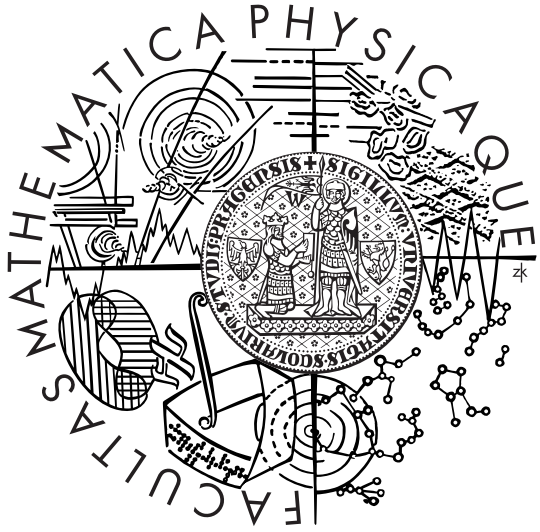
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jiřina Münsterová

Modely se smíšenými efekty pro toxikokinetická data

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: *Ing. Marek Brabec, PhD.*

Studijní program: *Matematická statistika*

Tímto bych chtěla poděkovat vedoucímu mé diplomové práce Ing. Marku Brabcovi, PhD. za věnovaný čas a RNDr. Jaroslavu Mrázovi, CSc. za poskytnutá data a cenné informace k pozadí problému.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 16.4.2008 Jiřina Münsterová

Obsah

1 Úvod **4**
- 1.1 Motivace . . . 4
- 1.2 Data . . . 4
- 1.3 Vážené průměry v datech . . . 5

2 Kinetika dimethylformamidu **6**
- 2.1 Dimethylformamid a jeho vlastnosti . . . 6
- 2.2 Rozklad dimethylformamidu . . . 6

3 Modely se smíšenými efekty **8**
- 3.1 Formulace lineárního modelu . . . 8
- 3.2 Odhady parametrů LME modelu . . . 9
  - 3.2.1 Odhady metodou maximální věrohodnosti . . . 9
  - 3.2.2 Odhady metodou REML . . . 10

4 Konvoluce **12**

5 Analýza reálných pokusných dat **13**
- 5.1 Regresní modely pro jednotlivé subjekty . . . 14
  - 5.1.1 Polynomické modely . . . 15
  - 5.1.2 Modely s goniometrickými funkcemi . . . 18
  - 5.1.3 Srovnání různých modelů . . . 19
- 5.2 Regresní modely a konvoluce . . . 20
  - 5.2.1 Konvoluce v polynomických modelech . . . 20
  - 5.2.2 Konvoluce v modelech s goniometrickými funkcemi . . . 23
- 5.3 Modely se smíšenými efekty pro reálná data . . . 26
  - 5.3.1 Funkce `lme` v `R` . . . 26
  - 5.3.2 LME model pro DMF . . . 27
  - 5.3.3 LME model pro AMCC a konvoluce . . . 30
  - 5.3.4 Upravený LME model pro AMCC . . . 33

6 Závěr **37**

Název práce: Modely se smíšenými efekty pro toxikokinetická data
Autor: Jiřina Münsterová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Ing. Marek Brabec, PhD.
e-mail vedoucího: mbrabec@cs.cas.cz, mbrabec@szu.cz
Abstrakt: Opakovaná měření prováděná na každém objektu ze skupiny zkoumaných jedinců vedou k přítomnosti korelace. Získaná data tudíž nejsou zcela nezávislá. Jedná se tedy o dva typy variability, jednak variabilita mezi jedinci a na druhé straně variabilita mezi měřeními jednoho daného jedince. Pro taková data je vhodné použití modelů se smíšenými efekty. Práce se zaměřuje především na situaci, kdy v důsledku přeměny toxické látky v lidském organismu, neodpovídají měření hodnotám, jež jsou předmětem zájmu. Za pomoci procesu, který lze pozorovat, se snažíme odhadnout parametry modelu procesu, který ve skutečnosti probíhá v lidském organismu. Hodnoty, které máme k dispozici, odpovídají váženým průměrům skutečných měření. Uvedená problematika je diskutována na reálných datech a k jejich zpracování byl použit program R.
Klíčová slova: modely se smíšenými efekty, konvoluce, toxikokinetika

Title: Mixed-effects models for toxic-kinetical data
Author: Jiřina Münsterová
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Marek Brabec, PhD.
Supervisor's e-mail address: mbrabec@cs.cas.cz, mbrabec@szu.cz
Abstract: Multiple measurements performed at each individual of a related group lead to a presence of correlation. Thus the data aren't fully independent. Therefore we have two types of variability, partly variability among the individuals and on the other hand variability among measurements performed at given subject. For such data fits the use of mixed-effect models. This graduation thesis focuses escpecially on situation when measurements do not respond with values that we consider as a object of interest due to a toxic substance transformation in human organism. With the use of process that can be observed we try to estimate the parameters of a model of process that actually runs in human organism. Processed values correspond to weighted average of a real measurements. Presented problem is discussed with the use of real data and for their processing was used the R software.
Keywords: mixed-effects models, convolution, toxicokinetics

Kapitola 1

Úvod

1.1 Motivace

Ve farmakokinetice, biologii, biomedicíně a dalších oblastech se často vyskytují data sestávající z opakovaných měření daného jevu na každém ze skupiny, či více skupin, objektů. Tímto objektem mohou být lidé, zvířata, stroje, apod. Například v longitudinálních klinických studiích se provádějí měření na řadě jedinců opakovaně po určitém čase. Přítomnost opakovaných pozorování vede k rozlišení dvou typů variability. Jedná se o variabilitu mezi objekty a variabilitu mezi jednotlivými měřeními objektu samotného. Pro taková, v praxi obvykle nevyvážená, data je vhodné použití modelu se smíšenými efekty.

1.2 Data

Datový soubor, který budeme zpracovávat, laskavě poskytl RNDr. Jaroslav Mráz, CSc. Data v tomto souboru pocházejí z reálného prostředí. Jejich shromažďování probíhalo v roce 2003 ve čtyřech Italských továrnách na zpracování koženky. Říkejme jim Italská data. Zaměstnanci těchto továren jsou během výkonu své práce vystaveni působení toxické látky dimethylformamid (chemický vzorec - $C_3H_7N_1O_1$). Charakter a vlastnosti dimethylformamidu popíšeme v Kapitole 2.

Italská data obsahují hodnoty koncentrací dimethylformamidu (DMF), který zaměstnanci při práci vdechují, a látky acetyl-methyl-carbamoylcystein vznikající přeměnou DMF. Data byla získávána opakovaně od čtyřiceti pracovníků různých profesí ze čtyř továren. Rozpětí sledovaného období je 85 dní. Jak už je u takovýchto sběrů dat obvyklé, nejsou měření kompletní, například proto, že zaměstnanec z jakýchkoliv důvodů nepřišel do práce. Počty jedinců v jednotlivých továrnách s odpovídajícím počtem pozorování pro obě sledované látky jsou uvedeny v Tabulce 1.1. Vidíme, že zpravidla máme k dispozici šest hodnot pro každého jednotlivce.

Data zpracujeme v programu `R`, který je volně k dispozici na internetových stránkách [1], s využitím knihovny `nlme`, kterou speciálně pro modely se smíšenými efekty vytvořili Jose Pinheiro a Douglas Bates a popsali ji v knize [3].

	DMF					AMCC				
	počet měření					počet měření				
	3	4	5	6	$\sum$	3	4	5	6	$\sum$
továrna I	1	0	1	6	44	1	0	1	6	44
továrna III	2	0	2	8	64	2	0	4	6	62
továrna IV	0	0	2	8	58	0	1	1	8	57
továrna V	0	1	1	8	57	0	1	1	8	57
$\sum$					223					220

Tabulka 1.1: Počty subjektů v jednotlivých továrnách se 3, 4, 5 či 6-ti pozorováními.

1.3 Vážené průměry v datech

Tato práce se zaměřuje především na zkoumání situace, kdy pozorování neodpovídají přímo hodnotě, která je předmětem zájmu. Známe-li však zákonitosti, které platí mezi pozorovanou a skutečnou hodnotou, lze vhodnou korekcí dat z daných hodnot získat charakteristiky pro skutečná data. V následujících kapitolách se pokusíme tyto korekce odvodit. Naše situace je taková, že hodnoty v datech neudávají jedno konkrétní měření, ale vážené průměry koncentrací za posledních několik dní. Váhy máme k dispozici a jsou známé z externích studií. Tyto váhy zohledňují přírůstky předchozích dnů.

Máme-li váhy $w_0, w_1, \ldots, w_p$, pak jednu konkrétní hodnotu y_t z dat, můžeme zapsat ve tvaru

$$y_t = \sum_{i=0}^{p} w_i z_{t-i} + \epsilon_t, \;\; \epsilon_t \sim N(0, \sigma^2) \tag{1.1}$$

kde t je časový okamžik, v našem případě je to den. Váha w_0 odpovídá přírůstku aktuálního dne, tj. dne t. A váhy $w_1, \ldots, w_p$ odpovídají přírůstkům předchozích p dnů $t-1, \ldots, t-p$. Hodnoty $z_t, z_{t-1}, \ldots, z_{t-p}$ odpovídají skutečné hodnotě koncentrace sledované látky, jejíž průběh nás zajímá.

Kapitola 2

Kinetika dimethylformamidu

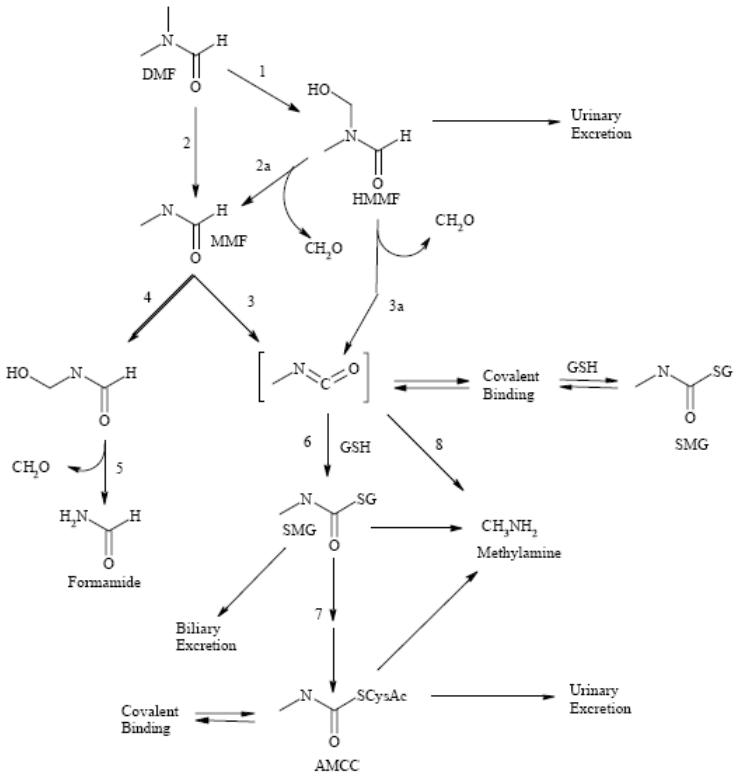
Kinetika je vědní obor zabývající se rychlostí a mechanismem chemických reakcí za rozličných podmínek. V této práci představují toxikokinetická data látku, která v průběhu času prochází lidským organismem a rozkládá se v něm.

2.1 Dimethylformamid a jeho vlastnosti

Dimethylformamid (DMF), v průmyslu používán jako rozpouštědlo, je bezbarvá až lehce nažloutlá kapalina, jemného čpavkového zápachu. Způsobuje zejména hepatotoxicitu, dráždí sliznice dýchacích cest, spojivky a kůži. Do organismu se vstřebává plícemi, zažívacím traktem i pokožkou. V případě styku s pokožkou hraje důležitou roli teplota okolního prostředí, neboť s potem se lépe vstřebává. Do krve se však dostává rychleji vdechováním výparů a především se lze vystavení tomuto působení nejobtížněji vyhnout. Představu o množství vdechnuté sledované látky nám poskytnul přístroj, který sledovaní zaměstnanci nosili po celou dobu směny u sebe, a který měřil koncentraci DMF ve vzduchu.

2.2 Rozklad dimethylformamidu

DMF však v organismu nezůstává ve své původní podobě, ale přeměňuje se a rozpadá na mnoho látek a vedlejších produktů. Všimněme si složitosti tohoto procesu na Obrázku 2.1, který na svých internetových stránkách uvádí U.S. Environmental Protection Agency [9]. Jednou z látek vznikajících přeměnou dimethylformamidu je mimo jiné acetyl-methyl-carbamoylcystein (AMCC). Na AMCC se dimethylformamid rozpadá po několika mezikrocích. Při jednorázové expozici je AMCC v malém množství zjistitelný až na konci směny a dosahuje maxima až na druhý den. Jeho poločas rozpadu je přibližně 24 hodin. Při pravidelných směnách, nebo-li při dlouhodobém vystavení působení DMF, koncentrace AMCC roste, až se ustálí na nějaké hladině. Je tedy třeba brát v úvahu historii expozice. Zde přicházejí na řadu dané váhy, jak již bylo řečeno, známé z externích studií. Díky nim získáme představu o přírůstcích koncentrace AMCC v organismu a okamžité hodnotě koncentrace v daném okamžiku, jejíž průběh nás zajímá. Koncentrace AMCC je zjistitelná ze vzorků moči a je



Obrázek 2.1: Rozklad dimethylformamidu.

tedy evidentní, že na rozdíl od DMF pochází hodnota AMCC ze všech způsobů absorpce, ne jen z vdechování.

Musíme si ještě uvědomit, že koncentrace látky v moči je značně ovlivněna tím, kolik zaměstnanec vypije během dne tekutin. Před samotnou analýzou bychom měli uvést data do stavu, aby se vůbec dala porovnávat. Z tohoto důvodu byla měřena v organismu ještě jedna látka, kreatinin. V Lékařském slovníku [7] se uvádí, velmi zjednodušeně, že kreatinin je látka vznikající ve svalech a její koncentrace v krvi odráží funkci ledvin. Stejně jako koncentrace sledované látky, je i meření koncentrace kreatininu zatíženo chybou. Mohli bychom se ptát, zda by nebylo lepší předpokládat, že všichni pracovníci vypijí přibližně stejné množství tekutin. Avšak koncentrace kreatininu se mezi pracovníky značně liší a domníváme se tedy, že biologická stránka věci převáží nad statistickou chybou. Od této chvíle, kdykoliv budeme mluvit o AMCC, budeme mít na mysli AMCC upravené podle vzorce

$$\text{AMCC} = \frac{\text{původní AMCC}}{\text{kreatinin}}. \tag{2.1}$$

Kapitola 3

Modely se smíšenými efekty

Klasické regresní modely mají všechny parametry pevné, což znamená, že jsou shodné pro všechny sledované objekty. Tyto parametry jsou obvykle označovány vektorem $\boldsymbol{\beta}$. Modely se smíšenými efekty jsou oproti klasickým modelům rozšířeny o parametry s náhodnými efekty, které jsou označovány $\boldsymbol{b}_i$. Parametry modelu se smíšenými efekty pak můžeme zapsat jako

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i. \tag{3.1}$$

V kontextu modelů se smíšenými efekty se parametry s pevnými efekty (zkráceně pevné efekty) vztahují k celé sledované populaci a parametry s náhodnými efekty (náhodné efekty) se vztahují k náhodně vybrané jednotce z této populace. Jednotkou může být například skupina objektů se shodnými znaky, nebo jediný objekt, na kterém jsou prováděna opakovaná měření. Pomocí modelů se smíšenými efekty lze popsat náhodnou variabilitu i na jiných úrovních než mezi jednotlivými pozorováními. Můžeme tak analyzovat data, ve kterých nejsou jednotlivá pozorování úplně nezávislá, jako např. zmíněná opakovaná měření na stejném jedinci. Tyto modely nabízejí i další rozšíření, jakým je časová korelace. V následujících kapitolách si uvedeme modely se smíšenými efekty jak je popisují autoři [3]. Značení přizpůsobíme typu dat, která budeme zpracovávat v Kapitole 5.3.

3.1 Formulace lineárního modelu

Lineární modely se smíšenými efekty (LME) jsou modely, ve kterých se pevné i náhodné efekty vyskytují v lineárním tvaru.

Mějme pro každého z M jedinců sadu dat. Pro i-tého jedince mějme n_i opakovaných měření, tj. n_i-rozměrný vektor odezvy $\boldsymbol{y}_i$. Celkem tedy máme $\sum_{i=1}^{M} n_i$ pozorování. Vektor $\boldsymbol{y}_i$ je popsán lineárním modelem se smíšenými efekty

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,\ i = 1, ..., M, \\ & \\ & \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi}),\ \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_i) \end{aligned} \tag{3.2}$$

kde $\boldsymbol{\beta}$ je p-rozměrný vektor pevných efektů, $\boldsymbol{b}_i$ je q-rozměrný vektor náhodných efektů, $\boldsymbol{X}_i$ $(n_i \times p)$ a $\boldsymbol{Z}_i$ $(n_i \times q)$ jsou známé regresní matice pevných a náhodných efektů a $\boldsymbol{\epsilon}_i$ je n_i-rozměrný chybový vektor i-té skupiny. Chyba $\boldsymbol{\epsilon}_i$ má normální rozdělení s nulovou střední hodnotou a rozptyl $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2\boldsymbol{\Lambda}_i$.

Náhodné efekty $\boldsymbol{b}_i$ a chyby $\boldsymbol{\epsilon}_i$ předpokládáme nezávislé pro každé dva jedince a navzájem pro jednoho daného jedince. Rozdělení náhodných efektů $\boldsymbol{b}_i$ rovněž předpokládáme normální s nulovou střední hodnotou a tudíž je plně charakterizováno varianční-kovarianční maticí $\boldsymbol{\Psi}$. Tato matice tedy musí být přinejmenším symetrická a pozitivně semidefinitní. Prvky matic $\boldsymbol{\Psi}$ a $\boldsymbol{\Lambda}_i$ shrňme do vektoru, který označíme $\boldsymbol{\theta}$.

Sloupce matice $\boldsymbol{Z}_i$ jsou obvykle podmnožinou sloupců matice $\boldsymbol{X}_i$. V našem případě tomu tak skutečně je.

3.2 Odhady parametrů LME modelu

Zaměříme se na dvě obecné metody pro odhadování parametrů, metodu maximální věrohodnosti (ML, maximum likelihood) a z ní odvozenou REML (restricted maximum likelihood).

3.2.1 Odhady metodou maximální věrohodnosti

Věrohodnostní funkce je tvaru

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= p\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \\ &= \prod_{i=1}^{M} p\left(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \end{aligned} \tag{3.3}$$

kde L je věrohodnost, p je hustota a $\boldsymbol{y}$ je celkový N-rozměrný vektor pozorování, $N = \sum_{i=1}^{M} n_i$.

Uvažujme hustotu normálního rozdělení s nulovou střední hodnotou a varianční-kovarianční maticí $\boldsymbol{\Sigma}_i$. Model (3.2) lze přepsat na

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^*,\ i = 1, ..., M, \tag{3.4}$$

kde $\boldsymbol{\epsilon}_i^* = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$. Vidíme, že $\boldsymbol{\epsilon}_i^*$ jsou součty dvou nezávislých mnohorozměrně normálně rozdělených náhodných vektorů. A $\boldsymbol{\epsilon}_i^*$ jsou tudíž také nezávislé mnohorozměrně normálně rozdělené s nulovou střední hodnotou a varianční-kovarianční maticí $\sigma^2\boldsymbol{\Sigma}_i$, kde $\boldsymbol{\Sigma}_i = \boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T/\sigma^2$. Pak z (3.4) plyne, že $\boldsymbol{y}_i$ jsou

nezávislé mnohorozměrně normální náhodné vektory se střední hodnotou $\boldsymbol{X}_i\boldsymbol{\beta}$ a varianční maticí $\sigma^2\boldsymbol{\Sigma}_i$. Hustota $\boldsymbol{y}_i$ je tedy

$$p\ (\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2) = (2\pi\sigma^2)^{-\frac{n_i}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\exp\Big(\frac{(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})}{-2\sigma^2}\Big), \quad (3.5)$$

kde $|\boldsymbol{\Sigma}_i|$ je determinant matice $\boldsymbol{\Sigma}_i$. Věrohodnostní funkce (3.3) je pak

$$\begin{aligned} &L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = \\ &= \prod_{i=1}^{M}(2\pi\sigma^2)^{-\frac{n_i}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\exp\Big(\frac{(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})}{-2\sigma^2}\Big). \end{aligned} \quad (3.6)$$

Maximálně věrohodné odhady získáme z logaritmické věrohodnostní funce

$$\begin{aligned} &\ell(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = \log L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) \\ &= \sum_{i=1}^{M}\Big\{\Big(\frac{(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})}{-2\sigma^2}\Big)\log\Big((2\pi\sigma^2)^{-\frac{n_i}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\Big)\Big\}. \end{aligned} \quad (3.7)$$

Položíme

$$\begin{aligned} \frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y})}{\partial\boldsymbol{\beta}} &= 0 \\ \frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y})}{\partial\sigma^2} &= 0 \end{aligned}$$

a pro dané hodnoty $\boldsymbol{\theta}$ jsou maximálně věrohodné odhady $\boldsymbol{\beta}$ a σ^2

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\sum_{i=1}^{M}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)^{-1}\sum_{i=1}^{M}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i, \quad (3.8)$$

$$\hat{\sigma}^2(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{M}(\boldsymbol{y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))}{N}. \quad (3.9)$$

3.2.2 Odhady metodou REML

Metoda maximální věrohodnosti má tendenci podhodnocovat odhady komponent rozptylu a podceňovat velikosti náhodných efektů. Proto mnoho analytiků preferuje metodu REML, která tuto vlastnost nemá. I funkce v knihovně `nlme` standardně uvádějí odhady metodou REML.

Věrohodnostní funkce má v tomto případě tvar

$$L_R(\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta}. \quad (3.10)$$

Pro pevné efekty $\boldsymbol{\beta}$ předpokládáme, na základě Bayesovského přístupu, lokálně rovnoměrné apriorní rozdělení, a integrujeme je mimo věrohodnost.

Logaritmická věrohodnostní funkce v případě REML je pak

$$\ell_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) \quad = \quad \log L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}). \tag{3.11}$$

Kapitola 4

Konvoluce

O konvoluci hovoříme, když nová náhodná veličina vzniká jako součet nezávislých náhodných veličin. Někdy má součet rozdělení stejného typu jako mají jednotlivé sčítance, pouze s jinými parametry. Jedná se úlohu určit rozdělení součtu dvou či více náhodných veličin. Uveďme si tvrzení podle [2], která se přitom používají.

Věta 4.1 *Nechť veličiny X a Y jsou nezávislé a nechť mají po řadě distribuční funkce F_1 a F_2. Pak veličina $Z = X + Y$ má distribuční funkci*

$$F(z) = \int F_2(z - x) dF_1(x). \tag{4.1}$$

Funkci F, danou vzorcem (4.1), nazýváme konvolucí distribučních funkcí F_1 a F_2 a značíme ji symbolem $F = F_1 * F_2$. Platí $F_1 * F_2 = F_2 * F_1$.

Věta 4.2 *Nechť X a Y jsou nezávislé náhodné veličiny, které nabývají pouze celočíselných hodnot. Pro každé celé k označme $p_k = P(X = k)$, $q_k = P(Y = k)$. Pak veličina $Z = X + Y$ nabývá rovněž celočíselných hodnot a platí*

$$P(Z = k) = \sum_{j=-\infty}^{+\infty} p_j q_{k-j}. \tag{4.2}$$

Vztah (4.2) je obdobou vztahu (1.1). V našem případě nejde o konvoluci pravděpodobnostních rozdělení, ale Větu 4.2 lze na naši situaci aplikovat.

Kapitola 5

Analýza reálných pokusných dat

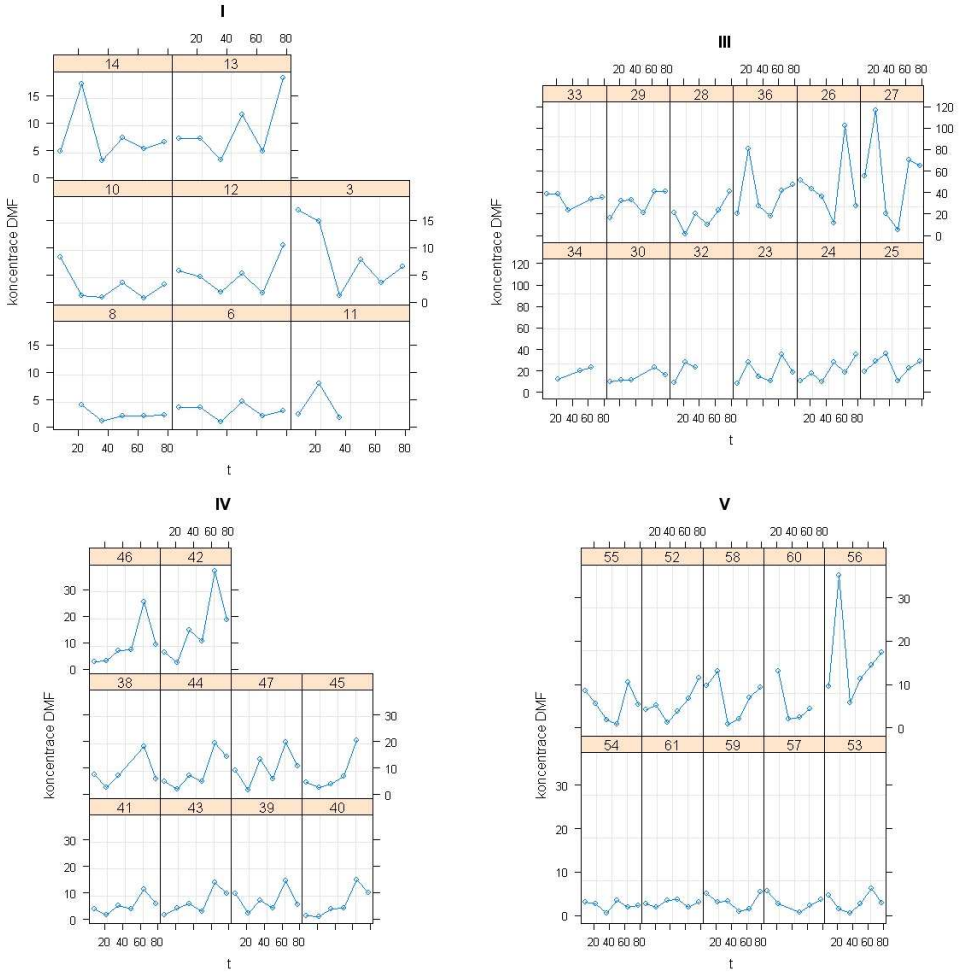
V této kapitole přistoupíme k samotnému zpracování Italských dat. K tomu využijeme program `R` a knihovnu `nlme`, jejímiž autory jsou Jose Pinheiro a Douglas Bates. Než vybudujeme smíšený model, zpracujeme data nejprve na úrovni jednotlivých subjektů. V kapitole 5.1 budeme modelovat koncentraci DMF. Zde pozorovaná hodnota odpovídá skutečné hodnotě, na rozdíl od naměřených koncentrací AMCC, na které se zaměříme v kapitole 5.2, kde data budeme modelovat se zohledněním konvolučních vah. V obou případech budeme data modelovat dvěma způsoby, jednak na základě polynomů, § 5.1.1 resp. § 5.2.1, a vyzkoušíme také, zda nebude vhodnější použít jiný model, např. s goniometrickými funkcemi, § 5.1.2 resp. § 5.2.2. V kapitole 5.3 vypracujeme smíšené modely opět pro DMF a AMCC zvlášť.

	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7
DMF	1	0	0	0	0	0	0	0
AMCC	1	0.4867	0.2369	0.1153	0.0561	0.0273	0.0133	0.0065

Tabulka 5.1: Konvoluční váhy známé z externích experimentálních studií.

Jak v části, kde zpracováváme koncentraci AMCC pro jednotlivé subjekty zvlášť, tak v části, kde budujeme smíšený model, budeme pracovat s konvolučními vahami. První váha $w_0 = 1$ odpovídá příspěvku dne, pro který máme měření uvedeno, w_1 odpovídá příspěvku dne předchozího, atd. Příspěvky od osmého předchozího dne dál jsou již zanedbatelné, formálně tedy platí $w_i = 0$ pro $i > 7$. K dispozici máme tedy osm vah pro přírůstky aktuálního a sedmi předcházejících dnů. Velikosti těchto vah, které laskavě poskytl RNDr. Jaroslav Mráz, CSc. [8], jsou uvedeny v Tabulce 5.1. Měření DMF není nijak ovlivněno předchozími dny a kromě w_0 jsou všechny ostatní váhy nulové. Uvedené koncentrace AMCC pak můžeme zapsat ve tvaru $y_t = \sum_{i=0}^{7} w_i z_{t-i}$, kde $z_t, \ldots, z_{t-7}$ jsou skutečné hodnoty koncentrace, které nás zajímají.

Uveďme ještě malou poznámku k datům. U jednoho ze sledovaných subjektů, konkrétně se jedná o subjekt s číslem 29, je uvedena koncentrace kreatininu, která je v porovnání s ostatními hodnotami velmi nepravděpodobná. Konkrétně je to 0.04, zatímco ostatní se v průměru pohybují kolem hodnoty 1.2. Odpovídající hodnota AMCC je pak 2520, což je pětkrát více než druhá



Obrázek 5.1: Koncentrace DMF čtyřiceti subjektů ze čtyř různých továren.

nejvyšší hodnota. Proto bylo toto měření vynecháno. Z Tabulky 5.2 si můžeme udělat přibližnou představu o modelovaných koncentracích DMF a AMCC. Z rozdílu mezi mediánem a průměrem obou sledovaných látek je vidět, že data nejsou příliš vyvážená.

	Minimum	1.kvartil	Medián	Průměr	3.kvartil	Maximum
DMF	0.600	3.155	7.070	13.080	17.420	116.400
AMMC	0.5217	6.5000	16.8600	40.4100	42.7800	560.9000

Tabulka 5.2: Shrnutí koncentrací DMF a AMCC.

5.1 Regresní modely pro jednotlivé subjekty

V této části je zkoumána koncentrace látky DMF. Všechna měření jsou vynesena v Obrázku 5.1. Panely představují jednotlivé továrny a jsou dále

	Subjekt č.24				Subjekt č.42	
k	SR(k)	HQ(k)	A(k)	k	SR(k)	HQ(k)	A(k)
0	4.926901	4.822673	173.8533	0	5.300438	5.196211	252.5854
1	4.498182	4.289728	118.5601	1	5.135436	4.926982	224.2303
2	4.982541	4.669859	184.3668	2	5.713842	5.401160	383.0740
3	5.634720	5.217811	321.8158	3	5.951901	5.534992	441.9337
4	6.572689	6.051553	722.2196	4	6.597806	6.076670	740.5895

Tabulka 5.3: Statistiky $SR(k), HQ(k)$ a $A(k)$ pro subjekty č.24 a č.42, kde k je stupeň polynomu. ($c = 1$, $\alpha = 0.2$)

rozděleny na části představující subjekty z dané továrny. Tyto části jsou označeny čísly konkrétního subjektu. Jelikož ale budeme zpracovávat jednotlivé subjekty zvlášť, není v tuto chvíli informace o továrnách podstatná. Měření byla prováděna v přibližně ekvidistantních časových intervalech po dobu 85-ti dní. Přesto, jak již bylo uvedeno v Tabulce 1.1, pro některé subjekty nemáme kompletní sadu pozorování.

Data budeme modelovat ze dvou různých hledisek, uveďme si, jak lze takové modely porovnávat. K porovnání různých modelů se často používá funkce založená na odhadu rozptylu zvětšeného o penalizaci počtu odhadovaných parametrů

$$AIC = -2\ell(\hat{\theta}) + 2(k+1), \tag{5.1}$$

kde ℓ je logaritmická věrohodnostní funkce a $k+1$ je počet parametrů modelu. Za vhodnější považujeme model s menší hodnotou AIC.

5.1.1 Polynomické modely

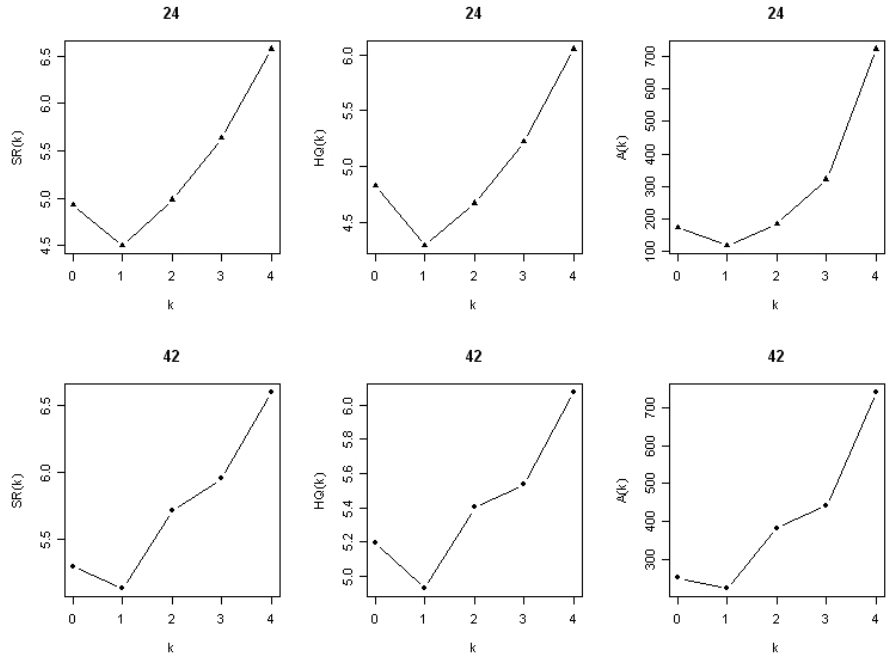
Popišme nejprve závislost koncentrace DMF na čase polynomem. K volbě stupně polynomu použijeme postup z přednášky Regrese prof. Zváry, [4] . Aby odhad stupně polynomu k byl konzistentní, je třeba penalizovat počet parametrů. Nechť pozorování y_i je v čase t_i popsáno polynomem

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + ... + \beta_k t_i^k + e_i, \tag{5.2}$$

kde $e_i \sim N(0, \sigma^2)$ a počet pozorování je $n > k + 1$.

Ke konzistentnímu odhadu vede minimalizace například těchto funkcí

$$\begin{aligned} SR(k) &= log S_k^2 + (k+1)\frac{log\ n}{n} \\ HQ(k) &= log S_k^2 + 2c(k+1)\frac{log\ log\ n}{n},\ c > 0 \\ A(k) &= S_k^2(1 + c(k+1)n^{-\alpha}),\ \alpha \in (0, 0.5),\ c > 0, \end{aligned} \tag{5.3}$$



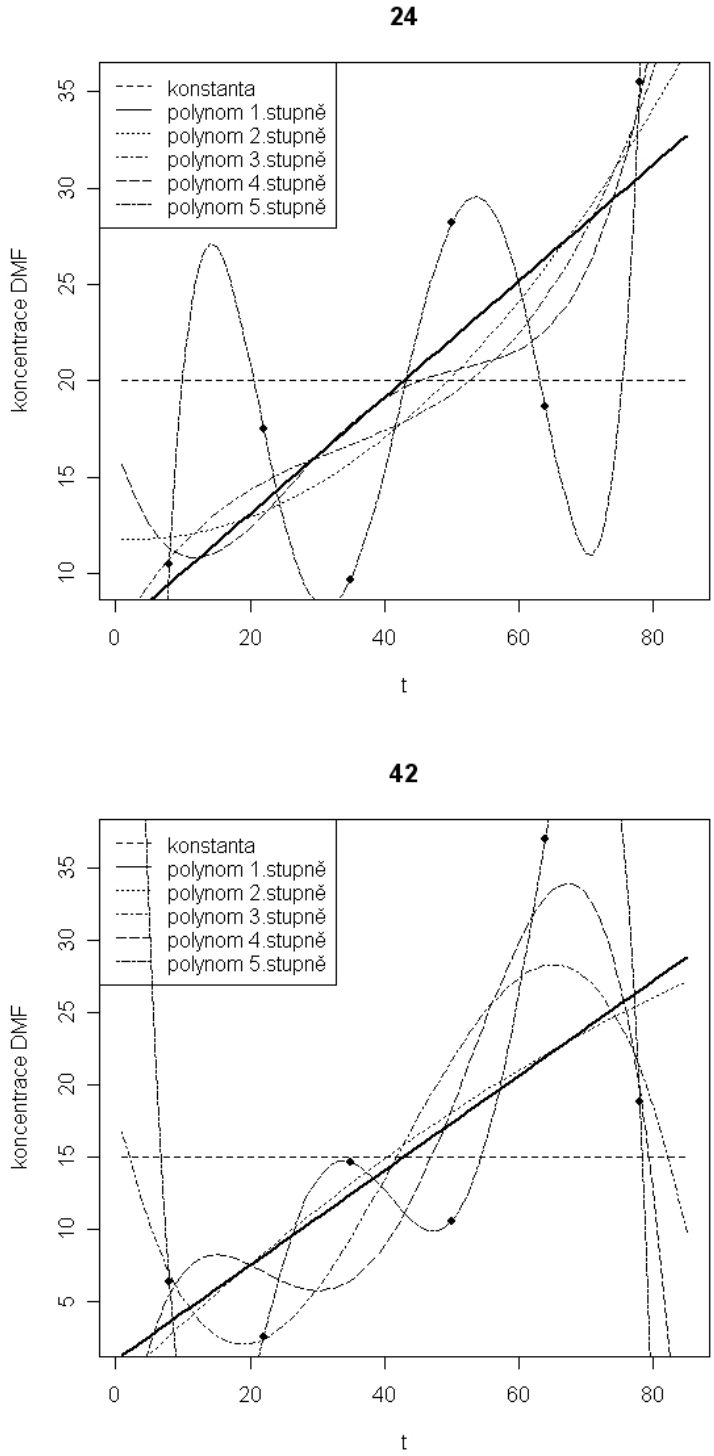
Obrázek 5.2: Statistiky $SR(k), HQ(k)$ a $A(k)$ pro subjekty č.24 a č.42, kde k je stupeň polynomu.

kde S_k^2 jsou reziduální rozptyly.

Od každého jednotlivce máme k dispozici $n \leq 6$ pozorování. Z toho plyne, že stupeň polynomu bude nejvýše čtvrtý (pokud neuvažujeme saturovaný model, pro který budou reziduální rozptyly nulové). Pro každý subjekt jsme spočítali všechny tři statistiky (5.3) s namátkovou volbou $c = 1$ a $\alpha = 0.2$. Při této volbě vycházejí stupně polynomu vypočtené pro všechny tři statistiky v téměř všech případech stejné. Pro ilustraci jsou vypočtené hodnoty statistik (5.3) pro dva náhodně vybrané subjekty uvedeny v Tabulce 5.3 a graficky na Obrázku 5.2. Na Obrázku 5.3 jsou pro ilustraci polynomy prvního až pátého stupně proloženy daty těchto dvou subjektů. Pro oba subjekty máme k dispozici všech šest pozorování. Z tabulky je zřejmé, že statistiky $SR(k), HQ(k)$ a $A(k)$ jsou pro tyto dva subjekty minimální pro stupeň polynomu $k = 1$. Ne vždy se jeví jako nejvhodnější přímka, ale průměrně je stupeň polynomu přibližně 1.4. Přidržíme se tedy volby $k = 1$ a na každou sadu dat použijeme lineární model. Pro každého jedince zapišme jeho j-té pozorování

$$y_j = \beta_0 + \beta_1 t_j + \epsilon_j, \ \epsilon_j \sim N(0, \sigma^2), \tag{5.4}$$

$j = 1, ..., n_i$, kde n_i je počet pozorování i-tého jedince.



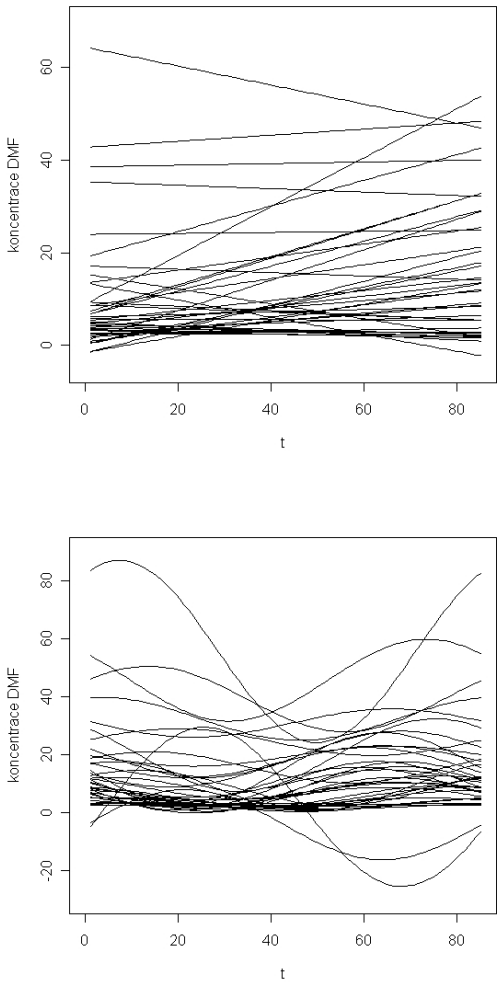
Obrázek 5.3: Polynomy stupně $k = 0, 1, ..., 5$ pro dva náhodně vybrané subjekty číslo 24 a 42.

5.1.2 Modely s goniometrickými funkcemi

V této části budeme zkoumat, zda při modelování koncentrace DMF nebude vhodnější použít goniometrické funkce, konkrétně $\sin(x)$ a $\cos(x)$. Pro lepší srovnání zvolme pouze jednu dvojici goniometrických funkcí, abychom neodhadovali o moc více parametrů než v lineárním modelu. Popišme j-té pozorování y_j jako

$$y_j = \beta_0 + \beta_1 \cos(ut_j) + \beta_2 \sin(ut_j) + e_j,\ e_j \sim N(0, \sigma^2), \tag{5.5}$$

$j = 1, \ldots, n_i$ a u je frekvence goniometrických funkcí. Z hlediska AIC se jeví jako nejvhodnější $u = 2k\pi/85$ a zároveň interval sledovaného období je právě 85 dní.



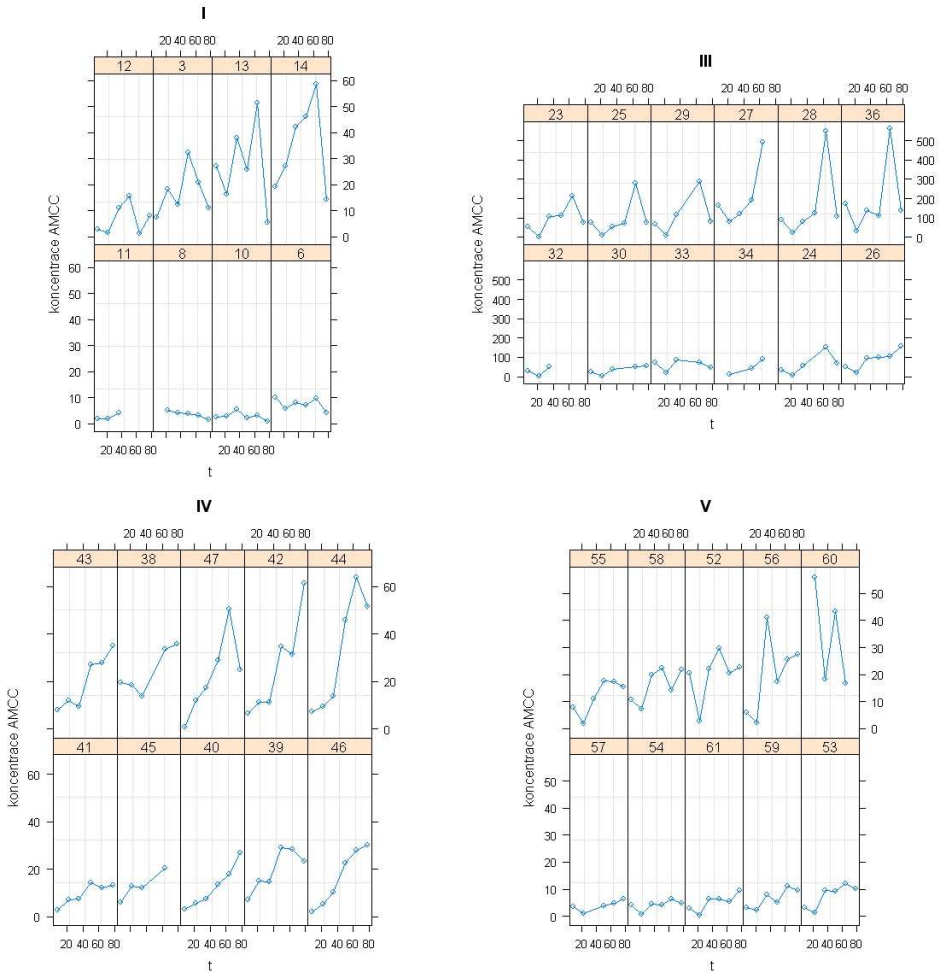
Obrázek 5.4: První graf zobrazuje lineární modely pro každý subjekt zvlášť, na druhém grafu jsou znázorněny modely s goniometrickými funkcemi, taktéž pro každý model zvlášť.

č. subjektu	lineární model	goniometrický model	č. subjektu	lineární model	goniometrický model
3	40.092673	40.65385	38	35.807549	32.74687
6	25.045758	27.03035	39	39.404137	39.44544
8	19.013507	18.85953	40	35.207585	35.46703
10	33.101813	34.14615	41	33.675623	32.85678
11	20.661274	$-\infty$	42	47.823569	47.57966
12	35.435581	34.89004	43	35.010787	38.35382
13	40.036284	41.94975	44	34.077860	32.10314
14	40.875568	41.94975	45	39.102853	39.36614
23	49.196459	51.57683	46	44.177871	43.42817
24	44.000046	48.70484	47	42.299134	42.97499
25	48.076587	48.97637	52	33.050174	31.24768
26	63.132616	64.33840	53	30.147646	27.13286
27	65.887905	65.21609	54	22.180792	23.85295
28	50.085657	50.23966	55	37.713679	35.74152
29	45.475889	50.82319	56	49.994070	50.98888
30	31.762318	26.73241	57	24.111188	15.18776
32	24.915575	$-\infty$	58	40.260728	35.74114
33	37.004410	33.74020	59	28.971878	23.45166
34	9.892296	$-\infty$	60	27.306573	22.29547
36	59.676097	60.75811	61	18.562003	19.80313

Tabulka 5.4: Hodnoty AIC pro jednotlivé modely, jak pro lineární tak pro modely s goniometrickými funkcemi.

5.1.3 Srovnání různých modelů

Na Obrázku 5.4 jsou pro ilustraci všechny modely zaneseny do grafu. Horní graf zobrazuje odhady z lineárního modelu (5.4) zkonstruovaného po jednotlivých subjektech. Dolní zobrazuje odhady z modelu (5.5), kdy je DMF modelováno pomocí goniometrických funkcí $\sin(x)$ a $\cos(x)$. V Tabulce 5.4 můžeme porovnat hodnoty AIC lineárního modelu a modelu s goniometrickými funkcemi pro každý subjekt zvlášť. U některých modelů s goniometrickými funkcemi nabývá hodnota AIC mínus nekonečna. To je způsobeno tím, že u těchto jedinců máme pouze tři pozorování na tři odhadované parametry, jedná se tedy o saturovaný model. Za vhodnější se považuje model s nižší hodnotou AIC. Podle Tabulky 5.4 vychází v 54% případů lépe lineární model. A jelikož je to zároveň model jednodušší, budeme z něj vycházet i při tvorbě modelu smíšeného.



Obrázek 5.5: Koncentrace AMCC 40-ti sledovaných subjektů ze čtyř různých továren.

5.2 Regresní modely a konvoluce

Nyní se podívejme na koncentraci látky AMCC, která vzniká rozkladem dimethylformamidu. Měření AMCC jsou uvedena na Obrázku 5.5. Zopakujme, že se jedná o koncentrace AMCC upravené kreatininem. Každý panel představuje měření od jednoho daného subjektu označeného číslem a shluky panelů opět představují čtyři sledované továrny. Všimněme si, že v továrně III jsou hodnoty koncentrací u některých subjektů výrazně vyšší než v ostatních třech továrnách. Stejně jako při modelování DMF, budeme i nyní zkoumat závislost AMCC na čase jak z hlediska polynomů, tak z hlediska goniometrických funkcí. Ale v tomto případě je třeba zohlednit konvoluční váhy z Tabulky 5.1.

5.2.1 Konvoluce v polynomických modelech

Uvažujme nejprve, že $\tilde{y}_t$ a $\tilde{z}_t$ jsou polynomy a platí $\tilde{y}_t = \sum_{i=0}^{p-1} w_i \tilde{z}_{t-i},\ p > 1$. Následující tvrzení ukazuje, jak vypadá vztah mezi koeficienty těchto dvou

polynomů $\tilde{y}_t$ a $\tilde{z}_t$.

továrna	subjekt	B_0	B_1	b_0	b_1
I	3	12.89828	0.09533	6.686717	0.04908604
I	6	9.18850563	-0.03936721	4.712512	-0.02027043
⋮	⋮				
III	23	28.670849	1.513067	15.48192	0.7790881
III	24	8.357697	1.317328	4.929519	0.678301
⋮	⋮				
IV	38	12.3890432	0.2883177	6.516228	0.1484567
IV	39	7.6795817	0.2780963	4.086438	0.1431936
⋮	⋮				
V	52	13.5079164	0.1457851	7.024602	0.07506569
V	53	1.7221045	0.1365699	0.9516304	0.07032073
⋮	⋮				

Tabulka 5.5: Odhadnuté a vypočtené koeficienty pro polynomické modely.

Tvrzení 5.1 *Mějme $p > 1$ vah $w_0, w_1, \ldots, w_{p-1}$ a polynomy*

$$\begin{aligned} \tilde{z}_t &= b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k, \\ \tilde{y}_t &= \sum_{i=0}^{p-1} w_i \tilde{z}_{t-i}. \end{aligned}$$

Pak koeficienty polynomu $\tilde{y}_t$ jsou

$$B_0 = \sum_{i=0}^{p-1} w_i (b_0 - b_1 i + b_2 i^2 - \cdots + (-1)^k b_k i^k), \tag{5.6}$$

$$\begin{aligned} B_j = \sum_{i=0}^{p-1} w_i \Big(& b_j \tbinom{j}{0} i^0 - b_{j+1} \tbinom{j+1}{1} i^1 + \\ & + b_{j+2} \tbinom{j+2}{2} i^2 - \cdots + b_k (-1)^{k-j} \tbinom{k}{k-j} i^{k-j} \Big), \end{aligned} \tag{5.7}$$

$j = 1, 2, \ldots, k$.

Důkaz: Polynom $\tilde{z}_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k$ dosadíme do vztahu pro $\tilde{y}_t$

$$\begin{aligned} \tilde{y}_t &= \sum_{i=0}^{p-1} w_i \tilde{z}_{t-i} = \\ &= \sum_{i=0}^{p-1} w_i (b_0 + b_1 (t-i) + b_2 (t-i)^2 + \cdots + b_k (t-i)^k), \end{aligned}$$

Použijeme binomickou větu

$$\begin{aligned}
\tilde{y}_t &= \sum_{i=0}^{p-1} w_i\Big(b_0 + b_1\Big(\tbinom{1}{0}t^1i^0 - \tbinom{1}{1}t^0i^1\Big) + b_2\Big(\tbinom{2}{0}t^2i^0 - \tbinom{2}{1}t^1i^1 + \tbinom{2}{2}t^0i^2\Big) + \\
&\quad + b_3\Big(\tbinom{3}{0}t^3i^0 - \tbinom{3}{1}t^2i^1 + \tbinom{3}{2}t^1i^2 - \tbinom{3}{3}t^0i^3\Big) + \\
&\quad \vdots \\
&\quad + b_{k-1}\Big(\tbinom{k-1}{0}t^{k-1}i^0 - \tbinom{k-1}{1}t^{k-2}i^1 + \tbinom{k-1}{2}t^{k-3}i^2 - \cdots + (-1)^{k-1}\tbinom{k-1}{k-1}t^0i^{k-1}\Big) \\
&\quad + b_k\Big(\tbinom{k}{0}t^ki^0 - \tbinom{k}{1}t^{k-1}i^1 + \tbinom{k}{2}t^{k-2}i^2 - \cdots + (-1)^k\tbinom{k}{k}t^0i^k\Big)\Big)
\end{aligned}$$

Koeficienty u $t^0, t^1, t^2, ..., t^k$ jsou postupně

$$\begin{aligned}
B_0 &= b_0 - b_1\tbinom{1}{1}i^1 + b_2\tbinom{2}{2}i^2 + \cdots + (-1)^k b_k\tbinom{k}{k}i^k \\
B_1 &= b_1\tbinom{1}{0}i^0 - b_2\tbinom{2}{1}i^1 + b_3\tbinom{3}{2}i^2 - \cdots + (-1)^{k-1}b_k\tbinom{k}{k-1}i^{k-1} \\
B_2 &= b_2\tbinom{2}{0}i^0 - b_3\tbinom{3}{1}i^1 + b_4\tbinom{4}{2}i^2 - \cdots + (-1)^{k-2}b_k\tbinom{k}{k-2}i^{k-2} \\
B_3 &= b_3\tbinom{3}{0}i^0 - b_4\tbinom{4}{1}i^1 + b_5\tbinom{5}{2}i^2 - \cdots + (-1)^{k-3}b_k\tbinom{k}{k-3}i^{k-3} \\
&\vdots \\
B_{k-1} &= b_{k-1}\tbinom{k-1}{0}i^0 - b_k\tbinom{k}{1}i^1 \\
B_k &= b_k\tbinom{k}{0}i^0
\end{aligned}$$

Absolutní člen již máme, jen si jej upravíme

$$B_0 = \sum_{i=0}^{p-1} w_i(b_0 - b_1 i + b_2 i^2 - \cdots + (-1)^k b_k i^k),$$

a pro $j = 1, \ldots, k$ dostáváme obecně j-tý koeficient ve tvaru

$$\begin{aligned}
B_j &= \sum_{i=0}^{p-1} w_i\Big(b_j\tbinom{j}{0}i^0 - b_{j+1}\tbinom{j+1}{1}i^1 + \\
&\quad + b_{j+2}\tbinom{j+2}{2}i^2 - \cdots + b_k(-1)^{k-j}\tbinom{k}{k-j}i^{k-j}\Big).
\end{aligned}$$

◇

Jelikož pro DMF jsme zvolili model lineární, "očištěný" model pro AMCC by měl být také lineární a j-té pozorování zapišme jako

$$\begin{aligned}
z_j &= b_0 + b_1 t_j, \\
y_j &= \sum_{k=0}^{p-1} w_k(b_0 + b_1 t_{j-k}) + \epsilon_j,\ \epsilon_j \sim N(0, \sigma_\epsilon^2).
\end{aligned} \tag{5.8}$$

V Tabulce 5.5 jsou uvedeny odhadnuté a vypočtené koeficienty procesů (5.8) pro dva subjekty z každé továrny. Avšak, jak si uvedeme dále, takovéto přepočítání koeficientů není korektní.

5.2.2 Konvoluce v modelech s goniometrickými funkcemi

Podobné tvrzení jako Tvrzení 5.1 si můžeme uvést i pro případ, kdy uvažujeme gonimetrické funkce.

Tvrzení 5.2 *Mějme $p > 1$ vah $w_0, w_1, \ldots, w_{p-1}$ a funkce $\tilde{z}_t$ a $\tilde{y}_t$ tvaru*

$$\begin{aligned}\tilde{z}_t &= b_0 + b_1\cos(u_1 t) + b_2\sin(u_1 t) + b_3\cos(u_2 t) + b_4\sin(u_2 t) + \\ &\quad + \cdots + b_{2k-1}\cos(u_k t) + b_{2k}\sin(u_k t),\\ \tilde{y}_t &= \sum_{i=0}^{p-1} w_i \tilde{z}_{t-i},\end{aligned}$$

kde $u_1, u_2, \ldots, u_k$ jsou frekvence goniometrických funkcí. Pak koeficienty funkce $\tilde{y}_t$ jsou

$$B_0 = b_0 \sum_{i=0}^{p-1} w_i, \tag{5.9}$$

$$\begin{aligned}B_j &= b_j \sum_{i=0}^{p-1} w_i \cos(u_{\frac{j+1}{2}} i) - b_{j+1} \sum_{i=0}^{p-1} w_i \sin(u_{\frac{j+1}{2}} i) \ \textit{pro } j \ \textit{liché}\\ &= b_{j-1} \sum_{i=0}^{p-1} w_i \sin(u_{\frac{j}{2}} i) + b_j \sum_{i=0}^{p-1} w_i \cos(u_{\frac{j}{2}} i) \ \textit{pro } j \ \textit{sudé},\end{aligned} \tag{5.10}$$

$j = 1, 2, \ldots, 2k$.

Důkaz: Nové koeficienty dostaneme dosazením $\tilde{z}_t$ a úpravami s použitím vzorců pro sin a cos rozdílu dvou argumentů

$$\begin{aligned}\sin(x-y) &= \sin x\cos y - \cos x\sin y\\ \cos(x-y) &= \cos x\cos y + \sin x\sin y\end{aligned}$$

$$\begin{aligned}\tilde{y}_t &= \sum_{i=0}^{p-1} w_i \tilde{z}_{t-i} =\\ &= \sum_{i=0}^{p-1} w_i(b_0 + b_1\cos(u_1(t-i)) + b_2\sin(u_1(t-i)) + \\ &\quad + \cdots + b_{2k-1}\cos(u_k(t-i)) + b_{2k}\sin(u_k(t-i)) =\end{aligned}$$

$$
\begin{aligned}
&= b_0 \sum_{i=0}^{p-1} w_i + b_1 \sum_{i=0}^{p-1} w_i \cos(u_1(t-i)) + b_2 \sum_{i=0}^{p-1} w_i \sin(u_1(t-i)) + \\
&\quad + \cdots + b_{2k-1} \sum_{i=0}^{p-1} w_i \cos(u_k(t-i)) + b_{2k} \sum_{i=0}^{p-1} w_i \sin(u_k(t-i)) = \\
&= b_0 \sum_{i=0}^{p-1} w_i + b_1 \sum_{i=0}^{p-1} w_i [\cos(u_1 t)\cos(u_1 i) + \sin(u_1 t)\sin(u_1 i)] + \\
&\quad + b_2 \sum_{i=0}^{p-1} w_i [\sin(u_1 t)\cos(u_1 i) - \cos(u_1 t)\sin(u_1 i)] + \cdots + \\
&\quad + b_{2k-1} \sum_{i=0}^{p-1} w_i [\cos(u_k t)\cos(u_k i) + \sin(u_k t)\sin(u_k i)] + \\
&\quad + b_{2k} \sum_{i=0}^{p-1} w_i [\sin(u_k t)\cos(u_k i) - \cos(u_k t)\sin(u_k i)].
\end{aligned}
$$

Absolutní člen je $B_0 = b_0 \sum_{i=0}^{p-1} w_i$, ostatní koeficienty přísluší postupně $\cos(u_1 t), \sin(u_1 t), \ldots, \cos(u_k t), \sin(u_k t)$ a mají tedy tvar

$$
\begin{aligned}
B_1 &= b_1 \sum_{i=0}^{p-1} w_i \cos(u_1 i) - b_2 \sum_{i=0}^{p-1} w_i \sin(u_1 i) \\
B_2 &= b_1 \sum_{i=0}^{p-1} w_i \sin(u_1 i) + b_2 \sum_{i=0}^{p-1} w_i \cos(u_1 i) \\
&\vdots \\
B_{2k-1} &= b_{2k-1} \sum_{i=0}^{p-1} w_i \cos(u_k i) - b_k \sum_{i=0}^{p-1} w_i \sin(u_k i) \\
B_{2k} &= b_{2k-1} \sum_{i=0}^{p-1} w_i \sin(u_k i) + b_{2k} \sum_{i=0}^{p-1} w_i \cos(u_k i),
\end{aligned}
$$

Označíme-li $j = 2k - 1$ je j liché a koeficient

$$
B_j = b_j \sum_{i=0}^{p-1} w_i \cos(u_{\frac{j+1}{2}} i) - b_{j+1} \sum_{i=0}^{p-1} w_i \sin(u_{\frac{j+1}{2}} i)
$$

a označíme-li $j = 2k$ je j sudé a koeficient je tvaru

$$
B_j = b_{j-1} \sum_{i=0}^{p-1} w_i \sin(u_{\frac{j}{2}} i) + b_j \sum_{i=0}^{p-1} w_i \cos(u_{\frac{j}{2}} i).
$$

$\diamond$

Italská data modelujeme pouze jednou dvojicí goniometrických funkcí se

stejnou frekvencí. Označme $u_1 = u$ a j-té pozorování AMCC zapišme jako

$$\begin{aligned} z_j &= b_0 + b_1 \cos(ut_j) + b_2 \sin(ut_j) \\ y_j &= \sum_{k=0}^{7} w_k (b_0 + b_1 \cos(ut_{j-k}) + b_2 \sin(ut_{j-k})) + \epsilon_j,\ \epsilon_j \sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{5.11}$$

továrna	subjekt	koeficient	i=0	i=1	i=2
I	3	B_i	16.901200	-7.552078	-5.023650
		b_i	8.702538	-11.714636	-8.175989
I	6	B_i	7.49292407	-0.22076678	-0.09769363
		b_i	3.8581556	-0.3379725	-0.1657262
⋮	⋮		⋮	⋮	⋮
III	23	B_i	94.06781	-25.21090	-72.94668
		b_i	48.43613	-44.22219	-111.03051
III	24	B_i	74.61858	-23.61797	-67.08064
		b_i	38.42159	-41.31359	-102.14151
⋮	⋮		⋮	⋮	⋮
IV	38	B_i	24.647948	4.852382	-9.446310
		b_i	12.691390	6.372791	-13.638859
IV	39	B_i	19.665681	-3.827025	-9.624852
		b_i	10.125988	-6.581044	-14.695360
⋮	⋮		⋮	⋮	⋮
V	52	B_i	19.746811	-2.122614	-7.508885
		b_i	10.167762	-3.847756	-11.386079
V	53	B_i	7.576853	-1.251623	-4.747290
		b_i	3.901371	-2.297975	-7.190314
⋮	⋮		⋮	⋮	⋮

Tabulka 5.6: Odhadnuté a vypočtené koeficienty pro modely s goniometrickými funkcemi.

Tabulka 5.6 uvádí přepočtené koeficienty tentokrát pro procesy (5.11). Jak jsme se již zmínili, pouhé přepočítání koeficientů není korektní. Nevíme totiž nic o chybách a dalších vlastnostech modelu pro z_j. Vhodnější je upravit regresní matici. Do modelu pak můžeme vkládat data, která máme k dispozici a získáme odhady "očištěného" modelu, které nás zajímají. To si ukážeme již přímo na modelech se smíšenými efekty v § 5.3.3 a § 5.3.4.

5.3 Modely se smíšenými efekty pro reálná data

V této části se již zaměříme na samotné modely se smíšenými efekty pro Italská data. Modely vybudujeme pro obě sledované látky, DMF a AMCC, zvlášť. V případě smíšeného modelu pro AMCC musíme opět brát v úvahu konvoluce.

Podívejme se ještě jednou blíže na naše data. Zpracováváme měření od 40-ti subjektů, tj. $M = 40$. Pro i-tého jedince máme n_i-rozměrný vektor odezvy $\boldsymbol{y}_i$. Počet pozorování je obvykle $n_i = 6$, ale jak je vidět v Tabulce 1.1, vyskytly se i výjimky se třemi, čtyřmi či pěti pozorováními. Celkem máme $\sum_{i=1}^{M} n_i$ pozorování. Celkový počet pozorování se pro obě látky liší. Opět z Tabulky 1.1 lze vyčíst, že pro DMF máme 223 pozorování a pro AMCC o tři méně.

Během zpracovávání dat pro jednotlivé subjekty nebylo podstatné, ve které továrně pracují. Nyní se budeme věnovat také rozdílům mezi továrnami. Původně mělo být továren pět, nicméně jedna továrna data neposkytla, tudíž máme ke zpracování data ze čtyř továren. Pro větší přehlednost si uveďme převodní Tabulku 5.7 mezi čísly továren a indexy veličin v modelech uvedených dále.

Továrna	index
I	1
II	neposkytla data
III	2
IV	3
V	4

Tabulka 5.7: Převodní tabulka mezi čísly továren a indexy veličin.

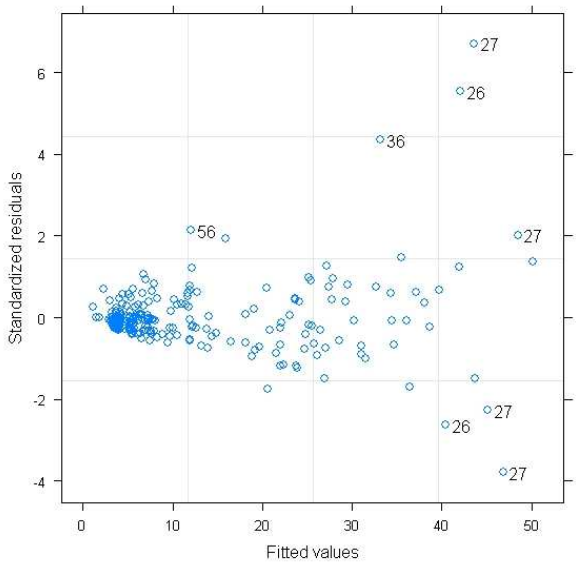
V předešlých kapitolách jsme usoudili, že pro modelování jednotlivých subjektů je vhodnější, a jistě také pohodlnější, použít lineární model, oproti vyšším polynomům či modelům s goniometrickými funkcemi. Z lineárního modelu budeme tedy vycházet i nyní, při tvorbě modelů smíšených. Efekty továren budou v modelech vystupovat jako efekty pevné. Kdykoliv hledáme závislost na nějakém faktoru, použijeme reparametrizaci pomocí kontrastů. V `R` je dostupných několik možných voleb kontrastů. Standardním nastavením v `R` je `contr.treatment`. Jedna z úrovní faktoru se zvolí jako základní a ostatní se s touto úrovní porovnávají. V našem případě je základní úrovní továrna I.

5.3.1 Funkce `lme` v `R`

Pro zápis lineárních modelů se smíšenými efekty slouží v `R` funkce `lme` z již zmiňované knihovny `nlme`.

```
lme(fixed, data, random),
```

kde parametr `fixed` představuje pevné a `random` náhodné efekty. V zápisu modelu lze definovat i další volitelné parametry. Například, chceme-li porovnávat modely s různými pevnými efekty, je nutné odhadovat parametry metodou



Obrázek 5.6: Rozložení standardizovaných reziduí ukazuje na nestejné rozptyly.

maximální věrohodnosti (ML). Ve funkci `lme` je standardním nastavením REML a lze jej změnit v definici modelu formulací `method="ML"`.

5.3.2 LME model pro DMF

Naším cílem je modelovat DMF v závislosti na čase a na tom, ve které továrně je subjekt, na němž je prováděno meření, zaměstnán. Přičemž chceme odhadovat parametry pevné, které budou stejné pro všechny subjekty, a parametry náhodné, které se budou mezi jednotlivci lišit.

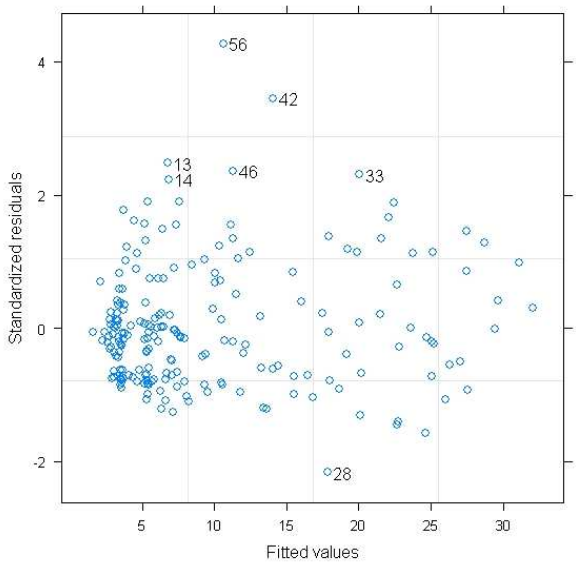
Zapišme j-té pozorování na i-tém jedinci jako

$$\begin{aligned} y_{ij} \quad = \quad & (\beta_0 + \xi_{02}D_{2i} + \xi_{03}D_{3i} + \xi_{04}D_{4i} + b_{0i}) + \\ & (\beta_1 + \xi_{12}D_{2i} + \xi_{13}D_{3i} + \xi_{14}D_{4i})t_{ij} + \epsilon_{ij}, \\ & \\ & b_{0i} \sim N(0, \sigma_b^2),\ \epsilon_{ij} \sim N(0, \sigma^2\lambda_{ijj}), \\ & j = 1, \ldots, n_i,\ i = 1, \ldots, 40, \end{aligned} \tag{5.12}$$

kde $D_{mi} = 1$ jestliže i-tý jedinec pracuje v m-té továrně a $D_{mi} = 0$ jestliže pracuje v jiné továrně, $m = 1, 2, 3, 4$. Pro modelování koncentrace DMF je postačující, když za náhodný považujeme pouze absolutní člen b_{0i}.

Na Obrázku 5.6 vidíme, podle rozložení standardizovaných reziduí, že je porušen předpoklad homoskedasticity. K modelování heteroskedasticity se používá varianční funkce, která je definována jako

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2 g^2(\mu_{ij}, \boldsymbol{\upsilon}_{ij}, \boldsymbol{\delta}),\ i = 1, ..., M,\ j = 1, ..., n_i, \tag{5.13}$$



Obrázek 5.7: Rozložení reziduí po vynechání subjektů 26, 27, 36 a po použití funkce `varPower`.

kde $\mu_{ij} = E[y_{ij}] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, $\boldsymbol{\upsilon}_{ij}$ je vektor kovariát, $\boldsymbol{\delta}$ je vektor variančních parametrů a $g(.)$ je varianční funkce spojitá v $\boldsymbol{\delta}$.

Při zápisu modelu v `R` si lze zvolit i jiný než konstantní rozptyl pomocí parametru `weights` v definici modelu. Jednou z možností je varianční funkce `varIdent`, která reprezentuje varianční model s různými rozptyly pro každou úroveň stratifikační proměnné s, nabývající hodnot z množiny $\{1, 2, ..., S\}$,

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2\delta_{s_{ij}}^2,\ g(s_{ij}, \boldsymbol{\delta}) = \delta_{s_{ij}}. \tag{5.14}$$

Tento model používá $S + 1$ parametrů k reprezentaci S rozptylů, k dosažení jednoznačnosti je třeba omezit $\boldsymbol{\delta}$. Položíme $\delta_1 = 1$, takže $\delta_l,\ l = 2, \ldots, S$ reprezentuje poměr mezi standardními odchylkami l-té a první úrovně, $\delta_l > 0,\ l = 2, \ldots, S$.

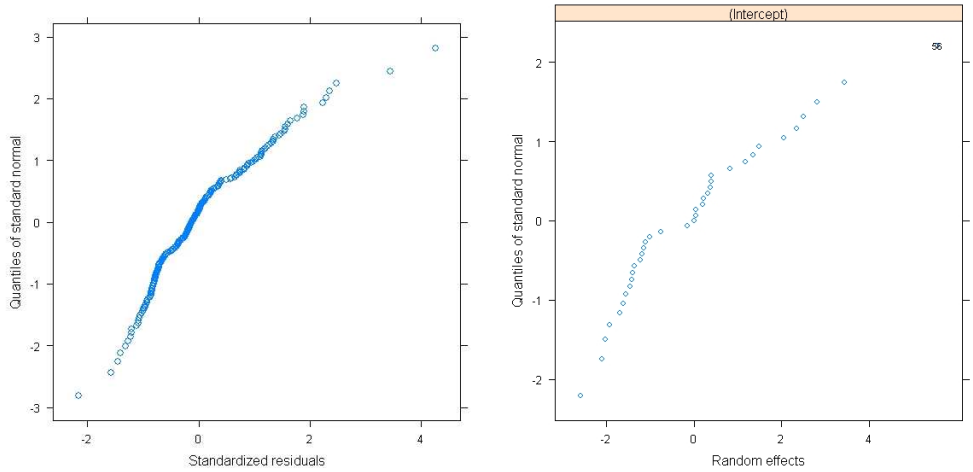
Jinou možností je zvolit `varPower`, kde

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2|\upsilon_{ij}|^{2\delta},\ g(\upsilon_{ij}, \delta) = |\upsilon_{ij}|^{\delta}. \tag{5.15}$$

A zvolit si lze například i model s exponenciální varianční funkcí `varExp`

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2\exp(2\delta\upsilon_{ij}),\ g(\upsilon_{ij}, \delta) = \exp(\delta\upsilon_{ij}), \tag{5.16}$$

kde parametr δ je libovolný.



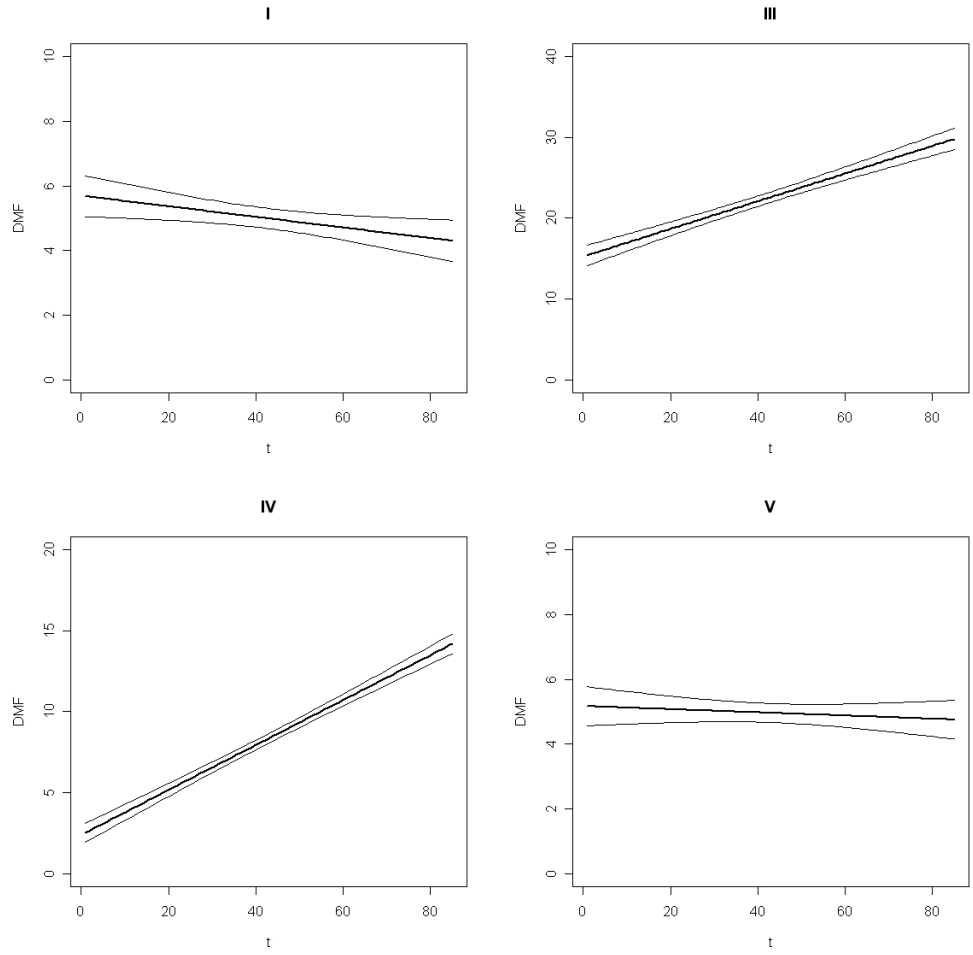
Obrázek 5.8: Ověření předpokladů normality.

Podívejme se ještě jednou na Obrázek 5.6. Standardizovaná rezidua, která přesáhla mez 0.975 v absolutní hodnotě, jsou v grafech označena čísly subjektu. Všimněme si subjektů 26, 27 a 36 s největšími rezidui. Tyto subjekty pocházejí všechny z továrny III a vykonávají stejnou práci. Jedná se o jakési dokončovací práce, to však není podstatné, podstatné v tuto chvíli je, že až na tyto tři subjekty nikdo jiný stejnou práci nevykonává ani v této ani v jiné z ostatních továren. Zkusme proto tyto subjekty z našich dat vynechat. Po jejich vynechání se jeví jako nejvhodnější použít varianční funkci `varPower` s volbou parametru $\delta = 1/2$. To znamená, že rozptyl roste lineárně s fitovanými hodnotami. Rozložení standardizovaných reziduí, jak je vidět na Obrázku 5.7, je pak výrazně lepší. Podle Obrázku 5.8 můžeme (s jistou rezervou) říci, že předpoklady normality jsou splněny.

Pevné efekty	parametr	odhad
továrna I	$\gamma_{01} = \beta_0$	5.696637
	$\gamma_{11} = \beta_1$	-0.016439
továrna III	$\gamma_{02} = \beta_0 + \xi_{02}$	15.259055758
	$\gamma_{12} = \beta_1 + \xi_{12}$	0.171327778
továrna IV	$\gamma_{03} = \beta_0 + \xi_{03}$	2.388600950
	$\gamma_{13} = \beta_1 + \xi_{13}$	0.138838414
továrna V	$\gamma_{04} = \beta_0 + \xi_{04}$	5.172073994
	$\gamma_{14} = \beta_1 + \xi_{14}$	-0.004903646
Komponenty	σ	1.766298
rozptylu	σ_b	2.450896

Tabulka 5.8: Odhady parametrů a komponenty rozptylu modelu (5.12).

V Tabulce 5.8 jsou shrnuty odhady parametrů modelu (5.12) s varianční funkcí (5.15) a volbou $\delta = 1/2$. Parametry $\gamma_{0i},\ i = 1, \ldots, 4$ odpovídají odhadům středních hodnot v příslušných továrnách a parametry $\gamma_{1i},\ i = 1, \ldots, 4$ říkají,



Obrázek 5.9: Odhady z modelu (5.12) pro všechny čtyři továrny s konfidenčními intervaly.

jak se změní koncentrace DMF, když uplyne jeden den. A na Obrázku 5.9 jsou tyto odhady spolu s konfidenčními intervaly vyneseny do grafu.

5.3.3 LME model pro AMCC a konvoluce

Jak jsme již uvedli v Kapitole 5.2, pouhé přepočítání odhadnutých parametrů není korektní. My chceme pomocí koncentrací AMCC odhadnout parametry procesu, o kterém víme, díky modelu pro DMF, že je lineární. To znamená, že je potřeba upravit si jak regresní matici pro pevné, tak pro náhodné efekty.

V modelu (5.12), tj. v situaci, kdy y_{ij} je skutečně y_{ij}, vypadá regresní matice $\boldsymbol{X}$ o rozměrech $(\sum_{i=1}^{M} n_i \times 8)$ takto

$$\boldsymbol{X} = \begin{pmatrix} 1 & D_{21} & D_{31} & D_{41} & t_{11} & D_{21}t_{11} & D_{31}t_{11} & D_{41}t_{11} \\ 1 & D_{21} & D_{31} & D_{41} & t_{12} & D_{21}t_{12} & D_{31}t_{12} & D_{41}t_{12} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & D_{21} & D_{31} & D_{41} & t_{1n_1} & D_{21}t_{1n_1} & D_{31}t_{1n_1} & D_{41}t_{1n_1} \\ 1 & D_{22} & D_{32} & D_{42} & t_{21} & D_{22}t_{21} & D_{32}t_{21} & D_{42}t_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & D_{22} & D_{32} & D_{42} & t_{2n_2} & D_{22}t_{2n_2} & D_{32}t_{2n_2} & D_{42}t_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & D_{2M} & D_{3M} & D_{4M} & t_{Mn_M} & D_{2M}t_{Mn_M} & D_{3M}t_{Mn_M} & D_{4M}t_{Mn_M} \end{pmatrix},$$

kde $D_{mi} = 1$ jestliže i-tý jedinec pracuje v m-té továrně a $D_{mi} = 0$ jestliže pracuje v jiné továrně, n_i je počet pozorování i-tého jedince, M je celkový počet jedinců. Jedná se o blokovou matici

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{D}_1^* \\ \boldsymbol{D}_2 & \boldsymbol{D}_2^* \\ \vdots & \vdots \\ \boldsymbol{D}_M & \boldsymbol{D}_M^* \end{pmatrix}, \tag{5.17}$$

kde

$$\boldsymbol{D}_i = \begin{pmatrix} 1 & D_{2i} & D_{3i} & D_{4i} \\ 1 & D_{2i} & D_{3i} & D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & D_{2i} & D_{3i} & D_{4i} \end{pmatrix} \tag{5.18}$$

a

$$\boldsymbol{D}_i^* = \begin{pmatrix} t_{i1} & D_{2i}t_{i1} & D_{3i}t_{i1} & D_{4i}t_{i1} \\ t_{i2} & D_{2i}t_{i2} & D_{3i}t_{i2} & D_{4i}t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ t_{in_i} & D_{2i}t_{in_i} & D_{3i}t_{in_i} & D_{4i}t_{in_i} \end{pmatrix}, \tag{5.19}$$

$i = 1, ..., M$.
Chceme-li přímo odhady regresních parametrů pro jednotlivé továrny, pak

$$\boldsymbol{D}_i = \begin{pmatrix} D_{1i} & D_{2i} & D_{3i} & D_{4i} \\ D_{1i} & D_{2i} & D_{3i} & D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ D_{1i} & D_{2i} & D_{3i} & D_{4i} \end{pmatrix}, \tag{5.20}$$

$$\boldsymbol{D}_i^* = \begin{pmatrix} D_{1i}t_{i1} & D_{2i}t_{i1} & D_{3i}t_{i1} & D_{4i}t_{i1} \\ D_{1i}t_{i2} & D_{2i}t_{i2} & D_{3i}t_{i2} & D_{4i}t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ D_{1i}t_{in_i} & D_{2i}t_{in_i} & D_{3i}t_{in_i} & D_{4i}t_{in_i} \end{pmatrix}, \tag{5.21}$$

$i = 1, ..., M$.

Ale y_{ij} jsou nyní $y_{ij} = \sum_{k=0}^{p} w_k(z_{ij} - k) + \epsilon_{ij}$. Pro přehlednost si označme

$$A_1 := \sum_{k=0}^{p} w_k, \ A_2 := \sum_{k=0}^{p} k w_k, \tag{5.22}$$

a upravenou regresní matici $\boldsymbol{X}$ zapišme rovnou ve formě blokové matice (5.17), ale bloky nyní vypadají takto

$$\boldsymbol{D}_i = \begin{pmatrix} A_1 & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \\ A_1 & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ A_1 & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \end{pmatrix}, \tag{5.23}$$

$$\boldsymbol{D}_i^* = \begin{pmatrix} A_1 t_{i1} - A_2 & D_{2i}(A_1 t_{i1} - A_2) & \dots & D_{4i}(A_1 t_{i1} - A_2) \\ A_1 t_{i2} - A_2 & D_{2i}(A_1 t_{i2} - A_2) & \dots & D_{4i}(A_1 t_{i2} - A_2) \\ \vdots & \vdots & & \vdots \\ A_1 t_{in_i} - A_2 & D_{2i}(A_1 t_{in_i} - A_2) & \dots & D_{4i}(A_1 t_{in_i} - A_2) \end{pmatrix}, \tag{5.24}$$

$i = 1, ..., M$.

Analogicky jako v (5.20) a (5.21) můžeme bloky přepsat

$$\boldsymbol{D}_i = \begin{pmatrix} A_1 D_{1i} & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \\ A_1 D_{1i} & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ A_1 D_{1i} & A_1 D_{2i} & A_1 D_{3i} & A_1 D_{4i} \end{pmatrix}, \tag{5.25}$$

$$\boldsymbol{D}_i^* = \begin{pmatrix} D_{1i}(A_1 t_{i1} - A_2) & \dots & \dots & D_{4i}(A_1 t_{i1} - A_2) \\ D_{1i}(A_1 t_{i2} - A_2) & \dots & \dots & D_{4i}(A_1 t_{i2} - A_2) \\ \vdots & & & \vdots \\ D_{1i}(A_1 t_{in_i} - A_2) & \dots & \dots & D_{4i}(A_1 t_{in_i} - A_2) \end{pmatrix}, \tag{5.26}$$

$i = 1, ..., M$.

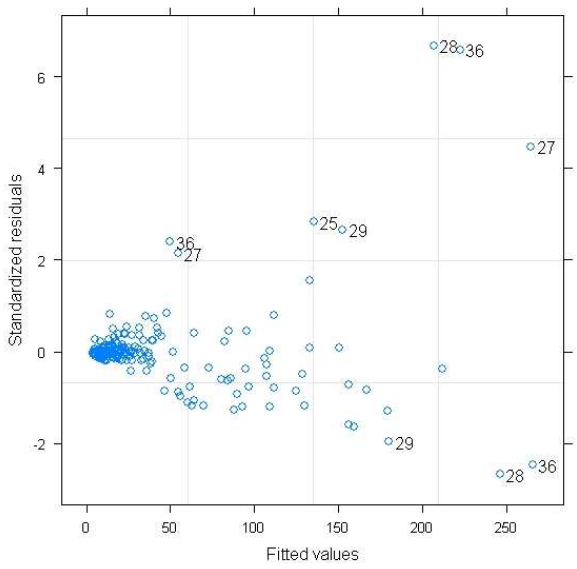
Takto upravenou matici si označme $\boldsymbol{X}_A$. Stejně upravíme i matici náhodných efektů

$$\boldsymbol{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}, \tag{5.27}$$

$i = 1, ..., M$.

Upravenou matici náhodných efektů označme $\boldsymbol{Z}_{Ai}$,

$$\boldsymbol{Z}_{Ai} = \begin{pmatrix} A_1 & A_1 t_{i1} - A_2 \\ A_1 & A_1 t_{i2} - A_2 \\ \vdots & \vdots \\ A_1 & A_1 t_{in_i} - A_2 \end{pmatrix}, \tag{5.28}$$



Obrázek 5.10: Rozložení reziduí ukazuje na nestejné rozptyly.

$i = 1, ..., M$.

Model pro AMCC zapišme maticově

$$\boldsymbol{y} \;=\; \boldsymbol{X}_A\boldsymbol{B} + \boldsymbol{Z}_A\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{5.29}$$

kde

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta} \\ \vdots \\ \boldsymbol{\beta} \end{bmatrix}, \boldsymbol{Z}_A = \begin{bmatrix} \boldsymbol{Z}_{A1} \\ \boldsymbol{Z}_{A2} \\ \vdots \\ \boldsymbol{Z}_{AM} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_M \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix},$$

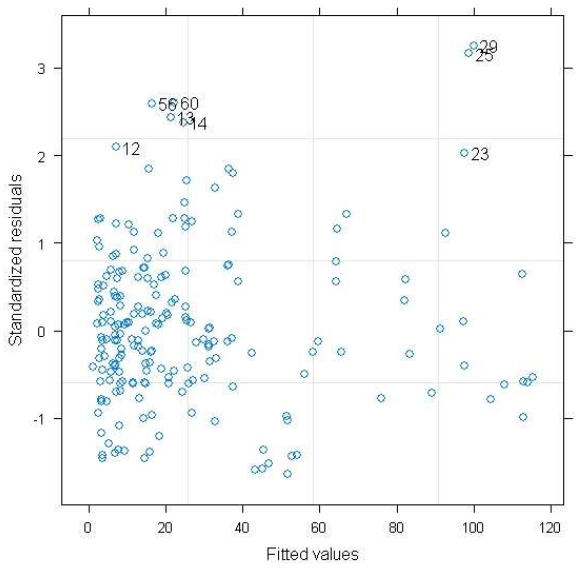
$$\boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}),\ \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_i),\ i = 1, \ldots, M,$$

kde $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$ a matice $\boldsymbol{\Lambda}_i$ je také diagonální s prvky λ_{ijj}. V případě AMCC je vhodnější zahrnout do modelu oba náhodné parametry.

Model (5.29) opět nesplňuje předpoklad homoskedasticity. Tentokrát mají na Obrázku 5.10 největší rezidua subjekty 27, 28 a 36. To, že nyní je v této trojici místo subjektu 26 subjekt 28, si lze vysvětlit například tak, že při sběru vzorků došlo k záměně. Ať už to tak bylo či nikoliv, vyzkoušejme tyto subjekty opět vynechat. A použijme vhodnou varianční funkci. Matice $\boldsymbol{X}_A$ a $\boldsymbol{Z}_{Ai}$ budeme ještě dále upravovat, proto si odhady parametrů modelu uvedeme až v další kapitole.

5.3.4 Upravený LME model pro AMCC

Hodnotu koncentrace AMCC ovlivňuje ještě jeden podstatný faktor. Tímto faktorem je přítomnost či nepřítomnost subjektu v zaměstnání v předchozích



Obrázek 5.11: Rozložení standardizovaných reziduí po vynechání subjektů 27, 28, 36 a použití funkce `varPower` s volbou $\delta = 1$.

sedmi dnech ode dne, pro který máme pozorování. Pokud v některém z těchto dnů nebyl subjekt v práci, nebudeme příslušný přírůstek započítávat. Indikátor absence v zaměstnání označme I_p, kdy $I_p(t) = 1$ pokud byl v čase t subjekt v práci a $I_p(t) = 0$ pokud nebyl. Sumy (5.22) se pak pro každé pozorování budou lišit. Označme

$$\begin{aligned} G_{ij} &:= \sum_{k=0}^{p} w_k I_p(t_{ij} - k) \\ H_{ij} &:= \sum_{k=0}^{p} k w_k I_p(t_{ij} - k) \end{aligned} \tag{5.30}$$

Bloky matice (5.17) pak budou

$$\boldsymbol{D}_i = \begin{pmatrix} G_{i1} & G_1 D_{2i} & G_{i1} D_{3i} & G_{i1} D_{4i} \\ G_{i2} & G_1 D_{2i} & G_{i2} D_{3i} & G_{i2} D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ G_{in_i} & G_1 D_{2i} & G_{in_i} D_{3i} & G_{in_i} D_{4i} \end{pmatrix}, \tag{5.31}$$

$$\boldsymbol{D}_i^* = \begin{pmatrix} G_{i1} t_{i1} - H_{i1} & \dots & \dots & D_{4i}(G_{i1} t_{i1} - H_{i1}) \\ G_{i2} t_{i2} - H_{i2} & \dots & \dots & D_{4i}(G_{i2} t_{i2} - H_{i2}) \\ \vdots & & & \vdots \\ G_{in_i} t_{in_i} - H_{in_i} & \dots & \dots & D_{4i}(G_{in_i} t_{in_i} - H_{in_i}) \end{pmatrix}, \tag{5.32}$$

Pevné efekty	parametr	odhad
továrna I	γ_{01}	6.244892
	γ_{11}	-0.009847
továrna III	γ_{02}	13.652653
	γ_{12}	0.575407
továrna IV	γ_{03}	1.211696
	γ_{13}	0.229091
továrna V	γ_{04}	4.224222
	γ_{14}	0.044261
Komponenty	σ	0.5813620
rozptylu	σ_0	3.83056379
	σ_1	0.00004783

Tabulka 5.9: Odhady parametrů modelu (5.36).

$i = 1, ..., M$.
Analogicky

$$\boldsymbol{D}_i = \begin{pmatrix} G_{i1}D_{1i} & G_{i1}D_{2i} & G_{i1}D_{3i} & G_{i1}D_{4i} \\ G_{i2}D_{1i} & G_{i2}D_{2i} & G_{i2}D_{3i} & G_{i2}D_{4i} \\ \vdots & \vdots & \vdots & \vdots \\ G_{in_i}D_{1i} & G_{in_i}D_{2i} & G_{in_i}D_{3i} & G_{in_i}D_{4i} \end{pmatrix}, \tag{5.33}$$

$$\boldsymbol{D}_i^* = \begin{pmatrix} D_{1i}(G_{i1}t_{i1} - H_{i1}) & \dots & \dots & D_{4i}(G_{i1}t_{i1} - H_{i1}) \\ D_{1i}(G_{i2}t_{i2} - H_{i2}) & \dots & \dots & D_{4i}(G_{i2}t_{i2} - H_{i2}) \\ \vdots & & & \vdots \\ D_{1i}(G_{in_i}t_{in_i} - H_{in_i}) & \dots & \dots & D_{4i}(G_{in_i}t_{in_i} - H_{in_i}) \end{pmatrix}, \tag{5.34}$$

$i = 1, ..., M$. Matici náhodných efektů upravíme takto

$$\boldsymbol{Z}_{Ci} = \begin{pmatrix} G_{i1} & G_{i1}t_{i1} - H_{i1} \\ G_{i2} & G_{i2}t_{i2} - H_{i2} \\ \vdots & \vdots \\ G_{in_i} & G_{in_i}t_{in_i} - H_{in_i} \end{pmatrix}, \tag{5.35}$$

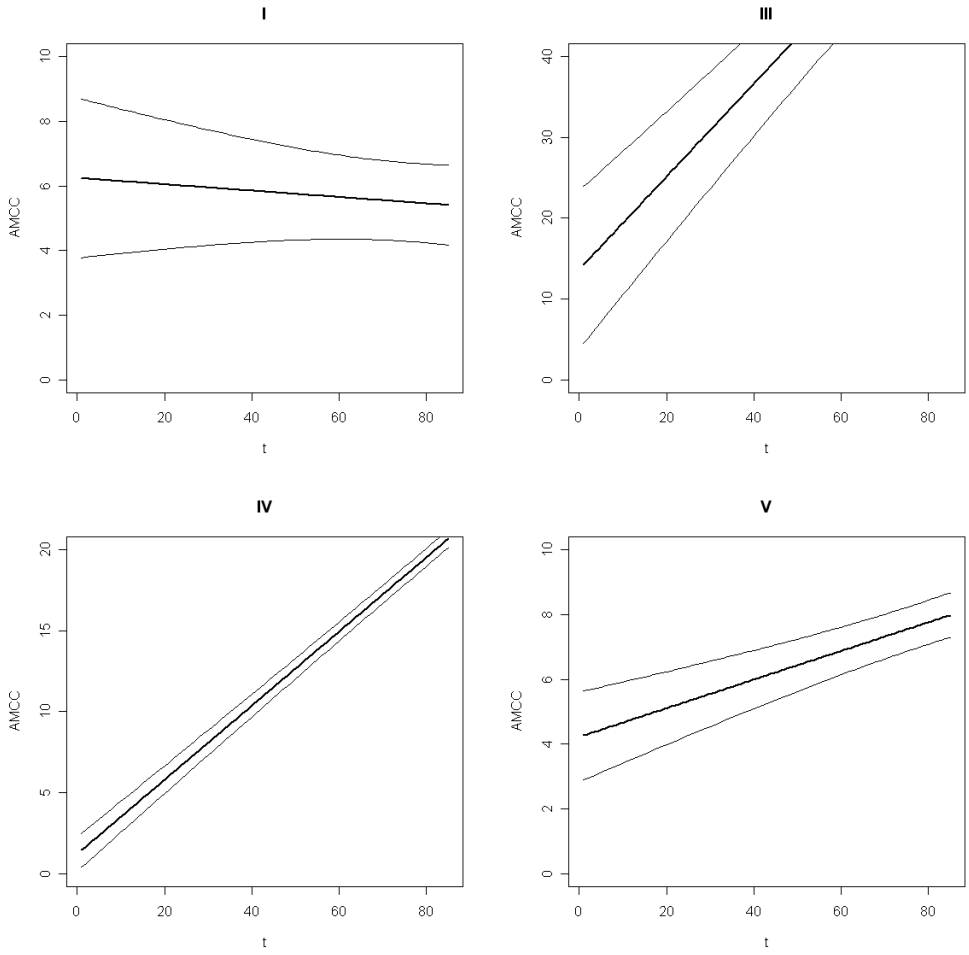
$i = 1, ..., M$.

Model pro AMCC nyní zapišme takto

$$\boldsymbol{y} \;=\; \boldsymbol{X}_C\boldsymbol{B} + \boldsymbol{Z}_C\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{5.36}$$

kde

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta} \\ \vdots \\ \boldsymbol{\beta} \end{bmatrix}, \boldsymbol{Z}_C = \begin{bmatrix} \boldsymbol{Z}_{C1} \\ \boldsymbol{Z}_{C2} \\ \vdots \\ \boldsymbol{Z}_{CM} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_M \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix},$$



Obrázek 5.12: Odhady z modelu (5.36) pro všechny čtyři továrny a odpovídající konfidenční intervaly.

$$\boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{\Psi}),\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i),\ i = 1, \ldots, M,$$

$\boldsymbol{\Psi} = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$ a $\boldsymbol{\Lambda}_i$ je opět diagonální. $\boldsymbol{X}_C$ je upravená regresní matice (5.17) s bloky (5.31) a (5.32), resp. (5.33) a (5.34).

Jak již víme z Obrázku 5.10 není splněn předpoklad homoskedasticity. Opět použijeme varianční funkci `varPower`, ale tentokrát s volbou $\delta = 1$. Na Obrázku 5.11 je vidět, že model zohledňující heteroskedasticitu je výrazně lepší.

Odhady parametrů modelu (5.36) a komponenty rozptylu jsou shrnuty v Tabulce 5.9.

Kapitola 6

Závěr

AMCC vzniká rozkladem dimethylformamidu, pro který jsme použili lineární model se smíšenými efekty (5.12). V modelech (5.29) a (5.36) jsme AMCC "očistili" od přírůstků předchozích dnů. Model (5.36) je vhodnější díky tomu, že navíc zohledňuje, zda byl či nebyl sledovaný subjekt v zaměstnání v předchozích dnech.

To, co subjekt vdechne (tj. koncentrace DMF), a to, co u něj naměříme (tj. koncentrace AMCC), by si mělo odpovídat. Modely (5.12) a (5.36) by tedy měly dávat přibližně stejné výsledky. Můžeme je porovnat v Tabulkách 5.8 a 5.9. Nejvýraznějším rozdílem je, že v továrně V vychází dokonce opačná směrnice. To, že jsou v odhadech patrné odlišnosti, je způsobeno především tím, že dimethylformamid se rozkládá nejenom na AMCC, ale na mnoho dalších látek, viz. Obrázek 2.1. Z těch měřitelných je to například ještě v krvi zjistitelné MVH, které má jiný průběh rozkladu než zpracovávané AMCC. Rozdíly v odhadech mohou být také způsobeny nevhodností použitých modelů. Bylo by jistě vhodné zkusit v modelech uvažovat korelaci.

Dalším krokem by bylo vytvořit společný model pro všechny měřené látky.

Literatura

[1] *R*. http://www.r-project.org/.

[2] Jiří Anděl. *Základy matematické statistiky*. UK, Matematicko-fyzikální fakulta, Praha, 2002.

[3] J. C. Pinheiro, D. M. Bates. *Mixed-effects models in S and S-plus*. Springer, New York, 2000.

[4] Karel Zvára. *RaR*. http://www.karlin.mff.cuni.cz/zvara/.

[5] M. D. Davidian, D. M. Giltinan. *Nonlinear models for repeated measurement data*. Chapman and Hall, London, 1995.

[6] Petr Šmilauer. *Moderní regresní metody*. Biologická fakulta JU. http://regent.jcu.cz/MRM.pdf.

[7] Mudr. Martin Vokurka CSc., Mudr. Jan Hugo a kol. *Velký lékařský slovník*. Maxdorf, Praha, 2002.

[8] CSc. RNDr. Jaroslav Mráz. Ústní sdělení. Státní zdravotní ústav.

[9] U.S.Environmental Protection Agency. *Overall summary for N-methyl formamide*. http://www.epa.gov/hpv/pubs/summaries/monomthf/c15159tp.pdf.