.5d}$ |
|---|---|---|---|
| Baek1991 | 0.91948 | 0.96912 | 0.95780 |
| Bult2002_che | 0.90570 | 0.96208 | 0.95827 |
| Bult2002_hir | 0.90499 | 0.96251 | 0.95062 |
| Bult2002_mk | 0.91310 | 0.96499 | 0.96519 |
| Bult2002_mul | 0.91575 | 0.96844 | 0.96825 |
| Bult2002_npa | 0.91804 | 0.95981 | 0.95949 |
| Bult2004_aim | 0.90175 | 0.95862 | 0.95055 |
| Chaves2006 | 0.91916 | 0.96848 | 0.96666 |
| Jir2008_dft | 0.91475 | 0.95350 | 0.94811 |
| Jir2008_hf | 0.91543 | 0.95544 | 0.94951 |
| Mort1986 | 0.91419 | 0.95519 | 0.96178 |
| Ouy2009 | 0.90943 | 0.95122 | 0.94819 |
| Ouy2009_elem | 0.91317 | 0.95741 | 0.94690 |
| Ouy2009_elemF | 0.88658 | 0.95856 | 0.89133 |
| Svob2007_cbeg2 | 0.91792 | 0.95947 | 0.95208 |
| Svob2007_chal2 | 0.92027 | 0.96020 | 0.95016 |
| Svob2007_cmet2 | 0.91889 | 0.95967 | 0.94508 |
| Svob2007_hm2 | 0.91794 | 0.95977 | 0.95113 |

Tabulka 3.10: Srovnání modelů g3.qH.5d a g2.qOD.5d

Hodnoty koeficientů determinace ukazují, že model g3.qH.5d dává lepší výsledky než model g2.qOD.5d. Podíváme se tedy na ANOVA tabulku.

```
qH<-dat$qH
s3.qH<-Mclust(qH, G=3)$classification
```

```
s3.qH<-as.factor(s3.qH)
dat3.qH<-cbind(dat,s3.qH)
g3.qH.5d<-gls(pKa~(qH+qO+qC1+qOD+qC1D)*s3.qH, data=dat3.qH,
              weights=varIdent(form=~1|s3.qH), method='ML')

qOD<-dat$qOD
s2.qOD<-Mclust(qOD, G=2)$classification
s2.qOD<-as.factor(s2.qOD)
dat2.qOD<-cbind(dat, s2.qOD)
g2.qOD.5d<-gls(pKa~(qH+qO+qC1+qOD+qC1D)*s2.qOD, data=dat2.qOD,
               weights=varIdent(form=~1|s2.qOD), method='ML')
anova(g3.qH.5d, g2.qOD.5d)

         Model df      AIC      BIC    logLik   Test
g3.qH.5d     1 21  78.6031 126.9885 -18.30155
g2.qOD.5d    2 14 131.2573 163.5142 -51.62866 1 vs 2

          L.Ratio p-value
g3.qH.5d
g2.qOD.5d 66.65422  <.0001
```

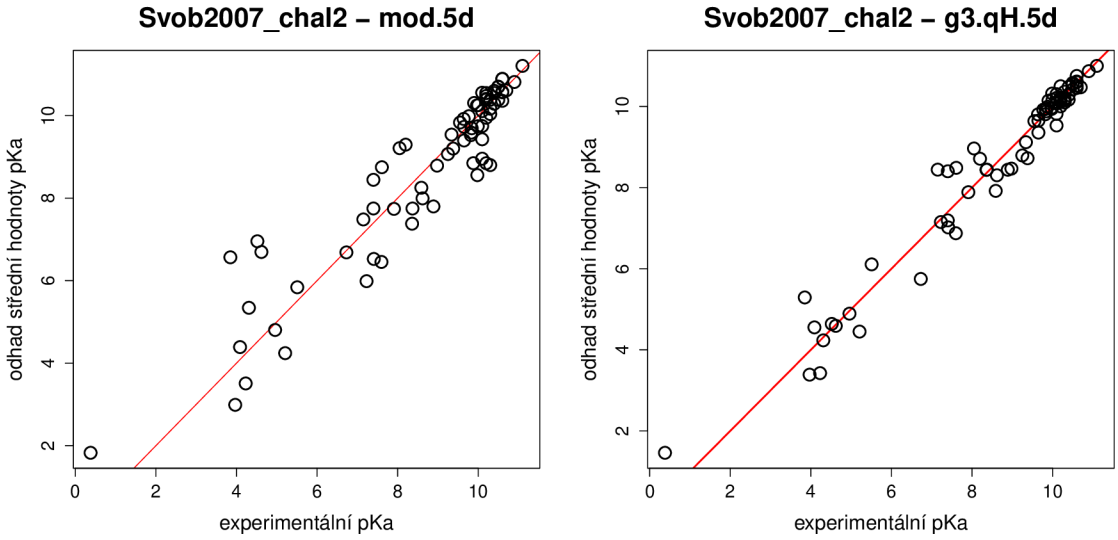
Protože *p-value* $< 0.0001$ zamítáme nulovou hypotézu o rovnocennosti modelů g3.qH.5d a g2.qOD.5d na hladině významnosti 0.0001. Nulovou hypotézu jsme zamítli na základě hodnoty *p-value* také pro zbylých sedmnáct souborů.

### Závěr pro gls 5d.modely

Za nejvhodnější 5d.model zvolíme `gls`-model

$$pKa \sim (qH + qO + qC1 + qOD + qC1D) * s3,$$

kde $s3$ je faktor nesoucí informace o rozdělení datového souboru do tří skupin podle hodnot nábojů $qH$.




Obrázek 3.21: Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ pro modely: a) mod.5d (vlevo), b) g3.qH.5d (vpravo)

*Poznámka.* Grafy pro ostatní soubory viz složka ***13-g3.qH.5d.predict.exp.pKa***.

### Výpočet kritérií kvality modelu

Pro nový `gls`-model g3.qH.5d vypočítáme kritéria kvality modelu

- $RMSE$,
- $\Delta$,
- $\sigma^2$.

```
#RMSE
  g3.qH.5d<-gls(pKa~(qH+qO+qC1+qOD+qC1D)*s3, data=dat3,
                weights=varIdent(form=~1|s3))
  ph_g3.qH.5d<-predict(g3.qH.5d)
  SE_g3.qH.5d<-(pKa-ph_g3.qH.5d)%*%(pKa-ph_g3.qH.5d)
  RMSE_g3.qH.5d<-sqrt((SE_g3.qH.5d)/74)
  RMSE_g3.qH.5d
[1] 0.4443015

#DELTA
  delta_g3.qH.5d<-sum(abs(pKa-ph_g3.qH.5d))
  delta_g3.qH.5d
[1] 21.97376

#SIGMA
  sigma_g3.qH.5d<-summary(g3.qH.5d)$sigma
  sigma_g3.qH.5d
[1] 0.1480267
```

Výsledky hodnot $R^2$, RMSE, $\Delta$ a $\sigma^2$ spolu s hodnotami kritérií původního modelu mod.5d jsou uvedeny v tabulkách 3.11 a 3.12.

| | $R^2_{mod.5d}$ | $R^2_{g3.qH.5d}$ | $RMSE_{mod.5d}$ | $RMSE_{g3.qH.5d}$ |
|---|---|---|---|---|
| Baek1991 | 0.91948 | 0.96912 | 0.63198 | 0.39137 |
| Bult2002_che | 0.90570 | 0.96208 | 0.68392 | 0.43368 |
| Bult2002_hir | 0.90499 | 0.96251 | 0.68651 | 0.43126 |
| Bult2002_mk | 0.91310 | 0.96499 | 0.65656 | 0.41674 |
| Bult2002_mul | 0.91575 | 0.96844 | 0.64646 | 0.39568 |
| Bult2002_npa | 0.91804 | 0.95981 | 0.63761 | 0.44652 |
| Bult2004_aim | 0.90175 | 0.95862 | 0.69812 | 0.45309 |
| Chaves2006 | 0.91916 | 0.96848 | 0.63326 | 0.39542 |
| Jir2008_dft | 0.91475 | 0.95350 | 0.65028 | 0.48028 |
| Jir2008_hf | 0.91543 | 0.95544 | 0.64771 | 0.47013 |
| Mort1986 | 0.91419 | 0.95519 | 0.65241 | 0.47149 |
| Ouy2009 | 0.90943 | 0.95122 | 0.67027 | 0.49188 |
| Ouy2009_elem | 0.91317 | 0.95741 | 0.65630 | 0.45962 |
| Ouy2009_elemF | 0.88658 | 0.95856 | 0.75007 | 0.45339 |
| Svob2007_cbeg2 | 0.91792 | 0.95947 | 0.63808 | 0.44840 |
| Svob2007_chal2 | 0.92027 | 0.96020 | 0.62889 | 0.44430 |
| Svob2007_cmet2 | 0.91889 | 0.95967 | 0.63432 | 0.44727 |
| Svob2007_hm2 | 0.91794 | 0.95977 | 0.63800 | 0.44671 |

Tabulka 3.11: Srovnání kritérií kvality modelů mod.5d a g3.qH.5d I

| | $\Delta_{mod.5d}$ | $\Delta_{g3.qH.5d}$ | $\sigma^2_{mod.5d}$ | $\sigma^2_{g3.qH.5d}$ |
|---|---|---|---|---|
| Baek1991 | 36.45381 | 20.00825 | 0.65927 | 0.13089 |
| Bult2002_che | 39.96043 | 23.04114 | 0.71346 | 0.18728 |
| Bult2002_hir | 38.60278 | 23.05237 | 0.71616 | 0.15976 |
| Bult2002_mk | 37.60964 | 20.54906 | 0.68491 | 0.12007 |
| Bult2002_mul | 37.26124 | 20.73713 | 0.67438 | 0.13070 |
| Bult2002_npa | 36.08075 | 21.95287 | 0.66514 | 0.12969 |
| Bult2004_aim | 42.24692 | 21.60004 | 0.72827 | 0.14379 |
| Chaves2006 | 36.17447 | 20.78204 | 0.66061 | 0.12919 |
| Jir2008_dft | 36.16314 | 25.42928 | 0.67836 | 0.15032 |
| Jir2008_hf | 36.08157 | 24.48598 | 0.67568 | 0.14626 |
| Mort1986 | 38.78844 | 24.49216 | 0.68059 | 0.15871 |
| Ouy2009 | 37.20294 | 25.94221 | 0.69921 | 0.14925 |
| Ouy2009_elem | 36.65900 | 23.15367 | 0.68465 | 0.15307 |
| Ouy2009_elemF | 38.46678 | 23.33357 | 0.78246 | 0.21254 |
| Svob2007_cbeg2 | 35.57102 | 22.22674 | 0.66564 | 0.14636 |
| Svob2007_chal2 | 35.01642 | 21.97376 | 0.65605 | 0.14803 |
| Svob2007_cmet2 | 35.23264 | 22.04147 | 0.66172 | 0.14787 |
| Svob2007_hm2 | 35.38753 | 21.92830 | 0.66555 | 0.14617 |

Tabulka 3.12: Srovnání kritérií kvality modelů mod.5d a g3.qH.5d II

Z tabulek vidíme, že použitím `gls`-modelu došlo ke zlepšení všech požadovaných kritérií. Na závěr analýzy si uvedeme tabulku intervalů rozdělení dat do shluků podle nábojů $qH$ zvlášť pro každý soubor.

| | 1. shluk | 2. shluk | 3. shluk |
|---|---|---|---|
| Baek1991 | $\langle 0; 0.0557\rangle$ | $\langle 0.0557; 0.0746\rangle$ | $\langle 0.0746; \infty\rangle$ |
| Bult2002_che | $\langle 0; 0.0715\rangle$ | $\langle 0.0715; 0.0967\rangle$ | $\langle 0.0967; \infty\rangle$ |
| Bult2002_hir | $\langle 0; 0.0359\rangle$ | $\langle 0.0359; 0.0502\rangle$ | $\langle 0.0502; \infty\rangle$ |
| Bult2002_mk | $\langle 0; 0.0804\rangle$ | $\langle 0.0804; 0.1022\rangle$ | $\langle 0.1022; \infty\rangle$ |
| Bult2002_mul | $\langle 0; 0.0979\rangle$ | $\langle 0.0979; 0.1275\rangle$ | $\langle 0.1275; \infty\rangle$ |
| Bult2002_npa | $\langle 0; 0.2450\rangle$ | $\langle 0.2450; 0.2960\rangle$ | $\langle 0.2960; \infty\rangle$ |
| Bult2004_aim | $\langle 0; 0.0663\rangle$ | $\langle 0.0663; 0.1123\rangle$ | $\langle 0.1123; \infty\rangle$ |
| Chaves2006 | $\langle 0; 0.0963\rangle$ | $\langle 0.0963; 0.1187\rangle$ | $\langle 0.1187; \infty\rangle$ |
| Jir2008_dft | $\langle 0; 0.2871\rangle$ | $\langle 0.2871; 0.3144\rangle$ | $\langle 0.3144; \infty\rangle$ |
| Jir2008_hf | $\langle 0; 0.3397\rangle$ | $\langle 0.3397; 0.3656\rangle$ | $\langle 0.3656; \infty\rangle$ |
| Mort1986 | $\langle 0; 0.0193\rangle$ | $\langle 0.0193; 0.0238\rangle$ | $\langle 0.0238; \infty\rangle$ |
| Ouy2009 | $\langle 0; 0.3689\rangle$ | $\langle 0.3689; 0.4120\rangle$ | $\langle 0.4120; \infty\rangle$ |
| Ouy2009_elem | $\langle 0; 0.3997\rangle$ | $\langle 0.3997; 0.4337\rangle$ | $\langle 0.4337; \infty\rangle$ |
| Ouy2009_elemF | $\langle 0; 0.7450\rangle$ | $\langle 0.7450; 0.8723\rangle$ | $\langle 0.8723; \infty\rangle$ |
| Svob2007_cbeg2 | $\langle 0; 0.1855\rangle$ | $\langle 0.1855; 0.1970\rangle$ | $\langle 0.1970; \infty\rangle$ |
| Svob2007_chal2 | $\langle 0; 0.1799\rangle$ | $\langle 0.1799; 0.1910\rangle$ | $\langle 0.1910; \infty\rangle$ |
| Svob2007_cmet2 | $\langle 0; 0.1899\rangle$ | $\langle 0.1899; 0.2028\rangle$ | $\langle 0.2028; \infty\rangle$ |
| Svob2007_hm2 | $\langle 0; 0.1832\rangle$ | $\langle 0.1832; 0.1951\rangle$ | $\langle 0.1951; \infty\rangle$ |

Tabulka 3.13: Tabulka intervalů pro rozdělení dat do shluků podle náboje $qH$

# Závěr

Cílem této bakalářské práce bylo seznámit se se dvěma typy lineárních regresních modelů, které předpovídají hodnotu disociační konstanty $pKa$ pomocí hodnot nábojů na určitých atomech v molekule:

- 3d.modelem . . . (mod.3d)
- 5d.modelem . . . (mod.5d)

a zvýšit přesnost jejich predikcí. K dispozici jsme měli 18 datových souborů, z nichž každý obsahoval informace o 74 navzájem různých molekulách fenolů. Předpokládaný postup práce byl provést analýzu residuí, najít, prozkoumat a případně odstranit odlehlá nebo vybočující pozorování a tím zvýšit přesnost predikce.

Analýza reziduí pro všech osmnáct sad však neprokázala významný vliv žádného z pozorování. Odstranění kterékoli molekuly z datových souborů se ukázalo zbytečné.

V průběhu analýzy jsme narazili na heteroskedasticitu dat, kterou jsme se rozhodli odstranit rozdělením molekul do shluků tak, aby rozptyl v každém shluku byl homogenní. Prozkoumali jsme možnosti rozdělení dat do dvou, tří a čtyř shluků podle nábojů $qH$, $qO$, $qC1$, $qOD$, $qC1D$ a okrajově i podle hodnoty $pKa$. Jako nejvhodnější řešení se ukázalo rozdělení dat do tří skupin podle hodnoty náboje $qH$ a to z následujících důvodů:

- pro všechny soubory se data podařilo rozdělit do shluků tak, aby se jednotlivé shluky nepřekrývaly (jednoznačnost),
- hodnota koeficientu determinace se oproti původním modelům mod.3d a mod.5d výrazně zvýšila (efektivnost),
- všechny datové soubory bylo možné jednoznačně rozdělit do skupin podle stejného náboje (univerzálnost).

Původní lineární regresní modely mod.3d a mod.5d jsme nahradili obecnými lineárními regresními modely g3.qH.3d a g3.qH.5d, které uvažují heterogenní rozptyl dat. Kvalitu nových modelů jsme zhodnotili pomocí koeficientů determinace ($R^2$), jejichž hodnoty jsme srovnali s hodnotami modelů původních. Přesnost predikce (vyjádřená koeficientem determinace) se zlepšila z původních hodnot 0.86–0.91 na 0.9-0.95 v případě 3d.modelů a z hodnot 0.89-0.92 na 0.95-0.97 v případě 5d.modelů.

Pro modely g3.qH.3d a g3.qH.5d jsme dopočítali další požadovaná kritéria $\Delta$, $\sigma^2$ a $RMSE$ a zjistili jsme, že všechny hodnoty se oproti původním hodnotám zlepšily.

Nakonec jsme pro každý soubor stanovili disjunktní intervaly hodnot nábojů $qH$ pro jednotlivé shluky. To umožňuje jednoznačné (s výjimkou hraničních bodů) zařazení nové neznámé molekuly do příslušné skupiny na základě jejího náboje na atomu vodíku $qH$ a jednoduché využití modelů g3.qH.3d a g3.qH.5d v praxi.

Na závěr tedy můžeme říci, že rozdělení molekul do tří shluků podle náboje na atomu vodíku a nahrazení lineárních regresních modelů obecnými lineárními regresními modely g3.qH.3d a g3.qH.5d se zdá být účinným způsobem, jak s vysokou přesností predikovat neznámou hodnotu disociační konstanty zadaných molekul.

# Příloha A - Informace o cd

Cd obsahuje celkem 13 složek s grafy, na které se odkazujeme v textu. Konkrétně jde o složky

- 1-3d.model.predict.exp.pKa 24
- 2-3d.grafy_predikovanych_rezidui 27
- 3-qH 31
- 4-qO 31
- 5-qC1 31
- 6-qOD 31
- 7-qC1D 31
- 8-pKa 31
- 9-grafy_smisene_hustoty 36
- 10-g3.qH.3d.predict.exp.pKa 42
- 11-5d.model.predict.exp.pKa 46
- 12-5d.grafy_predikovanych_rezidui 47
- 13-g3.qH.5d.predict.exp.pKa 52

Dále je na cd uložen datový soubor Svob2007_chal2, na kterém byly demonstrovány všechny kroky analýzy.

# Příloha B - Algoritmus generující grafy hustot smíšených rozdělení

Ve třetí kapitole si na straně 33 uvádíme algoritmus na generování grafů smíšené hustoty pro rozdělení dat do tří shluků podle náboje $qH$. Tento algoritmus si zde teď rozebereme a vysvětlíme si jednotlivé příkazy, které jsou v něm využity. Protože algoritmus je poměrně dlouhý, rozdělíme si ho na několik částí, kterým se budeme postupně věnovat.

```
qH<-dat$qH
qH.Mclust<-Mclust(qH,G=3)
pocet_shluku<-qH.Mclust$G
prumery_shluku<-qH.Mclust$parameters$mean
pomery_shluku<-qH.Mclust$parameters$pro
rozptyly_shluku<-qH.Mclust$parameters$variance$sigmasq
barvy<-c("deepskyblue","red","gold3")
```

Nejprve si zadefinujeme pomocné proměnné. Zde by nás mohly překvapit tři příkazy:

- `prumery_shluku<-qH.Mclust$parameters$mean`,
- `pomery_shluku<-qH.Mclust$parameters$pro`,
- `rozptyly_shluku<-qH.Mclust$parameters$variance$sigmasq`.

Jak jsme si uvedli v kapitole 3 na straně 29, proměnná qH.Mclust je vlastně datová struktura, která obsahuje různé informace o odhadnutých parametrech směsi normálních hustot, jež jsou rozděleny do sekcí (jejich seznam můžeme získat příkazem `ls(qH.Mclust)`). Jednou z nich je sekce *parameters*, která obsahuje další čtyři podsekce, z nichž nejdůležitější pro nás budou tyto:

- *mean* - vektor průměrů jednotlivých shluků,
- *pro* - vektor vah jednotlivých shluků,
- *variance* - sekce nesoucí informace o rozptylech shluků. Její podsekce se mění s počtem kritérií, podle kterých data rozdělíme. V našem případě, kdy data dělíme pouze podle jedné proměnné (náboje $qH$), je jednou z jejích podsekcí sekce *sigmasq*, což je vektor směrodatných odchylek jednotlivých shluků.

Nakonec vytvoříme vektor barev, který využijeme při vykreslování samostatných hustot každé skupiny.

```
USR<-par('usr')
xx<-seq(USR[1],USR[2],length=1000)
fx<-matrix(rep(0,1000*pocet_shluku),ncol=pocet_shluku)
for(k in 1:pocet_shluku){
  fx[,k]<-pomery_shluku[k]*dnorm(xx,mean=prumery_shluku[k],
                                 sd=sqrt(rozptyly_shluku[k]))}
fxx<-apply(fx,1,sum)
```

`par()` je obecně funkce, která umožňuje měnit nastavení jednotlivých parametrů vytvořeného grafu. Příkaz `par(' ')` nám však poskytuje informace o hodnotách jednotlivých parametrů grafu. Konkrétně příkaz `par('usr')` má za výstup vektor tvaru $(x_1, x_2, y_1, y_2)$, kde $x_1$ je nejmenší možná $x$-ová hodnota, která může být v grafu zakreslená a $x_2$ je nejvyšší $x$-ová hodnota, která může být v grafu zakreslená. Obdobně pro hodnoty $y_1$ a $y_2$.

Dále si do proměnné xx vložíme posloupnost bodů, jež budou rozloženy v rozmezí od nejnižší možné zobrazitelné hodnoty na ose $x$ ($x_1$), do nejvyšší možné zobrazitelné hodnoty na ose $x$ ($x_2$). Příkaz `seq(min, max, length)` je příkaz pro vytváření posloupnosti od bodu `min` do bodu `max` o délce `length`. Vzdálenosti mezi jednotlivými body jsou ekvidistantní.

Do proměnné fx uložíme nulovou matici o dimenzi $1000 \times 3$. Příkaz `rep(x,y)` vezme hodnotu `x`, zopakuje ji `y`-krát a tyto hodnoty poskládá do vektoru (jeho délka bude tedy `y`). Příkaz `matrix(data, ncol=m)` vezme datový soubor data (v našem případě vektor o délce 3000) a udělá z něj matici o `m` (konkrétně `m`=3) sloupcích (tedy dimenze $dim(\mathsf{fx}) = 1000 \times 3$).

Nyní se zaměříme na cyklus. Nejdříve bychom měli poznamenat, že chceme-li z proměnné vybrat nějakou její část, zapíšeme to pomocí hranatých závorek. Tedy příkaz

```
fx[,k]
```

vybere z matice fx k-tý sloupec. Příkaz `dnorm(x,mean,sd)` vytvoří hustotu náhodného výběru `x` s průměrem `mean` a směrodatnou odchylkou `sd`. Výstupem cyklu tedy bude matice, kde v prvním sloupci budou souřadnice bodů určující hustotu rozdělení prvního shluku, v druhém sloupci body určující hustotu rozdělení druhého shluku, apod.

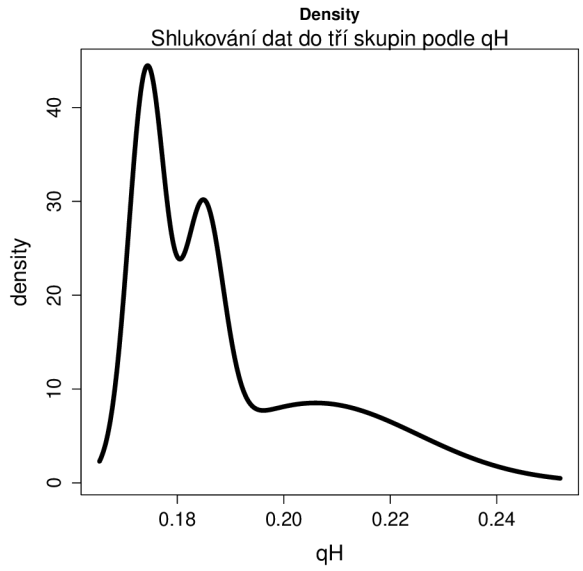
Konečně příkaz `fxx<-apply(fx, 1, sum)` vezme matici `fx` a sečte hodnoty v jednotlivých řádcích[7]. Sečte tedy funkční hodnoty smíšených hustot všech tří shluků v daném bodě. Výstupem tedy bude vektor `fxx` o délce 1000.

```
plot(qH.Mclust,what="density",ngrid=1000,lwd=5)
mtext('Shlukování dat do tří skupin podle qH')
```

[7] Že jde o řádky určujeme zadáním hodnoty druhého parametru (`1`). Kdybychom chtěli sečíst data po sloupcích, zadali bychom hodnotu druhého parametru rovnu `2`. Příkaz by měl potom tvar `fxx<-apply(fx, 2, sum)`.

```
lines(xx,fxx,col="black",lwd=4)

for(k in 1:pocet_shluku) {
lines(xx,fx[,k],col=barvy[k],lwd=2)
    rug(qH[qH.Mclust$classification==k],col=barvy[k])}
legend('topright',c('shluk č.1','shluk č.2'),fill=barvy,
       cex=1.5)
```

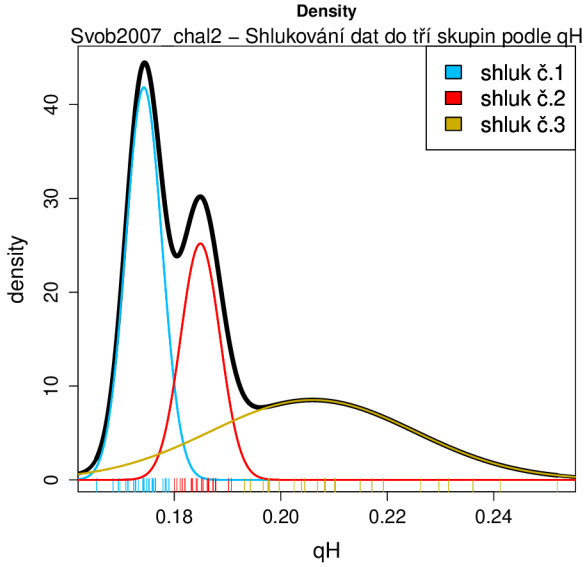
Příkaz `plot(qH.Mclust, what="density")` vykreslí křivku hustoty smíšeného rozdělení. Ta však není dotažená až do krajů grafu. Viz. graf 3.22




Obrázek 3.22: Křivka vykreslená příkazem `plot(qH.Mclust,what="density")`

Příkazem `lines(xx,fxx,col="black",lwd=5)` tuto křivku prodloužíme. Vezmeme vygenerované $x$-ové souřadnice xx a jim přiřadíme dopočítané hodnoty smíšené hustoty uložené v proměnné fxx. Získáme tedy dvourozměrné souřadnice tisíce bodů, které vykreslíme, a dostaneme tak křivku hustoty smíšeného rozdělení shodnou s křivkou vygenerovanou příkazem `plot`, ale s delšími konci.

Poslední cyklus potom vykreslí smíšené hustoty jednotlivých shluků a barevně je odliší. Příkazem `legend` vložíme do pravého horního rohu (parametr `'topright'`) legendu. Výsledkem je následující graf.




Obrázek 3.23: Graf smíšené hustoty pro rozdělení dat podle $qH$ na tři shluky

Obdobně bychom postupovali pro rozdělení dat do dvou shluků a podle ostatních typů nábojů.

# Příloha C - Přehled R-kových příkazů


`anova` 41
`apply` 58
`dnorm` 58
`gls` 37
`cooks.distance` 25
`dffits` 25
`hatvalues` 25
`influencePlot` 27
`library` 26
    **car** 26
    **mclust** 28
    **nlme** 37
`legend` 59
`lines` 59
`lm` 22
`ls` 29
`matrix` 58
`Mclust` 28
`names` 22
`plot` 24
`plot.mclust` 59
`predict` 24
`resid` 25
`read.delim` 22
`rep` 58
`rstandard` 25
`rstudent` 25
`seq` 58
`summary.gls` 39
`summary.lm` 23

# Příloha D - Seznam obrázků

# Seznam obrázků


1.1 Graf predikovaných reziduí . . . . . 10
1.2 Rankitový graf pro: a) normální rozdělení, b) rozdělení s dlouhými konci, c) rozdělení s krátkými konci . . . . . 11
1.3 Bublinkový graf . . . . . 12

3.1 Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ 24
3.2 a) Graf predikovaných reziduí, b) Bublinkový graf . . . . . 27
3.3 Graf závislosti odhadu střední hodnoty $pK$a na experimentální $pKa$ 28
3.4 Rozdělení dat podle $qH$: na dva shluky (vlevo), na tři shluky (vpravo) 30
3.5 Rozdělení dat: podle $qH$ na čtyři shluky (vlevo), podle $qO$ na dva shluky (vpravo) . . . . . 30
3.6 Rozdělení dat podle $qO$: na tři shluky (vlevo), na čtyři shluky (vpravo) 31
3.7 Rozdělení dat podle $qC1$:na dva shluky (vlevo), na tři shluky (vpravo) 31
3.8 Rozdělení dat podle $qOD$: na dva shluky (vlevo), na tři shluky (vpravo) 32
3.9 Rozdělení dat podle $qC1D$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 32
3.10 Rozdělení dat podle $pKa$: na dva shluky (vlevo), na tři shluky (vpravo) 32
3.11 Graf smíšené hustoty: rozdělení dat podle $qH$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 33
3.12 Grafy smíšené hustoty: rozdělení dat podle $qO$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 34
3.13 Grafy smíšené hustoty: rozdělení dat podle $qC1$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 34
3.14 Grafy smíšené hustoty: rozdělení dat podle $qOD$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 35
3.15 Grafy smíšené hustoty: rozdělení dat podle $qC1D$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 35
3.16 Grafy smíšené hustoty: rozdělení dat podle $pKa$: na dva shluky (vlevo), na tři shluky (vpravo) . . . . . 35

3.17 Graf smíšené hustoty: rozdělení dat podle $qOD$: na dva shluky . . . 36
3.18 Grafy závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ pro modely: a) mod.3d (vlevo), b) g3.qH.3d (vpravo) . . . . . 42
3.19 Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ 45
3.20 Grafy k analýze reziduí 5d.modelu . . . . . 46
3.21 Graf závislosti odhadu střední hodnoty $pKa$ na experimentální $pKa$ pro modely: a) mod.5d (vlevo), b) g3.qH.5d (vpravo) . . . . . 51
3.22 Křivka vykreslená příkazem `plot(qH.Mclust,what="density")` . 59
3.23 Graf smíšené hustoty pro rozdělení dat podle $qH$ na tři shluky . . . 60

# Příloha E - Seznam tabulek

# Seznam tabulek


3.1 Tabulka Cookových vzdáleností . . . . . 25
3.2 Hodnoty různých typů reziduí . . . . . 26
3.3 Srovnání koeficientů determinace pro modely g3.qH.3d a g3.qO.3d 39
3.4 Srovnání koeficientů determinace pro modely typu g2.3d . . . . . 40
3.5 Srovnání modelů g3.qH.3d a g2.qC1D.3d . . . . . 41
3.6 Srovnání kritérií kvality modelů mod.3d a g3.qH.3d I . . . . . 43
3.7 Srovnání kritérií kvality modelů mod.3d a g3.qH.3d II . . . . . 44
3.8 Srovnání koeficientů determinace pro modely g3.qH.5d a g3.qO.5d 49
3.9 Srovnání koeficientů determinace pro modely typu g2.5d . . . . . 50
3.10 Srovnání modelů g3.qH.5d a g2.qOD.5d . . . . . 50
3.11 Srovnání kritérií kvality modelů mod.5d a g3.qH.5d I . . . . . 52
3.12 Srovnání kritérií kvality modelů mod.5d a g3.qH.5d II . . . . . 53
3.13 Tabulka intervalů pro rozdělení dat do shluků podle náboje $qH$ . . 53

# Seznam použité literatury


[1] J. ANDĚL, *Statistické metody*, 2. přeprac.vydání, Praha: MATFYZPRESS, 1998, 274 s., ISBN 80-85863-27-8.

[2] J. ANDĚL, *Základy matematické statistiky*, 1. vydání., Praha: MATFYZPRESS, 2005, 358 s., ISBN 80-86732-40-1.

[3] J. A. BILMES, *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, (elektronicky), dostupné na: http://lasa.epfl.ch/teaching/lectures/ML_Phd/Notes/GP-GMM.pdf.

[4] T. BOUCHAL, *Predikce pKa s využitím QSPR modelů: Bakalářská práce*, Brno: Masarykova univerzita, Fakulta přírodovědecká, 2012, 35 s., Vedoucí bakalářské práce: Radka Svobodová Vařeková, (elektronicky), dostupné na: https://is.muni.cz/auth/th/357534/prif_b/bc_thesis.pdf.

[5] M. J. CRAWLEY, *The R Book*, 1. vydání, Chichester: John Wiley & Sons Ltd, 2007, 942 s., ISBN 978-0-470-51024-7, (elektronicky), dostupné na: www.montana.edu/dchristianson/teaching/conservation_biology/crawley2007.pdf.

[6] M. FORBELSKÁ, *Přednáškový text k předmětu Lineární statistické metody I.*

[7] M. FORBELSKÁ, *Přednáškový text k předmětu Lineární statistické metody II.*

[8] T. HANZLÍČEK, *Odhady parametrů smíšených distribučních funkcí náhodného vektoru: Diplomová práce*, Olomouc: Univerzita Palackého v Olomouci, Fakulta přírodovědecká, 2010, 69 s., Vedoucí diplomové práce: Lubomír Kubáček, (elektronicky), dostupné na: http://theses.cz/id/8r5zl0/108934-529171705.pdf.

[9] P. HEBÁK a J. HUSTOPECKÝ, *Vícerozměrné statistické metody 1*, 2. přeprac.vydání, Praha: INFORMATORIUM, 2007, 253 s., ISBN 978-80-7333-056-9.

[10] P. HEBÁK a J. HUSTOPECKÝ, *Vícerozměrné statistické metody 2*, 1. vydání, Praha: INFORMATORIUM, 2005, 239 s., ISBN 80-7333-036-9.

[11] A. KOMORA, *Aplikace EM algoritmu: Bakalářská práce*, Praha: Univerzita Karlova, Matematicko-fyzikální fakulta, 2011, 34 s., Vedoucí bakalářské práce: Marek Omelka, (elektronicky), dostupné na: https://is.cuni.cz/webapps/zzp/detail/90852/?lang=cs.

[12] K. KONEČNÁ, *Výuka jazyka R: Bakalářská práce*, Brno: Masarykova univerzita, Fakulta přírodovědecká, 2010, 50 s., Vedoucí bakalářské práce: Jan Koláček, (elektronicky), dostupné na:
https://is.muni.cz/auth/th/270073/prif_b/Bakalarska_prace.pdf.

[13] M. MELOUN a J. MILITKÝ, *Interaktivní statistická analýza dat*, 3. vydání, Praha: Karolinum, 2012, 953 s., ISBN 978-80-246-2173-9.

[14] S. PEKÁR a M. BRABEC, *Moderní analýza biologických dat 1*, 1. vydání, Praha: Scientia, 2009, 225 s., ISBN 978-80-86960-44-9.

[15] S. PEKÁR a M. BRABEC, *Moderní analýza biologických dat 2*, 1. vydání, Brno: Masarykova univerzita, 2012, 256 s., ISBN 978-80-210-5812-5.

[16] P. PÍCHOVÁ, *Vážená metoda nejmenších čtverců: Diplomová práce*, Brno: Masarykova univerzita, Fakulta přírodovědecká, 1995, 37 s., Vedoucí diplomové práce: Jaroslav Michálek.

[17] H. ŘEZANKOVÁ, *Shluková analýza dat*, 1. vydání, Praha: Professional Publishing, 2007, 196 s., ISBN 978-80-86946-26-9.

[18] J. VANĚK, *Parametrizace equalizačních metod: Diplomová práce*, Brno: Masarykova univerzita, Fakulta přírodovědecká, 2009, 57 s., Vedoucí diplomové práce: Radka Svobodová Vařeková, (elektronicky), dostupné na:
http://is.muni.cz/th/98536/fi_m/dp_jv.pdf.

[19] R. SVOBODOVÁ VAŘEKOVÁ, S. GEIDL, C.-M. IONESCU, O. SKŘEHOTA, M. KUDERA, D. SEHNAL, T. BOUCHAL, R. ABAGYAN, H. J. HUBER, and J. KOČA, Predicting pKa Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes, *Journal of Chemical Information and Modeling*[online], 2011, roč. 51, č. 8 [cit. 17. července 2011], ISSN 1549-9596, (elektronicky), dostupné na: http://pubs.acs.org/doi/abs/10.1021/ci200133w.

[20] R. SVOBODOVÁ VAŘEKOVÁ, S. GEIDL, C.-M. IONESCU, O. SKŘEHOTA, M. KUDERA, D. SEHNAL, T. BOUCHAL, R. ABAGYAN, H. J. HUBER, and J. KOČA, Predicting pKa Values of Substituted Phenols from Atomic Charges, *Journal of Chemoinformatics*[online], 2013, roč. 18, č. 5 [cit. 10. dubna 2013], ISSN 1758-2946, (elektronicky), dostupné na:
http://www.jcheminf.com/content/pdf/1758-2946-5-18.pdf.