Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

Zuzana ŘÍMSKÁ

MODELY PRO DISKRÉTNÍ LONGITUDINÁLNÍ DATA A JEJICH APLIKACE PŘI VYŠETŘOVÁNÍ DOTAZNÍKŮ

Magisterská práce

Vedoucí práce: Mgr. Pavel Tuček, Ph.D.

Olomouc 2012

Prohlášení

Prohlašuji, že jsem magisterskou práci magisterského studia oboru Geoinformatika vypracovala samostatně pod vedením Mgr. Pavla Tučka, Ph. D.

Všechny použité materiály a zdroje jsou citovány s ohledem na vědeckou etiku, autorská práva a zákony na ochranu duševního vlastnictví. Všechna poskytnutá i vytvořená digitální data nebudu bez souhlasu školy poskytovat.

14. duben 2012 Zuzana ŘÍMSKÁ

Anotace

Tato diplomová práce popisuje vybrané regresní modely a longitudinální analýzu. Větší část této diplomové práce je věnována lineárním mixovaným modelům a generalizovaným lineárním mixovaným modelům. Tyto modely jsou součástí longitudinální analýzy. Veškerá teorie je demonstrována na příkladech, které jsou spočteny pomocí implementace jazyka R.

Poděkování

Poděkování patří všem trpělivým lidem, kteří mi s touto diplomovou prací přímo i nepřímo pomohli.

Obsah

Úvod

V této diplomové práci jsou představeny a vysvětleny vybrané regresní modely a longitudinální analýza. Tyto statistické metody jsou důležitým nástrojem při analýze dat, což může být užitečné i pro obor geoinformatiky.

Pro lepší představu o oblasti uplatnění těchto statistických metod jsou v rešerši uvedeny praktické ukázky. Dále jsou regresní modely matematicky popsány a vysvětleny na příkladech. Podrobněji jsou objasněny náročnější regresní modely. Poslední část teorie se zabývá longitudinální analýzou. Případové studie ověřují správné pochopení teorie.

Vybrané jsou následující čtyři typy regresních modelů: model lineární regrese (LM), model zobecněné lineární regrese (GLM), model lineární mixované regrese (LMM) a model zobecněné mixované lineární regrese (GLMM).

Metoda regrese slouží k odhadnutí hodnoty náhodné veličiny na základě znalosti veličin jiných. Například ráno odhadujeme hodnotu náhodné veličiny vyjadřující počasí přes den, když známe aktuální stav počasí. Dalším příkladem je zkoumání životnosti automobilu na základě zkušeností z minulých let.

Lineární regresní model, dále jen LM, slouží k popisu dvou veličin. Například odhadujeme-li výšku synů podle výšky otce. Data vstupující do lineárního regresního modelu musí splňovat předpoklad nezávislosti a normality, pokud není jeden z těchto předpokladů splněn, je třeba užít jiný typ regresního modelu.

Jiným typem regresního modelu, který již dokáže popsat více veličin, je zobecněný lineární model (GLM). V něm jsou vstupní data nekorelovaná a v ne-normálním rozdělení. Dalším modelem je lineární mixovaný model (LMM), kde se vyskytují korelovaná data v normálním rozdělení. A pro ne-normální data, která jsou korelovaná, se používá zobecněný lineární mixovaný model (GLMM).

Zobecněné lineární modely, dále jen GLM, slouží pro práci s daty, jejichž rozdělení je exponenciálního typu (normální, binomické, Poissonovo,...).

V lineárních mixovaných modelech, dále jen LMM, se zavádějí náhodné i neměnné efekty, proto mixované modely. Tyto modely jsou vhodné pro závislé náhodné veličiny.

Zobecněný lineární mixovaný model, dále jen GLMM se používá pro závislá data s vybraným typem exponenciálního rozdělení.

S těmito modely úzce souvisí i longitudinální analýza, ve které je důležitý časový interval, kdy byla data opakovaně sbírána, a stejná nebo velmi podobná skupina subjektů, která byla opakovaně měřena. Například sledování úbytku humrů za 10 let či poruchovost automobilů za 20 let. Hlavním parametrem longitudinální analýzy je čas, zachycený jako náhodná veličina v datech u každého subjektu měření. Pro výpočet longitudinální analýzy je možné využít všech regresních modelů uvedených výše.

1. Cíle práce

Na základě dostupné literatury a za pomoci dostupných materiálů bude sestavena diplomová práce na téma: „Modely pro diskrétní longitudinální data a jejich aplikace při vyšetřování dotazníků“.

První kapitola diplomové práce bude věnována rešerši, která obeznámí čtenáře s využitím regresních modelů, zvláště pak těch složitějších, jako jsou mixované lineární modely (LMM) a generalizované mixované lineární modely (GLMM). V rešerši budou také uvedeny praktické příklady využití longitudinálních dat a jejich aplikace v geoinformatice.

V teoretickém bloku budou vysvětleny složitější regresní modely se zvláštním zaměřením na zobecněné mixované lineární modely GLMM. Zde bude také představena longitudinální analýza.

V praktické části se na datech získaných dotazníkovým šetřením demonstruje využití těchto statistických metod. Výsledky těchto statistických metod budou také vhodně vyjádřeny a interpretovány.

Na závěr práce bude připojeno resumé v anglickém jazyce.

O diplomové práci bude vytvořena webová stránka v souladu s pravidly dostupnými na stránkách katedry geoinformatiky.

Výstupy budou odevzdány v digitální podobě na CD - ROM. A budou také odevzdány údaje o všech datových sadách, které byly vytvořeny nebo získány v rámci diplomové práce, pro potřeby zaevidování do Metainformačního systému katedry geoinformatiky ve formě vyplněného dotazníku.

2. Rešerše

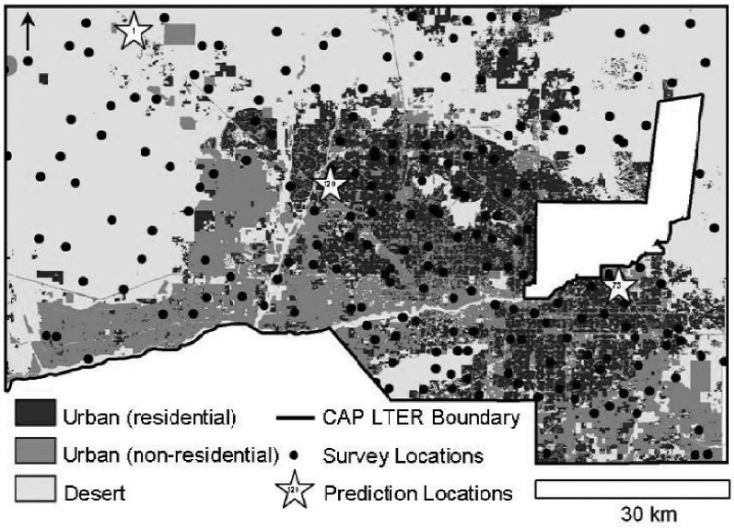
Jako důkaz využitelnosti regresní a longitudinální analýzy v oboru geoinformatiky jsou v této rešerši uvedeny studie, které poskytnou více praktických situací, kde se dají generalizované mixované lineární modely (GLMM) a longitudinální data použít. Využití regresní analýzy je velmi populární u veškerých oborů, kde se pracuje s daty, ve kterých jsou zachyceny vztahy náhodných veličin. Obor geoinformatiky není výjimkou, zde je tato analýza ještě prospěšnější, protože pomocí prostorových informací se dají výsledky lépe prezentovat. U většiny oborů, jako je chemie či fyzika, jsou výsledkem grafy, které je nutné zdlouhavě okomentovat. Díky prostorovým informacím lze vytvořit například mapu území, na kterém se hodnoty nasbíraly. Longitudinální analýza, se svým časovým hlediskem, je v oboru geoinformatiky také dobře využitelná, protože se s její pomocí dají lépe zhotovit časoprostorové analýzy dat.

První studie jménem „Disentangling complex fine-scale ecological patterns by path modelling using GLMM and GIS“ od V. Bakkestuen používá GLMM modely s GIS pro komplexnější pochopení ekologických systémů. Byly zkoumány tvary krajiny a jejich vztahy k určitým rostlinám. Byla použita data, která v sobě zahrnovala 140 druhů pozorování trvalek rostoucích v norských jehličnatých lesech za 11 let (1992 - 2002). Měřeny byly různé fyzické parametry rostlin a data také obsahovala topografické charakteristiky, jako je sklon či konvexnost. Pomocí metody GLMM se vytvářely odhady parametrů a modely, jejichž výsledkem bylo, že tvar a velikost rostliny záleží na sklonu svahu. [5]

Další studie, která se zabývá studiem výskytu či tvaru rostlin, které jsou pod vlivem okolí, je „Seedling interactions in a tropical forest in Panama“ od J. C. Svenning. Zde jsou navíc srovnávány metody uchycení dat pomocí modelů LMM a GLMM. V této studii byly také měřeny fyzické vlastnosti rostlin a topografické charakteristiky, ale s tím rozdílem, že se bral v potaz i další faktor. Což byla velikost a četnost jiných rostlinných druhů, které rostly poblíž. Výsledkem bylo, že topografické okolí má celkově větší vliv na velikost rostliny. Výskyty sousedních rostliny měly vliv menší.[29]

V článku z roku 2011 „A non-stationary spatial generalized linear mixed model approach for studying plant diversity“ od Anandamayee Majumdar se zkoumá druhová variabilita rostlin na daném území pomocí prostorových modelů GLMM. Zde jsou nejzajímavější mapové výstupy vytvořené pomocí softwaru ArcGIS. Zkoumanou oblastí je město Phoenix v USA, které je obklopeno pouští. Na rozdíl od předešlých dvou studií jsou zahrnuty socio-ekonomické faktory, které k urbanizované oblasti patří. Časový interval měření byl vždy po pěti letech. První byl v roce 2000, druhý 2005 a třetí, se kterým se v této studii ještě nepočítalo, byl v roce 2010. Data byla získána pomocí dálkového průzkumu země a z regionálních modelů. Na 1. obrázku je vidět město Phoenix s vyznačenými oblastmi, které se dělí na obydlené části, Urban (residential), neobydlené části,

Urban (non-residential), a poušť, Desert. Body měření jsou vyznačeny černou tečkou, Survey Location. Výsledky studie jsou prezentovány na 2. obrázku. Horní levá mapa prezentuje výsledek Shannon Weiner Diversity metric (SWDM), což je jedna z metod výpočtu GLMM. Pro horní pravou mapu byla použita metoda směrodatné odchylky pro SWDM. Spodní levá mapa ukazuje na rozmanitost rostlinných druhů v oblasti a spodní pravá mapa prezentuje směrodatnou odchylku druhové rozmanitosti. Světle šedá popisuje nízký výskyt, zatímco tmavě šedá ukazuje vysokou intenzitu výskytu. [16]

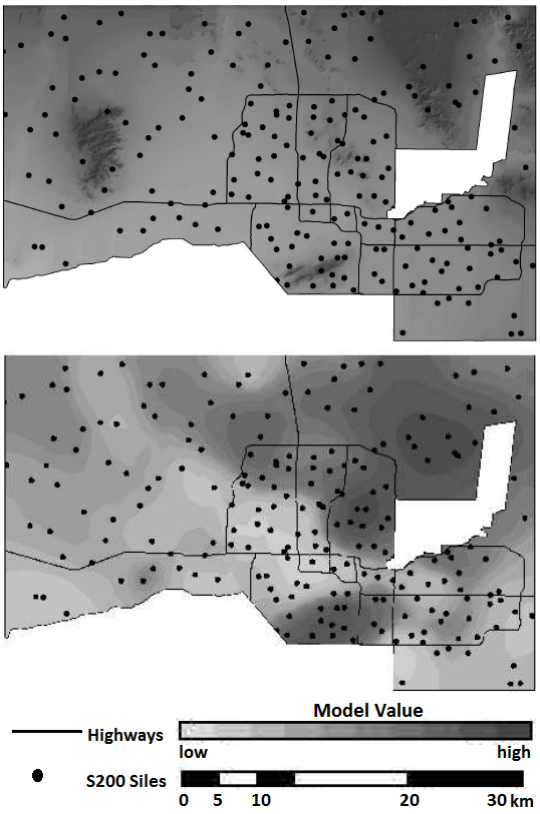


Obrázek 1.: Město Phoenix v USA.

Generalizovaný mixovaný lineární model byl také použit při studiu velikosti humrů v závislosti na okolních podmínkách v Jižní Africe. Článek se jmenuje „Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster, Jasus lalandii“ od A. Brandão. Zde se studoval každoroční pokles či nárůst velikosti humrů již od roku 1970. Zajímavá je datová sada, která byla rozsáhlá, ne-rovnovážná a která obsahovala značné množství náhodných efektů. Jedinou metodou, jak zpracovat takováto data, byla metoda GLMM. Výsledkem byla i mapa kvantitativního výskytu humrů.[7]

Výborný článek, který shrnuje metodu GLMM na analýze ekologických dat je „Generalized linear mixed models: a practical guide for ecology and evolution“ od Benjamin M. Bolker. Zde jsou vysvětleny jednotlivé kroky, které vedou k výpočtu a k ohodnocení GLMM modelů. Na praktické ukázce je demonstrováno užití anovy, výpočet odhadu parametrů, inference a testování hypotéz. Zásadní je také seznam pojmů a jejich zkratek, se kterými se dá setkat při analýzách GLMM. Je zde také názorně ukázaná vybavenost jednotlivých statistických softwarů pro práci s těmito modely.[6]

Ještě uveďme pár studií, které spojují šetření a vyhodnocování dotazníků a modely GLMM. Jednou z mnoha může být studie s názvem „The Smoking Con-



Obrázek 2.: Výsledky prostorových GLMM analýzy pro město Phoenix.

sequences Questionnaire: Factor structure and predictive validity among Spanish-speaking Latino smokers in the United States“ od Jennifer Irvin Vidrine z roku 2009, která pomocí dotazníkové formy zkoumá návyky spojené s kouřením na španělsky mluvících Latino - Američanech v USA. Jedním z výsledků bylo, že tato skupina přestává s kouřením velmi obtížně a pokud s kouřením přestanou, tak valná většina začne s kouřením znovu do 5 až 12 týdnů.[30]

Další zajímavou studií s využitím dotazníkového šetření a GLMM může být, „Possible influence of neighbours on stereotypic behaviour in horses“ od Krisztina Nagy z roku 2008. Tento článek se zaměřuje na stereotypní chování koní v závislosti na jejich ustájení a okolí, kde žijí, a agresivitu, kterou může stereotyp vyvolávat. V devíti jezdeckých školách (náhodný efekt) byl vyplněny dotazníky, které nesly data o 287 koních, a vypočítal se potenciál agrese. [21] Obor geoinformatiky by zde mohl vnést komplexnější pohled na okolí, kde se koně pohybují, naměřit určité charakteristiky okolí a vnést je do studie. Takto by se mohly naplánovat trasy, které by se koním mohly obměňovat, a zařídit, aby byly koně v co nejlepší kondici. Samozřejmostí by také byly mapové výstupy, třeba kvalitativní hodnocení okolí jednotlivých jezdeckých škol atd.

Longitudinální data se již vyskytla v předchozích příkladech, ale uveďme si ještě charakterističtější studie. První je „The ecology of childhood overweight: a 12-year longitudinal analysis“ od M. O'Brien, která popisuje vývoj dětské obezity během dvanácti let. Zkoumané děti byly v rozmezí od 2 do 12 let (960 subjektů) a zkoumalo se vzájemné působení kvality prostředí, ve kterém děti žijí, a výdej energie, který děti v takovém prostředí mají. Nezkoumal se rozdíl u pohlaví. Měřenými parametry byly „body mass index“ (BMI), který se naměřil sedmkrát po dobu studie, kvalita domácího prostředí, majetkové poměry rodičů a dětské zkušenosti v daných prostředích. Výsledkem bylo, že váha závisí spíše na domácím prostředí než na demografii. Děti, které byly obézní v předškolním věku, měly méně sensitivní matky než děti, které nikdy obézní nebyly. Děti, které trpěly obezitou ve školním věku, měly doma méně příležitostí na aktivity než děti, které nikdy obézní nebyly. Obézní děti se také více dívaly na televizi. Z této studie tedy obecně plyne, že pokud se dítě dívá po škole na televizi dlouho, má větší šanci být obézní do svých 12 let. Protože jsou zde mixované a náhodné efekty byla použita metoda GLMM. [22] Také zde by se daly prezentovat výsledky prostorově, třeba jako oblast s nejvyšší hustotou obézních dětí.

V další studii je měření prováděno každý druhý rok, tedy pro longitudinální analýzu se hodí i časy, které nejsou přímo následné. Pro longitudinální analýzu stačí minimálně dva časové údaje, abychom rozpoznali změnu u stejných nebo velmi podobných subjektů či pozorování. V Americe byla provedena studie „Longitudinal Monitoring of Public Reactions to the U.S“ od J.D. Absher, která si vzala za cíl zjistit, jak poplatky za vjezd do parků a služby ovlivňují venkovní rekreace. Data jsou z let 1999, 2001 a 2003 a poskytla je Lesní služba Spojených Států. Výsledkem bylo, že po mírném nárůstu cen zůstává reakce lidí na venkovní aktivity stejná. Pokud byla cena vyšší, tak pokles zájmu byl minimální. Zde byl také použit model pro GLMM. [4] Pomocí GIS by se daly vytvořit mapy parků, které by zachycovaly, jak vysoké poplatky je třeba na jednotlivých místech zaplatit.

Asi nejznámější organizací, která se zabývá sběrem longitudinálních dat v dotazníkové či jiné podobě je „The National Center for Analysis of Longitudinal Data in Education Research“ (CALDER) pod institucí „American Institutes for Reasearch“. Národní centrum pro analýzu longitudinálních dat ve vzdělávání, dále jen CALDER, shromažďuje a vyhodnocuje data o studentech a učitelích během let za účelem lepšího vzdělávání v USA. V datech jsou obsaženy mimo jiné i informace o státu, vládní politice vybraného státu, sociální a ekonomické skupině žáků a učitelů atd. Což umožní vyvarovat se kritickým problémům ve vzdělávání. Jejich webové stránky plně představují problematiku longitudinálních dat i s příkladem a s mnoha články. [17]

Jak je vidět z příkladů, jsou longitudinální a regresní analýzy velmi provázané.

3. Typy regresních modelů

V této kapitole se matematicky definují jednotlivé regresní modely (LM, GLM, LMM a GLMM). Ke každému modelu je vypočítán příklad v jazyku R.

3.1. Lineární regresní model (LM)

Klasický lineární regresní model je jedním z nejjednodušších typů regresních modelů. Do lineárního regresního modelu vstupují dvě proměnné, jedna je vysvětlující (známá, měřená) proměnná či kovariát a druhá vysvětlovaná (odhadovaná či cílová) proměnná. Příkladem klasického lineárního regresního modelu je analýza kovariance, která sleduje znak nebo vlastnost objektu a popisuje jejich změnu chování při změně podmínek. Například jak se budou vznášet různá tělesa v kapalině.

Pro zjištění vysvětlované proměnné, kde je n měřených objektů, J typů ošetření a k skupin, se použije tento model:

$$Y_{jk} \;=\; \mu \;+\; \alpha_j \;+\; \beta x_{jk} \;+\; \varepsilon_{jk},$$
$$j = 1, \ldots, J;\; k = 1, \ldots, n_j;\; n_1 \;+\; n_1 \;+ \cdots + \; n_n \;=\; n.$$

Y_{jk} ... odezva na k-tém objektu s typem ošetření j,
μ ... celkový průměr všech pozorování,
α_j... efekt typu ošetření j,
x_{jk}... hodnota kovariátu pro jk - tý objekt,
β ... regresní koeficient,
ε_{jk} ... náhodná odchylka.

Rovnice pro všechna n pozorování je v tomto tvaru:

$$\mathbf{Y} \;=\; \mathbf{X}\beta \;+\; \varepsilon,$$

$\mathbf{Y}$... vektor pozorování,
$\mathbf{X}$... matice plánu,
β... vektor parametrů,
ε... náhodná odchylka.

Tyto veličiny mají následující strukturu:

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{1n} \end{pmatrix},\; X = \begin{pmatrix} 1_{n1} & 1_{n1} & 0_{n1} & \cdots & 0_{n1} & x_{(1)} \\ 1_{n2} & 0_{n2} & 1_{n2} & \cdots & 0_{n2} & x_{(2)} \\ \vdots & \vdots & & \ddots & & \vdots \\ 1_{nj} & 0_{nj} & 0_{nj} & \cdots & 1_{nj} & X_{(j)} \end{pmatrix},\; \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_j \\ \beta \end{pmatrix}.$$

Příklad výpočtu LM:

Lineární regrese či lineární model se v prostředí implementace jazyka R vyvolá pomocí zkratky `lm` (linear model). Než bude představen příklad, je dobré si o této metodě něco přečíst v nápovědě (Help). Stačí napsat `?lm`.

```
## lm je funkce a uvnitř závorky jsou potřebné atributy
```

Model se vytvoří takto:

```
lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE,
x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset,
...)
```

Model se uchytí na příkladová data, která v sobě zahrnují informaci o věku v měsících a průměrnou výšku v centimetrech.[15]

```
## ukázka dat pomocí příkazu table
table(vek,vyska)
```

vyska/vek	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5
18	1	0	0	0	0	0	0	0	0	0	0	0
19	0	1	0	0	0	0	0	0	0	0	0	0
20	0	0	1	0	0	0	0	0	0	0	0	0
21	0	0	0	1	0	0	0	0	0	0	0	0
22	0	0	0	0	1	0	0	0	0	0	0	0

```
## pomocí funkce plot se vykreslí data do grafu
plot(vek,vyska)
```

```
## regresní lineární model je uložen do proměnné lm1
lm1<-lm(vyska ~ vek)
```

```
Call: lm(formula = vyska ~ vek)
```

```
Coefficients:
(Intercept) vek
   64.928   0.635
```

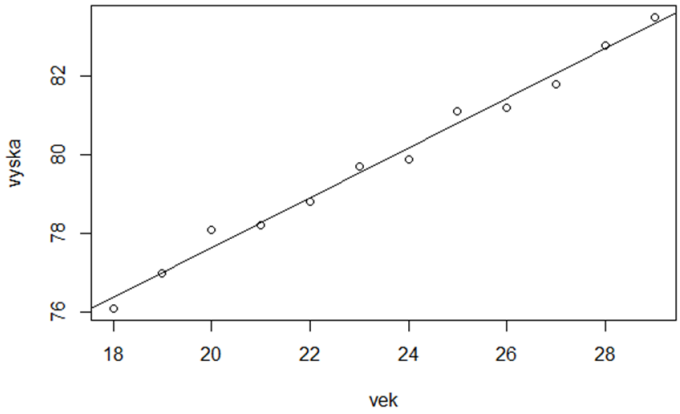
Výsledek:

Výsledkem modelu LM jsou dva koeficienty: Intercept (64,928) a sklon linie (0,635), která je nejlépe uchycena. Z těchto koeficientů lze vytvořit rovnici přímky, která má tuto podobu:

$$\text{výška} = 0.635 * \text{věk} + 64.928.$$

Z 1. grafu lze vyčíst, že s přibývajícím věkem se výška zvětšuje. Pomocí modelu se dají dopočítat i data, která nejsou změřena, nebo se dá odhadnout následující vývoj.

```
## proložení dat přímkou, pomocí funkce abline
abline(lm1)
```



Graf 1.: Znázornění dat pomocí funkce plot.

3.2. Lineární mixovaný model (LMM)

V předchozím modelu je jediný zdroj variability a to je náhodná složka ε. Pro více zdrojů variability, které je nutno popsat, lze použít lineární mixovaný model. Zde se jednotlivé objekty přiřazují do jedné z K skupin na základě určité vlastnosti (např. zařazení žáků do jednotlivých tříd). Pro výpočet LMM je použit tento model:

$$Y_{jk} = \mu + \alpha_j + U_k + \beta x_{jki} + \varepsilon_{jki},$$
$$j = 1, \ldots, J;\ k = 1, \ldots, K;\ i = 1, \ldots, n_{jk}.$$

Vysvětlení proměnných je stejné jako u předešlého modelu. Měřených objektů je n, typů ošetření je J, K je příslušnost objektu ke skupině a index i je pořadové číslo.

Rovnice pro všechna pozorování je v tomto tvaru:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZU} + \varepsilon,$$

$\mathbf{X}$ ... matice plánu pro neměnné efekty rozměrů $n * (J + 2)$,
β ... vektor parametrů,
$\mathbf{Z}$ ... matice plánu pro náhodné efekty rozměrů $n * K$,
$\mathbf{U}$ ... vektor náhodných efektů délky K,
ε ... vektor náhodných složek délky n.

Veličiny Z a U mají tuto strukturu:

$$Z = \begin{pmatrix} 1_{n1} & 0_{n1} & \cdots & 0_{n1} \\ 0_{n2} & 1_{n2} & \cdots & 0_{n2} \\ \vdots & \ddots & & \vdots \\ 0_{nK} & 0_{nK} & \cdots & 1_{nK} \end{pmatrix}, \; U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_K \end{pmatrix}.$$

Příklad výpočtu LMM:

Funkce `lmer`, která je pro tento výpočet použita, je obsažena v balíku lme4 či v lme4a(b), který si je nutný stáhnout. Další informace v nápovědě `?lmer`.

```
library„lme4"
install.packages("lme4a",repos="http://r-forge.r-project.org")
```

Balík lme4 obsahuje 4 metody pro uchycení dat: lmer, lmer2, glmer a nlmer. Protože data pro model LMM jsou v normálním rozdělení, lze použít všechny metody, pokud je parametr „family“ nastaven na hodnotu „gaussian“.

```
lmer(formula, data, REML = TRUE, sparseX = FALSE, control = list(), start =
NULL, verbose = 0L, doFit = TRUE, compDev = TRUE, subset, weights, na.action,
offset, contrasts = NULL, ...)
```

```
lmer2(formula, data, family = NULL, REML = TRUE,control = list(), start =
NULL, verbose = FALSE,subset, weights, na.action, offset, contrasts = NULL,
model = TRUE, x = TRUE,...)
```

```
glmer(formula, data, family = gaussian, start = NULL,verbose = FALSE, nAGQ
= 1, doFit = TRUE, subset, weights,na.action, offset, contrasts = NULL, model
= TRUE,control = list(),...)
```

```
nlmer(formula, data, start = NULL, verbose = FALSE, nAGQ = 1,doFit = TRUE,
subset, weights, na.action,contrasts = NULL, model = TRUE, control = list(),
...)
```

Data, která jsou pro tento příklad vybrána, obsahuje balík MEMSS, který je třeba si stáhnout. Data mají název „Rail“ a popisují rychlost zvukové vlny (travel), která se šíří od šesti typů železničních tratí (Rail).[14]

```
library(„MEMSS")
str(Rail)
'data.frame': 18 obs. of 2 variables:
Rail : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 2 2 2 3 3 3 4 ...
travel: num 55 53 54 26 37 32 78 91 85 92 ...
```

```
print(dotplot(reorder(Rail,travel)~travel,Rail,
xlab="Rychlost šíření (ms)",ylab="Typ tratě"))
```

```
## uchycení modelu, proč je formule zadána v této podobě, bude objasněno
## v následující kapitole
```

```
lmm1 <-lmer(travel ~ 1 + (1 | Rail),Rail, REML = FALSE)
```

```
Linear mixed model fit by maximum likelihood ['merMod']
Formula: travel ~ 1 + (1 | Rail)
   Data: Rail
  AIC      BIC    logLik deviance
134.5600 137.2312 -64.2800 128.5600
```

```
Random effects:
Groups    Name        Variance Std.Dev.
Rail    (Intercept)  511.86   22.624
Residual              16.17    4.021
Number of obs: 18, groups: Rail, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept) 66.500    9.285    7.162
```

Výsledek:

Důležitá je variabilita mezi daty, kterou tento model poskytuje. Variance (rozptyl) ukazuje na podobnost či nepodobnost dat v modelu. Standard deviation (směrodatná odchylka) je druhá odmocnina variance. Ta se dá lépe zobrazit než variance. Residual je část variability, která nemůže být modelována nebo vysvětlena.

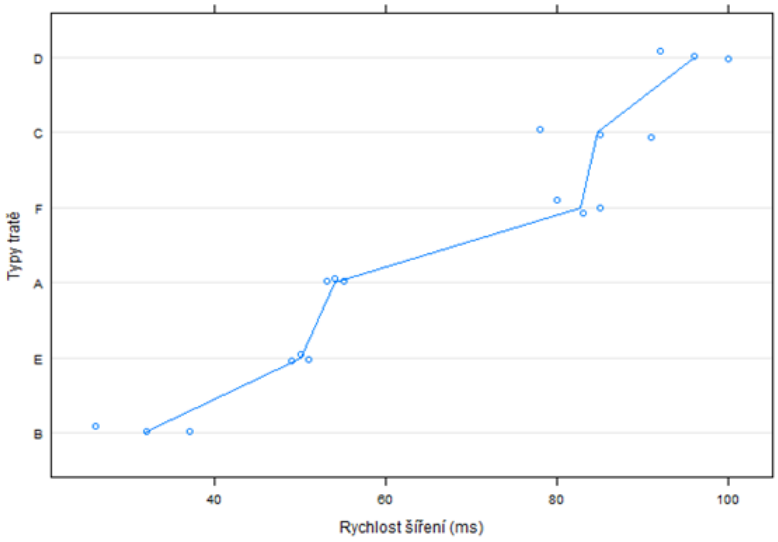
Zde lze z výsledků funkce lmer sestavit rovnici modelu, která spojitě charakterizuje průběh dat. Pro uchycení modelu se použije Intercept a sklon.

Výsledkem je také průměrná počáteční rychlost zvukové vlny podle typů trati, 66.500 (Intercept). Rovnice modelu:

$$\text{travel} = 22.624 * \text{Rail} + 66.500.$$

Z 2. grafu lze vyčíst rychlost zvukové vlny podle typů tratě. A pomocí rovnice modelu se odhadují a dopočítávají chybějící data a trend.

```
print(dotplot(reorder(Rail, travel) ~ travel, Rail,ylab = "Typy tratě",
jitter.y = TRUE, pch = 21,xlab = "Rychlost šíření (ms)",type = c("p", "a")))
```



Graf 2.: Rychlost šíření vln podle typů tratě.

3.3. Zobecněný lineární model (GLM)

Tento model je vhodný pro data, která nesplňují podmínku normálního rozdělení. V tomto modelu se pracuje s některým typem exponenciálního rozdělení, což je Poissonovo, binomické, gamma atd. Příladem GLM je logistická regrese.

Na n měření se aplikuje J typů ošetření a k skupin, neuvažuje se přiřazení do skupin pomocí indexu i jako u LMM.

Pro GLM je použit tento model:

$$E_{(Yjk)} = exp(\mu + \alpha_j + \beta x_{jki}),$$
$$j = 1, \ldots, J;\ k = 1, \ldots, n_j;\ n_1 + n_2 + \ldots + n_J = n.$$

Příklad výpočtu GLM:

V tomto výpočtu je použita funkce glm.

```
glm(formula, family = gaussian, data, weights, subset,na.action, start =
NULL, etastart, mustart, offset, control = list(...), model = TRUE, method =
"glm.fit",x = FALSE, y = TRUE, contrasts = NULL,...)
```

Data, která jsou v tomto příkladu použita, lze nalézt http://data.princeton.edu/wws509/datasets/cuse.dat. V datech jsou proměnné jako věk (age), vzdělání (education), zda chtějí dále studovat (wantsMore) a notUsing a using jsou proměnné, na které se budou předchozí proměnné vztahovat.[24]

```
cuse <-read.table("http://data.princeton.edu/wws509/datasets/
cuse.dat",header=TRUE)
## ukázka vybraných dat cuse
```

	age	education	wantsMore	notUsing	using
1	<25	low	yes	53	6
2	<25	low	no	10	4
3	<25	high	yes	212	52
4	<25	high	no	50	10
5	25-29	low	yes	60	atd.

```
## Jaké je rozložení dat, lze poznat z histogramu. Rozdělení je binomické.
hist(cbind(using, notUsing),xlab = "Užívají-neužívají", ylab="Četnost",
main = paste("Četnost užívají-neužívají"))
```

```
glm1<- glm(cbind(using, notUsing) ~ age + education + wantsMore , family =
binomial)
```

```
Call: glm(formula = cbind(using, notUsing) ~ age + education + wantsMore,
family = binomial)
```

```
Coefficients:
 (Intercept) age25-29 age30-39 age40-49 educationlow wantsMoreyes
 -0.8082    0.3894    0.9086    1.1892    -0.3250    -0.8330
```

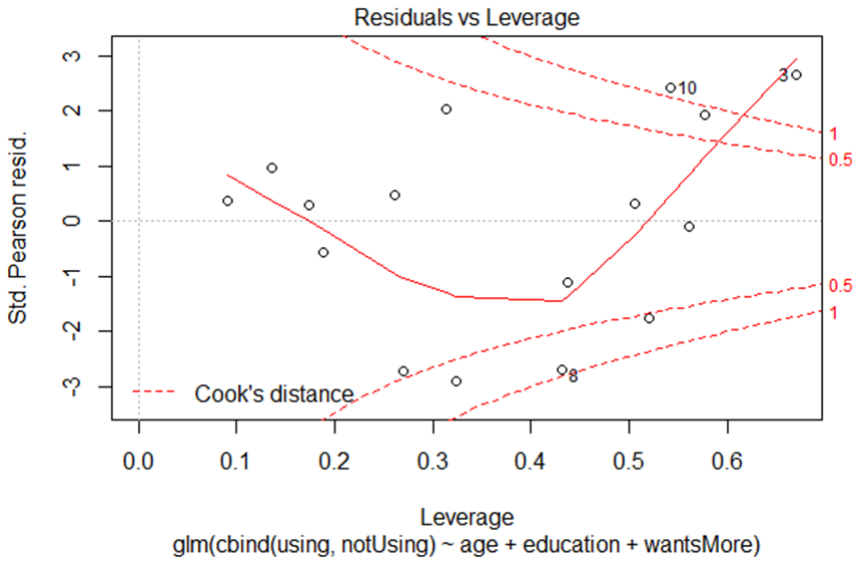
```
Degrees of Freedom: 15 Total (i.e. Null); 10 Residual
Null Deviance:      165.8
Residual Deviance: 29.92 AIC: 113.4
```

Výsledek:

Pomocí funkce `plot(glm1)` se zobrazí čtyři grafy. Ty popisují pomocí přímek, které prostupují daty, jednotlivé charakteristiky modelu. Nejdůležitější je 3. graf, který celistvou linií popisuje data se všemi nezávislými faktory.

Rovnice modelu:

$$\text{cbind(using,notUsing)} = 0.3894{*}\text{age25-29} + 0.9086{*}\text{age30-39} + 1.1892{*}\text{age40-49} + \\ +(-0.3250){*}\text{educationlow} + (-0.8330{*}\text{wantsMoreyes}) + (-0.8082).$$



Graf 3.: Přímka (plná) vyjadřující směrodatnou odchylku a residual.

3.4. Zobecněný mixovaný lineární model (GLMM)

Model GLMM je určen pro data, která nesplňují podmínku normality a nezávislosti. Data mají rozdělení exponenciálního typu a jsou závislá.

Je opět n měření, J možných typů ošetření (vybraných náhodně) a měření lze přiřadit do K skupin podle indexu i. Známá hodnota vysvětlující proměnné je x_{jki} a odhaduje se vysvětlovaná proměnná Y_{jki}.[1][2]

Model se vyjádří takto:

$$E_{(Yjki)} = \mu_{jki} = exp(\mu + \alpha_j + U_k + \beta x_{jki}),$$
$$j = 1, \ldots, J;\ k = 1, \ldots, K;\ i = 1, \ldots, n_{jk}.$$

Uchycení dat:

1. Pomocí balíku lme4 (lme4a, lme4b),

Funkce lmer, lmer2, glmer a nlmer (viz. LMM).

2. Pomocí balíku glmm

Pro výpočet generalizovaného lineárního mixovaného modelu, je vhodný i balík glmm s jednou funkcí glmm, která je velmi podobná funkci glmer. Tato funkce používá normální mixované rozdělení spočítané pomocí Gauss–Hermite integrace. Opět se nastaví family na příslušné rozdělení.

```
glmm(formula, family=gaussian, data=list(), weights=NULL,offset=NULL,
nest, delta=1, maxiter=20, points=10, print.level=0,
control=glm.control(epsilon=0.0001,maxit=10,trace=FALSE))
```

3. Pomocí balíku glmmPQL

Tento balík slouží také pro výpočet GLMM, ale s pomocí pravděpodobnosti PQL (Penalized Quasi–Likelihood). Podmínkou je nastavení family na správné rozdělení.

```
glmmPQL(fixed, random, family, data, correlation, weights, control,
niter = 10, verbose = TRUE,...)
```

Příklad výpočtu GLMM:

Data pocházejí ze studie „Beat the Blues“, která zkoumá chování jedinců trpících depresí v průběhu dní (longitudinální analýza). Data se jmenují BtheB a jsou připraveny v balíku HSAUR2. Každý jedinec podle určité stupnice v dotazníku zapisoval, jak se v daný den cítí. Proměnné, které se zde vyskytují, jsou Léky (drug), Doba léčení (length), Léčba (treatment, ma dva faktory TAU a BtheB), Předešlá léčba (bdi.pre) , Subjekt (subject), Čas (time) a Nynější léčba (bdi). [9]

```
install.packages("lme4a",repos="http://r-forge.r-project.org")
data("BtheB", package = "HSAUR2")

##data musela být před-připravena a uložena do BtheB_long
 drug length treatment bdi.pre subject time bdi
1.2m No >6m TAU 29 1 2 2
2.2m Yes >6m BtheB 32 2 2 16

## vytvoření modelu pomocí funkce lmer
BtheB_lmer1 <- lmer(bdi  ~  bdi.pre + time + treatment + drug + length +
(1 | subject), data = BtheB_long, REML = FALSE, na.action = na.omit)
```

```
## vytvoření modelu pomocí funkce glmmPQL
BtheB_glmmPQL1 <- glmmPQL(fixed = value ~ treatment + time + drug +
length + bdi.pre, random = 1 | subject, family = gaussian, data = btheb)

## vytvoření modelu pomocí funkce glmm
BtheB_glmm1 <- glmm(value ~ treatment + time + drug + length + bdi.pre,
nest = subject,data = na.omit(btheb))
```

Výpočty u všech modelů dopadly stejně, proto postačí podívat se jen na výsledek funkce lmer.

```
BtheB_lmer1

Linear mixed model fit by maximum likelihood
Formula: bdi ~ bdi.pre + time + treatment + drug + length + (1 |
subject)
Data: BtheB_long
AIC BIC logLik deviance REMLdev
1887 1916 -935.3 1871 1866

Random effects:
Groups Name Variance Std.Dev.
subject (Intercept) 48.304 6.9501
Residual            25.128 5.0127
Number of obs: 280, groups: subject, 97

Fixed effects:
 Estimate Std. Error t value
(Intercept) 5.94371 2.24911 2.643
bdi.pre    0.63819 0.07759 8.225
time       -0.71703 0.14605 -4.909
treatmentBtheB -2.37311 1.66365 -1.426
drugYes      -2.79786 1.71990 -1.627
length>6m     0.25639 1.63210 0.157

Correlation of Fixed Effects:
(Intr) bdi.pr time trtmBB drugYs
bdi.pre -0.678
time -0.264  0.023
tretmntBthB -0.389  0.121  0.022
drugYes    -0.071  -0.237  -0.025  -0.323
length>6m  -0.238  -0.242  -0.043  0.002  0.158
```

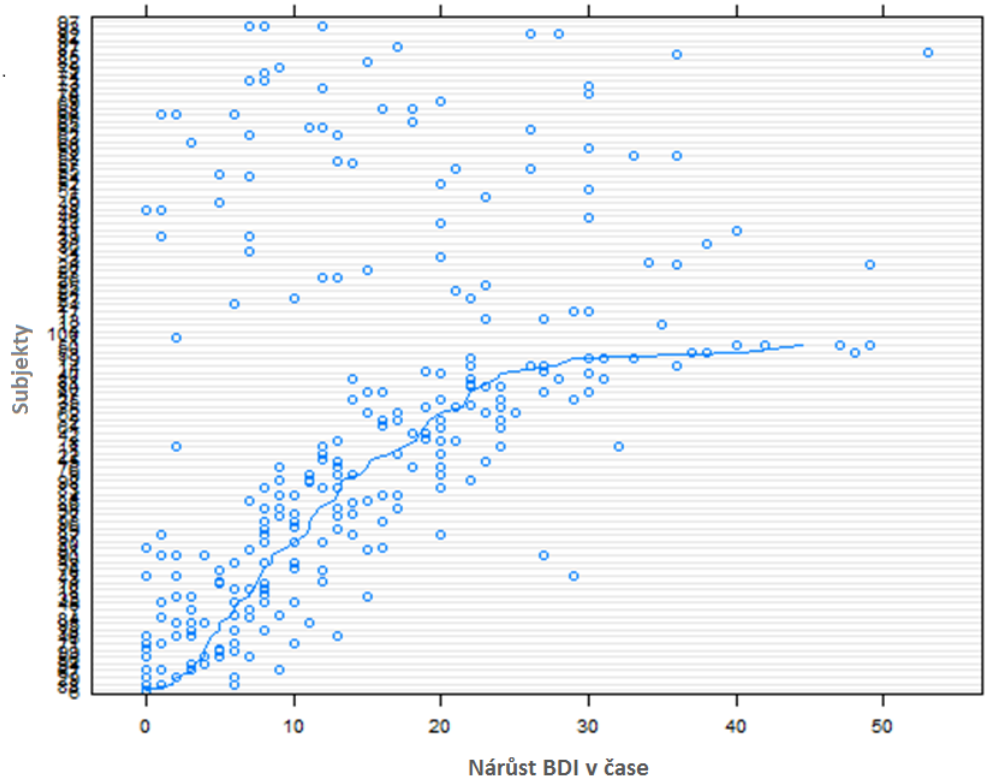
Výsledek:

Z výsledku funkce lmer lze odhadnout počáteční průměrnou hodnotu měření bdi (5.94371), bdi.pre (0.63819), času (-0.71703), léčby (-2.37311), léků (-2.79786) a doby léčení (0.25639). Velká variabilita u Subjektu znamená, že jsou si Subjekty málo podobné (48.304). Dokonce i Residual ukazuje na velkou část variability v datech, která nemůže být vysvětlena nebo modelována. Korelace vyjadřuje vztah mezi neměnnými efekty. Korelace je podrobněji vysvětlena v kapitole „Modely pro longitudinální data“.

Rovnice modelu je:

$$\text{bdi} = 6.9501{*}\text{subject} + 5.94371 + 0.63819{*}\text{bdi.pre} + (-0.71703){*}\text{time} + \\ + (-2.37311){*}\text{treatmentBtheB} + (-2.79786){*}\text{drugYes} + 0.25639{*}\text{length}(>6\text{m}).$$

```
print(dotplot(reorder(BtheB_long$subject,BtheB_long$bdi) ~ BtheB_long$bdi,
BtheB_long,ylab = "Subjekty", jitter.y = TRUE, pch = 21,xlab = "Nárůst hodnoty
BDI v čase",type = c("p", "a"))
```



Graf 4.: Rozmístění naměřených hodnot BDI a proložení přímkou.

4. Modely s mixovanými efekty – jeden náhodný efekt

Pro snadnější pochopení složitějšího generalizovaného lineárního mixovaného modelu (GLMM) je nejdříve vysvětlen jednodušší lineární mixovaný model (LMM). Ten pomůže objasnit základní principy chování modelu, které jsou ekvivalentní i pro GLMM. Vysvětleny jsou náhodné a neměnné efekty a pravděpodobnostní metody RELM a MLE. Pomocí příkladů bude popsán postup práce tvorby modelu a jeho výpočet v jazyku R na vybraných datových sadách.

Rozdíl mezi GLMM a LMM je jen v podmínce o ne-normalitě, oba modely mají tedy závislé veličiny. Proto se další práce s GLMM metodou bude lišit jen v použití jiného rozdělení, princip tvorby a popisu modelu je stejný.

Modely s mixovanými efekty popisují vztah mezi odezvou proměnné (faktoru či závislé proměnné) a kovariátu (nezávislá či vysvětlující proměnná) či více kovariátů, které byly naměřeny nebo pozorovány společně s danou odezvou. V modelech s mixovanými efekty je vždy alespoň jeden z kovariátů kategorickou reprezentací experimentální nebo pozorované jednotky v datové sadě. Příkladem může být chemie, kde je odpozorovaná jednotka dávka zkoumaného produktu používaného pro chemické pokusy. V medicíně nebo sociálních vědách je pozorovanou jednotkou člověk nebo zvíře a v zemědělství či geoinformatice je jednotkou například část země, kde se pěstují dané rostliny.

Ve všech těchto příkladech je kovariát v sadě s oddělenými úrovněmi (pohlaví - žena/muž, barva - modrá/zelená/bílá). Na popis úrovní se mohou použít číselné hodnoty, ale musí se mít na paměti, že to jsou pouze identifikátory a nelze s nimi provádět matematické operace jako s čísly. Nejdůležitější charakteristikou kategorického kovariátu je, že v každé naměřené pozorované hodnotě odezvy či reakce bude kovariátu, podle tohoto měření, přidělena pouze jedna hodnota ze sady jasně odlišitelných úrovní (př. subjekt1 byl naměřen jako muž). Parametry asociované s určitou úrovní kovariátu jsou nazývané jako efekty dané úrovně.

Pokud sada všech možných úrovní kovariátu je neměnná a reprodukovatelná modeluje se kovariát s pomocí parametrů, které jsou neměnnými efekty - „fixed effect“ (pohlaví vždy bude muž nebo žena, nezasahuje třetí hodnota). Pokud úrovně, které se pozorují, reprezentují náhodný či nahodilý vzorek z dané sady všech možných úrovní, tak jsou značeny jako náhodný efekt -„random effect“ (jeden člověk je jedním vzorkem z celého lidstva). Co je nutné si vysvětlit o neměnných a náhodných efektech parametrů, je to, že pojmenování může být občas zavádějící. Jelikož odlišnost mezi nimi je více v charakteru úrovní kategorického kovariátu než v charakteru souvisejících efektů. Dále se rozlišuje, že neměnné efekty parametrů jsou skutečnými parametry ve statistickém modelu a že náhodné efekty nejsou tak docela pravými parametry. Náhodné efekty jsou totiž nepozorovatelnými náhodnými veličinami.

Chceme-li modelovat, jak se v průběhu jednoho roku změnil výsledek testu psaní u studentů v jedné škole za předpokladu, že zaznamenané kovariáty (vysvětlující proměnné) zahrnují výsledek testů se studentským identifikačním číslem a pohlavím studenta. Výsledek testů, identifikátor i pohlaví studenta jsou kategorickými kovariáty a také neměnnými efekty. Pokud se vezmou do úvahy data i z jiných škol nebo se zahrnou i dřívější výsledky testů studentů, jsou vždy zachovány úrovně o pohlaví, tedy pohlaví je stále neměnným efektem. Avšak studenti a jejich identifikátory, jsou bráni jako vzorek ze sady všech možných studentů a identifikátorů, které je možno pozorovat, proto jsou náhodnými efekty. Náhodnými efekty jsou i výsledky testů.

Po shrnutí předchozích myšlenek lze tvrdit, že modely s mixovanými efekty jsou statistické modely, které zahrnují, jak parametry s neměnnými tak i náhodnými efekty. Přičemž modely s náhodnými efekty vždy zahrnují alespoň jeden parametr s neměnným efektem. Proto jsou modely s náhodným efektem nazývány také mixované modely.

Předchozí lze matematický popsat: pokud je q - dimenzionální vektor náhodných efektů zastoupený pomocí náhodných proměnných B a n - dimenzionální vektor odezvy reprezentovaný opět náhodnými proměnnými tentokrát Y. Pozorují se hodnoty y z vektoru Y, ale nedají se pozorovat hodnoty z B.

Při formulování modelu se popisuje nepodmíněné rozdělení náhodných proměnných B a podmíněné rozdělení $(Y \mid B = \mathbf{b})$. Tento popis podmíněnosti zahrnuje rozdělení a hodnoty určitých parametrů. Pozorované hodnoty odezvy a kovariáty budou použity k odhadnutí těchto parametrů a k vytvoření jejich dedukce.

Nyní se představí modelová ukázka, která je popsána v první kapitole o „Modelech s mixovanými efekty“. Tento výukový materiál připravila komunita open source jazyka R pro package lme4 http://lme4.r-forge.r-project.org/. Je možné si volně stáhnout kód tohoto příkladu, proto zde budou uvedeny jen zásadní části kódu, které je třeba podrobněji vysvětlit.

Knihovna pro lme4 obsahuje ukázkové datové sady Dyestuff a Dyestuff2. Tyto datové sady obsahují náhodné efekty, které jsou charakteristické svou variabilitou mezi dávkami v chemických procesech.

Mějme na paměti, že pokud je uvedena i podmínka ne-normality a je nutno použit GLMM, pak je princip stejný, jen se použije jiná hodnota ve „family“. Naopak pokud je podmínka jen pro nezávislost, použije se model GLM.

Nejdříve se použijí mixované modely LMM ke kvantifikaci variability ve Výnosu a mezi Dávkami s daty Dyestuff a Dyestuff2. Zde chceme zjistit, která Dávka zaručí vyšší nebo menší střední Výnos. Jinými slovy chceme zjistit, jaký bude mít Dávka „efekt“ na Výnos. Protože chceme predikovat Výnos i z budoucích Dávek, tak Dávky budeme pokládat za náhodný efekt. Dávky použité v experimentu jsou vzorky ze sady všech možných Dávek, které si přejeme zahrnout.

Dyestuff Data – obsahují 30 vzorků (pozorování) a 2 proměnné a to Dávku (Batch) se 6 úrovněmi a Výnos (Yield) s 30 úrovněmi. Proměnná Výnos je odezva (vysvětlované, závislé proměnné) kovariátu Dávka (nezávislá, vysvětlující proměnná).

Dyestuff2 Data – je velmi podobný Dyestuff, jen Výnos je uváděn s jinými hodnotami. Dávka je kategorickou proměnnou a je uložena v R prostředí jako faktor či štítek. Někdy implementace jazyka R rozpozná faktor automaticky, lze se na něj dotázat takto:

```
is.factor(Dyestuff$Batch)
[1] TRUE
```

Pokud je kategorická úroveň faktoru popsána čísly, zobrazí se FALSE a je nutné ji přepsat:

```
as.factor(Dyestuff$Batch)
```

Pokud by se to neudělalo, počítalo by se s numerickým kovariátem s nechtěnými následky.

Aby se nemusel prověřovat zvlášť každý atribut datové sady, zda je faktor, lze vyvolat souhrnný popis datové sady pomocí – `str(Dyestuff)`.

```
'data.frame': 30 obs. of 2 variables:
$ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
$ Yield: num 1545 1440 1440 1520 1580 ...
```

Pomocí funkce `summary(Dyestuff)` jde zjistit, zda jsou data vyvážená. To nastává, pokud jsou jednotlivé kovariáty, zde Dávky, stejně nebo podobně početné pro každou z neměnných úrovní (A:5 až F:5).

4.1. Popis modelu a metody odhadu parametrů

Na příkladová data Dyestuff a Dyestuff2 se upevnil model pomocí funkce lmer, což je jeden z možných modelů v balíku lme4.

MODEL PRO DYESTUFF

```
fm1 <- lmer(Yield ~  1 + (1|Batch), Dyestuff, REML=TRUE)
```

Do proměnné fm1 se uložil model vytvořený pomocí funkce lmer. Ten obsahuje formuli ve tvaru Yield $\sim$ 1 + (1|Batch), název datové sady Dyestuff a REML nastavený na hodnotu TRUE. Hodnota „1“ označuje lineární komponentu a (1|faktor) je seskupující faktor (viz následující kapitola).

Ačkoli originální příklad i nápověda v R uvádí, že REML je defaultně TRUE, tak při mých pokusech se nezobrazily výsledky shodné s tímto modelovým příkladem, pokud jsem do kódu nepřipsala REML=TRUE. Pokud definování REML

chybělo, model se choval jako s REML=FALSE. Jestli nepomůže ruční nastavení pravdivostní hodnoty RELM, je lepší restartovat počítač, protože R si uchovává v paměti předešlé výpočty a nepomůže ani restartování programu. Některé příklady fungují i s REML=0.

REML je zkráceně **Restricted maximum likelihood** a používá se k omezeným odhadům maximální pravděpodobnosti. Používá pravděpodobnostní funkci spočítanou z transformovaného datasetu tak, že nevýznamné parametry nemají efekt. V takovém případě nejsou variační komponenty odhadnuty z originálního datasetu, ale ten je nahrazen sadou kontrastů vypočítaných z dat a pravděpodobnostní funkce. Ta je vypočítána z pravděpodobnostního rozdělení kontrastů. REML se velmi často užívá jako jedna z metod v uchycení lineárního mixovaného modelu. Pokud je RELM = FALSE, model je počítán pomocí maximální pravděpodobnosti MLE.[27]

```
Linear mixed model fit by REML          1. Popis modelu
Formula: Yield  ~  1 + (1 | Batch)      2. Formule, data a metoda
Data: Dyestuff                             odhadu pravděpodobnosti
 REML
319.7

Random effects:                         3. Náhodné efekty
Groups Name Variance Std.Dev.
Batch ( Intercept) 1764.0 42.001
Residual           2451.3 49.510
Number of obs: 30, groups: Batch, 6

Fixed effects:                          4. Neměnné efekty
        Estimate Std. Error t value
(Intercept) 1527.50 19.38 78.8
```

1. Popis modelu, zde je název použitého modelu, ve kterém jsou parametry odhadnuty s minimálním REML.

2. Formule a Data jsou zobrazeny jako zmínka, pro lepší orientaci ve výsledcích. **REML** odhaduje varianci (odlišnost) komponentů (složek). Pokud REML=TRUE, tak je model upevněn pomocí REML, vytiskne se tedy hodnota REML, pokud je hodnota REML=FALSE, vypadá kód takto:

```
fm1ML <- lmer(Yield ~ 1 + (1|Batch), Dyestuff, REML = FALSE)

Linear mixed model fit by maximum likelihood1. 1. Popis modelu
Formula: Yield  ~  1 + (1 | Batch)             2. Formule, data a metoda
Data: Dyestuff                                    odhadu pravděpodobnosti
    AIC   BIC  logLik deviance REMLdev
 333.3    337.5    -163.7    327.3
```

```
Random effects:                                    3. Náhodné efekty
 Groups Name Variance Std.Dev.
 Batch (Intercept) 1388.3 37.26
 Residual          2451.3 49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:                                     4. Neměnné efekty
            Estimate Std. Error t value
(Intercept) 1527.50    17.69   86.33
```

Nyní je model uchycen pomocí maximální pravděpodobnosti.

1. Popis modelu, zde je název použitého modelu, ve kterém jsou parametry odhadnuty pomocí maximální pravděpodobnosti.

2. Formule a Data jsou zobrazeny jako zmínka, pro lepší orientaci ve výsledcích.

AIC – Akaikeho informační kritérium (vhodnost modelu).
BIC – Schwarz-Bayesianovo informační kritérium (vhodnost modelu).
logLik – log-likelihood – logaritmická pravděpodobnost odhadovaných parametrů.
deviance – odchylka (dvakrát negativní log-likelihood) odhadovaných parametrů.
AIC, BIC, logLik a deviance jsou statistiky spojené s upevněním modelu a jsou použitelné ke srovnání různých modelů k uchycení stejných dat.

Při srovnání více modelů je bezpečnější upevnit všechny modely pomocí maximální pravděpodobnosti tedy s REML = FALSE.

Maximum likelihood estimation – MLE, je metoda odhadu parametrů ve statistickém modelu. Tato metoda se používá, pokud není dostatečný počet měření. Hodnoty by měly mít gaussovo (normální) rozdělení. Průměr a variance budou z MLE odhadnuty i z malého počtu měření. MLE v tomto případě vezme střední hodnotu a varianci jako parametry a hledá konkrétní parametrické hodnoty, které určí pozorované výsledky jako nejpravděpodobnější.

Pro neměnná data a základní statistický model metoda vybírá hodnoty, které způsobují rozdělení dané pozorovanými daty s nejvyšší pravděpodobností (parametry maximalizují pravděpodobnostní funkci). MLE tedy podá jednotný odhad, který je nejlépe definovaný v daném případě s gaussovým (normálním) rozdělením a malým počtem základního měření.[20]

3. Náhodné efekty, zde lze vidět dva zdroje variability v upevněném modelu: variabilitu Dávky (Batch) k Dávce (Batch) v úrovni odezvy a Residual (pozůstatek).

Residual je část variability, která nemůže být vysvětlena nebo modelována s ostatními výrazy. Je to změna či odchylka od pozorovaných dat, která zbyla po určení odhadů parametrů v dalších částech modelu. Část této variability je v odezvě asociována s výrazy neměnných efektů.[8]

Intercept (průsečík) reprezentuje „typickou“ či střední úroveň odezvy v daném případě.[28]

Standard deviation (směrodatná odchylka) je další odhad variability. Směrodatná odchylka říká, jak moc se od sebe liší případy, které jsou typické ve zkoumaném souboru. Prvky jsou si tedy podobné, pokud je směrodatná odchylka malá a velká směrodatná odchylka vypovídá o velkých vzájemných odlišnostech. Také lze přibližně zjistit, jak daleko jsou hodnoty v souboru vzdáleny od průměru (střední hodnoty). Hodnota směrodatné odchylky je odmocnina variance (rozptylu).

Směrodatná odchylka se používá, protože se dá lépe prezentovat než variance (rozptyl). Směrodatná odchylka má měřítko odezvy avšak rozptyl má magnitudu.

Variance (rozptyl) popisuje variabilitu rozdělení pravděpodobnostní náhodné veličiny. Stejně jako směrodatná odchylka popisuje, jak si jsou hodnoty v souboru podobné.

Standard Error (střední chyba) je směrodatná odchylka výběrového rozdělení.[25]

Poslední řádek je počet pozorování a úrovně seskupeného „grouping“ faktoru. V obou modelech je použit jeden seskupující faktor (1|Batch), dohromady 6 náhodných efektů – Dávek.

4. Neměnné efekty, zde jsou vytištěny odhady a směrodatné chyby pro neměnné efekty. Jediným neměnným efektem pro model fm1 je „1“ z formule `Yield ~ 1 + (1|Batch)`.

Tento příklad měl objasnit rozdíl v použití metod REML A MLE. Pokud chceme vybrat ten lepší model z fm1 či fm1ML, lze je porovnat pomoci anovy, příkaz `anova(fm1,fm1ML)`, což je vysvětleno v další kapitole.

MODEL PRO DYESTUFF2

Příklad, kdy je variance a směrodatná odchylka rovna 0 v obou případech REML= TRUE/FALSE. Porovnání modelů:

```
fm2 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff2, REML=TRUE)

Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
Data: Dyestuff2
REML
161.8
```

```
Random effects:
Groups Name Variance Std.Dev.
Batch (Intercept) 0.000 0.0000
Residual          13.806 3.7157
Number of obs: 30, groups: Batch, 6

Fixed effects:
       Estimate Std. Error t value
(Intercept) 5.6656 0.6784 8.352

fm2ML <- update(fm2, REML = FALSE)

Linear mixed model fit by maximum likelihood
Formula: Yield  ~  1 + (1 | Batch)
Data: Dyestuff2
AIC BIC logLik deviance
168.9 173.1 -81.44 162.9

Random effects:
Groups Name      Variance Std.Dev.
Batch (Intercept) 0.000 0.0000
Residual          13.346 3.6532
Number of obs: 30, groups: Batch, 6

Fixed effects:
      Estimate Std. Error t value
(Intercept) 5.666 0.667 8.494
```

Pokud je variance a směrodatná odchylka rovna 0, tak to neznamená, že neexistuje variance mezi skupinami. Jen tento případ jednoduše indikuje, že „meziskupinová“ úroveň variability není schopna zahrnout náhodné efekty v modelu a se musí podstoupit numerická optimalizace.

4.2. Intervaly spolehlivosti

Parametry odhadu statistických modelů reprezentují „nejlepší odhad“ v neznámých hodnotách modelovaných parametrů, proto jsou tak důležité výsledky statistického modelování. Modely charakterizují variabilitu v datech a je nutné zhodnotit efekt této variability na odhady parametrů a na přesnost predikce z modelu.

Jedna z metod ohodnocení variability v odhadech parametrů je použití profilované odchylky. A další může být metoda charakteristiky podmíněné distribuce náhodných efektů v datech. Často používaný je také pravděpodobnostní model pro lineární-mixované efekty.

V příkladech pro Dyestuff a Dyestuff2 (modely fm1(2) a m1(2)ML) jsou tři parametry, pro které se obdržely odhady:

σ_1 - směrodatná odchylka náhodných efektů,
σ - směrodatná odchylka chybných termů v residual či v pozorování a
β_0 - parametry neměnných efektů (vyjádřeny jako Intercept).

Profilovací funkce (`profile()`) obměňuje parametry v modelu, aby dosáhla nejlepšího možného umístění modelu, které může být dosaženo s jedním neměnným parametrem se specifickou hodnotou. Potom dojde ke srovnání tohoto umístění s optimálním umístěním modelu (původním umístěním modelu). Modely jsou srovnány podle změny v odchylce, což je pravděpodobnostní poměrový test (LRT).

Intervaly spolehlivosti, podskupina intervalů odhadu, mají za cíl zjistit oblast, kde se skutečný parametr s velkou pravděpodobnostní nachází. Tedy odpovědět na otázku, jak blízko parametr leží či s jakou pravděpodobností se skutečný parametr odhadu nachází právě zde.

Intervaly spolehlivosti se do R prostředí zadávají takto:

```
pr1 <- profile(fm1ML)
```

Tento příkaz vytvoří plot funkce ξ, tedy pr1. Takovéto ploty se nazývají profilované zeta ploty.

```
confint(pr1)
```

	2.5 %	97.5 %	
.sig01	12.197461	84.063361	... σ_1
.lsig	3.643624	4.214461	... σ
(Intercept)	1486.451506	1568.548494	... β_0

```
xyplot(pr1, aspect = 1.3)
```

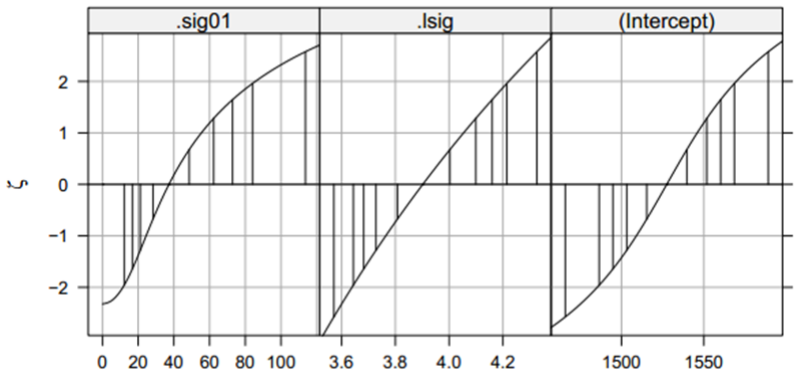
Hodnocením přesnosti parametrů σ_1, σ, β_0 je vytvořena funkce ξ. Ta bere každý parametr a aplikuje na ně transformaci odmocninou LRT, což vytvoří zeta funkci pro dané parametry. Díky tomu se dají parametry vykreslit (viz 5. graf).

Ideálně profilovaný zeta plot, bude vypadat jako rovná linie podél zájmových regionů, pak by se dalo říci, že se jedná o spolehlivý závěr založený na odhadech parametrů.

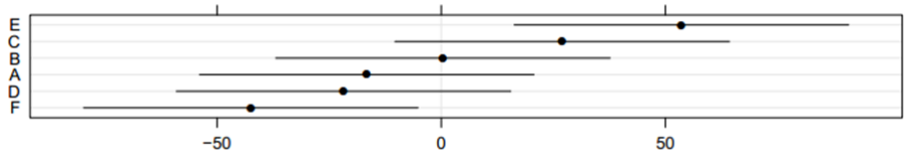
Pro práci s náhodnými efekty se často používá BLUP (bust linear unbiased estimators) – nejlepší lineární nestranný odhad. Připomeňme, že náhodné efekty nejsou tak docela parametry, neboť jsou nepozorovatelné náhodné proměnné, proto se bere raději v úvahu podmíněná distribuce a pozorovaná data. [26][11]

Zápis, který ukáže podmíněné módy:

```
ranef(fm1ML)
dotplot(ranef(fm1ML, postVar = TRUE))
```



Graf 5.: Funkce ξ pro fm1ML pro σ_1, σ, β_0.



Graf 6.: Odhad parametrů (95%) pro fm1ML.

4.3. Shrnutí

V této kapitole se vysvětlily neměnné a náhodné efekty, které jsou nedílnou součástí LMM a GLMM modelů. Díky jejich pochopení se velmi usnadní sestavení modelu a zároveň se pomocí nich definuje měřítko modelu.

Na ukázkových datech je ukázána transformace proměnných z numerických hodnot na faktory. Jsou uvedeny také osobní zkušenosti s metodami odhadů parametrů. Parametry se odhadují v R prostředí pomocí dvou metod. První je REML (restricted maximum likelihood) a druhá je MLE (maximum likelihood estimation).

Jednotlivé části výsledného modelu se rozebraly, porovnaly a popsaly. Možné interpretace výsledků lmer jsou uvedeny v následujících kapitolách. Toto je jen první seznámení.

Nad rámec jsou uvedeny i intervaly spolehlivosti parametrů, které se většinou dělají až u sofistikovanějších příkladů. Pro jednoduché ohodnocení modelu postačí metoda anovy a stepwise regrese. Pro ohodnocení modelu slouží také AIC a BIC kritérium.

5. Modely s mixovanými efekty – více náhodných efektů

V této kapitole jsou popsány lineární mixované modely (LMM) s více náhodnými efekty. Efekty mohou být jen křížené, tzv. „crossed“, nebo mohou být jen vnořené, tzv. „nested“ či mohou být částečně křížené (vnořené).

Zde je také představen příklad, kde se kombinují křížené a vnořené efekty s časovým hlediskem, tedy je lehce představena longitudinální analýza, která je lépe popsána v následující kapitole.

Důležitou součástí je i popis odhadu hypotéz, který slouží pro výběr vhodnějšího modelu. Představena je i metoda anovy, která určitým způsobem srovnává dva modely.

Všechny získané poznatky se ekvivalentně využijí i u GLMM modelu.

5.1. Modely s křížený mi náhodnými efekty

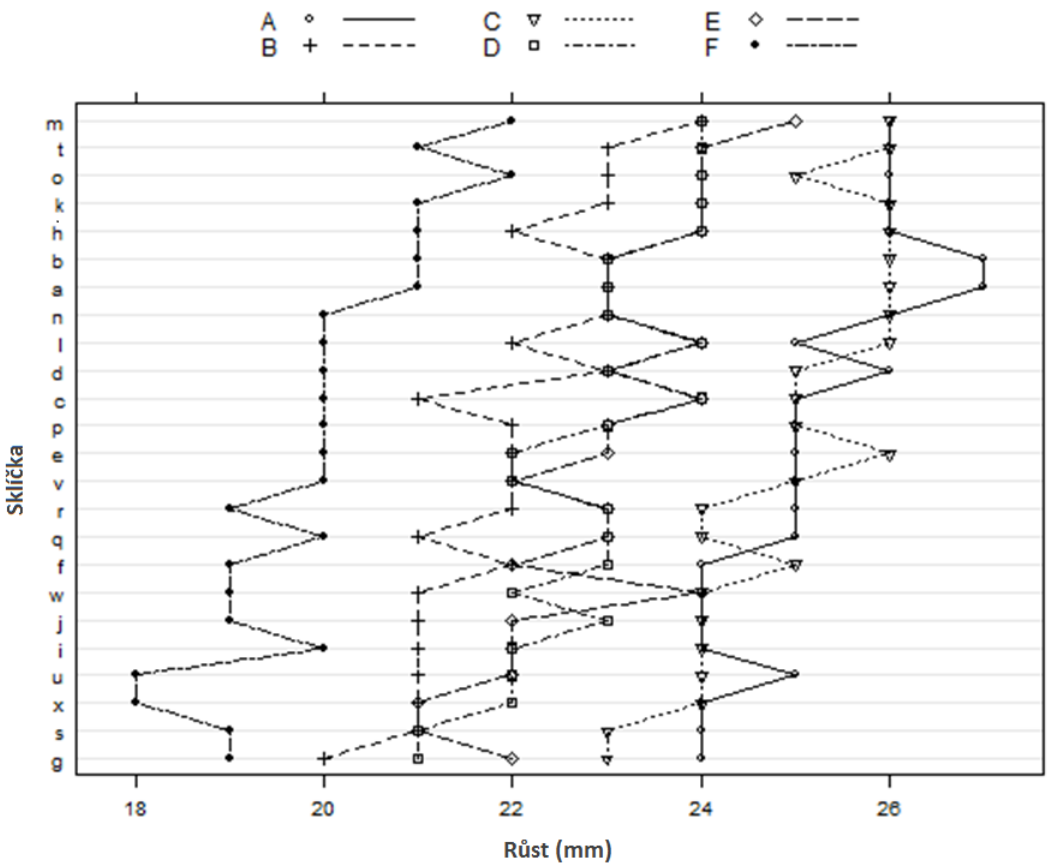
Modely s křížený mi náhodnými efekty lze například využít u zjišťování reakcí lidí na dané stimuly, kdy se zachytí vjem, přičemž každý jedinec vnímá stimuly jinak. Pokud je jedinec vnímán jako vzorek ze všech možných jedinců populace a stimuly jsou vzorkem z celé populace stimulů, tak jsou tyto dva faktory vnímány jako náhodné efekty. Stejně je to i u dotazníků, kdy každý jedinec odpovídá na otázky podle svého uvážení.

Pro takové výpočty se požívá balík lme4, který je schopen práce s rozsáhlými, nevyváženými daty s křížený mi efekty a větším počtem seskupujících faktorů. To je pro reálné analýzy velmi užitečné. Zde jsou však ukázány jednodušší příklady pro pochopení.

Pro první příklad jsou vybrána data s názvem **Penicillin**, v balíku BHH2, které obsahují 144 pozorování a 3 proměnné: Průměr (Diameter), Sklíčko (Plate) a Vzorek (Sample). Pomocí `str()` lze zjistit, zda jsou proměnné faktory. Ze `summary` je vidět, že je datová sada vyvážená.

```
str(Penicillin)
'data.frame': 144 obs. of 3 variables:
 diameter: num 27 23 26 23 23 21 27 23 26 23 ..
 plate : Factor w/ 24 levels "a","b","c","d",..: 1 1 1 1 1 1 2 2 2 2..
 sample : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2 3 4 ..
```

Variance v datech je propojená se Sklíčky a Vzorky. Každé Sklíčko je použito právě na šest Vzorků, proto je zajímavá změna mezi Sklíčky. Zajímavé je také zhodnocení potencionálu jednotlivých Vzorků na Sklíčku. Sklíčka i Vzorky jsou brány jako náhodný efekt, protože zde chceme zjistit variabilitu Vzorku ke Vzorku a ne účinnost jednotlivých Vzorků.



Graf 7.: Růst Vzorků A - F na jednotlivých Sklíčcích.

V tomto příkladu je každý Vzorek použit na každé Sklíčko, proto jsou tyto faktory křížené. Jsou také kompletně křížené, což znamená, že vždy existuje alespoň jedno pozorování pro každou úroveň Sklíček v kombinaci se Vzorkem. To ukazuje příkaz `xtabs()`. Hodnota „1“ je četnost použití Sklíčka na Vzorek.

```
xtabs(~  sample + plate, Penicillin)
Plate
sample a b c d e f g h i j k l m n o p q r s t u v w x
A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

atd.

Tyto dva faktory jsou vyvážené, protože je vždy jeden Vzorek na jedné z úrovní Sklíčka, ale ve skutečnosti takovou datovou sadu nemůžeme očekávat a častěji se setkáme s nevyváženými křížení faktory.

Model pro Penicillin

Model bude načten do proměnné fm2. V datech existují dva náhodné faktory a je třeba je připsat do formule pro vytvoření modelu.

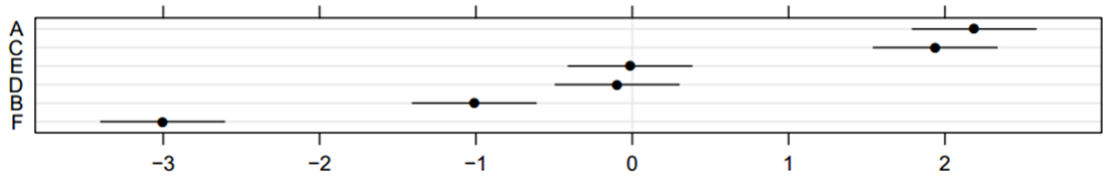
Zápis seskupujícího faktoru je v této podobě: (1|faktor). A hodnota „1“ za znakem „ $\sim$ “ zastupuje lineární komponentu.

```
fm2 <- lmer(diameter  ~  1 + (1|plate) + (1|sample), Penicillin)

Linear mixed model fit by REML
Formula: diameter  ~  1 + (1 | plate) + (1 | sample)
Data: Penicillin
 REML
330.9

Random effects: Groups Name    Variance Std.Dev.
plate (Intercept)  0.71691 0.84671
sample (Intercept) 3.73097 1.93157
Residual           0.30241 0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
       Estimate Std. Error t value
(Intercept) 22.9722 0.8086 28.41
```



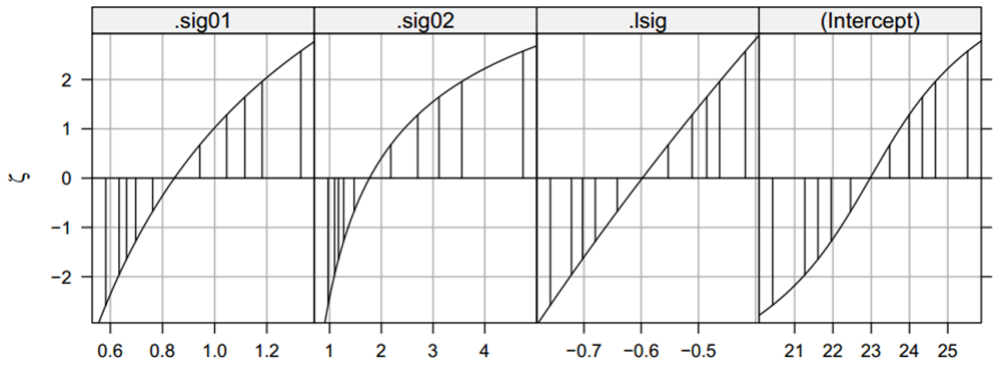
Graf 8.: Odhad parametrů (95%) pro náhodné efekty v modelu fm2.

Výsledek

Tento model ukazuje, že variabilita Vozku ke Vzorku (3,73097) má mnohem větší příspěvek než variabilita Sklíčka ke Sklíčku (0,71691). Větší variabilita znamená, že data jsou si méně podobná. Směrodatná odchylka (Std. Dev.) je druhá odmocnina variance a také ukazuje na podobnost rozložení dat. Variabilita pro Residual nemůže být přisouzená Vzorku ani Sklíčku. Malá hodnota Residual znamená, že jen malá část variability nemohla být modelována. Počáteční průměrná hodnota diametru je 22,9722.

Intervaly odhadu ukazují, že variabilita Sklíček ve spodní části grafu je menší než variabilita Sklíček v horní části grafu. Úroveň F má menší variabilitu než C.

Profilované zeta ploty pro parametry modelu fm2, jsou podobné těm, které jsou uvedeny u lineárních mixovaných modelů s jedním efektem. Zde je navíc jeden náhodný efekt. Připomeňme, že ideálním stavem by měla být přímka.



Graf 9.: Vertikální linie vyjadřují koncové body pro 50% až 99% pravděpodobnostního intervalu.

5.2. Modely s vnořenými náhodnými efekty

Vnořené náhodné efekty jsou přiblíženy na příkladu s daty **Pastes** v balíku lme4. Tyto data obsahují 60 pozorování a 4 proměnné: Sílu (strength), faktor Dávka (batch) s 10 úrovněmi, faktor Sud (cask) se 3 úrovněmi a faktor Vzorek (sample) se třiceti úrovněmi.

```
str(Pastes)
'data.frame': 60 obs. of 4 variables:
$ strength: num 62.8 62.6 60.1 62.3 62.7 63.1 60 61.4 57.5 56.9 ...
$ batch : Factor w/ 10 levels "A","B","C","D",..: 1 1 1 1 1 1 2 2 2 2..
$ cask : Factor w/ 3 levels "a","b","c": 1 1 2 2 3 3 1 1 2 2 ...
$ sample : Factor w/ 30 levels "A:a","A:b","A:c",..: 1 1 2 2 3 3 4 4 5..
```

Pro lepší orientaci je uvedeno `summary` se strukturou proměnných. U Vzorků jsou hodnoty ve tvaru A:a či A:b, to je však bráno za název jedné úrovně. Úrovně jsou pojmenovány takto, aby se naznačilo, že každá dávka A bude ve třech Sudech a, b a c, jak je vidět v `xtabs`. Každý vzorek je poté naměřen dvakrát.

```
summary(Pastes)
xtabs(~  batch + sample, Pastes, drop = TRUE, sparse = TRUE)
10 x 30 sparse Matrix of class "dgCMatrix"
A 2 2 2 . . . . . .
B . . . 2 2 2 . . . .
C . . . . . 2 2 2 . .
D . . . . . 2 2 2 .atd.
```

Protože se každá úroveň Vzorku vyskytuje pouze s jednou úrovní Dávky, tak se dá říci, že Vzorek je zahnízděný/vnořený u Dávky.

Pokud jsou náhodné efekty vnořené, použije se pro jejich odhad stejná funkce lmer jako pro odhad křížených náhodných efektů. Nezáleží na tom, zda jsou faktory křížené nebo z části křížené či vnořené vždy je na ně aplikována stejná funkce. V případě, že se vyskytne vnořený faktor ve slučovacím faktoru (1| faktor), tak je registrován, ale toto poznání nezmění postup výpočtu.

Je třeba zdůraznit jeden důležitý znak vnořených faktorů. Tím je možnost křížení tohoto faktoru v rámci jiného faktoru. Z příkazu `xtabs` pro Sud a Dávku je patrné křížení těchto faktorů. Proto by se měla vytvořit nová kategorická proměnná, která poskytne kombinaci Dávka Sud. Jednoduchou cestou jak vytvořit tento faktor je použití operátoru „:“, jak je vidět u batch:cask. Kombinace batch:cask je však stejná jako faktor Vzorek.

```
xtabs(~  cask + batch, Pastes)
    batch
cask A B C D E F G H I J
a 2 2 2 2 2 2 2 2 2 2
b 2 2 2 2 2 2 2 2 2 2
c 2 2 2 2 2 2 2 2 2 2
```

```
Pastes$sample <- with(Pastes, factor(batch:cask))
```

U malých datových sad, jako je tato, lze snadno rozpoznat faktory, které jsou vnořené a které jsou křížené. U velkých datových sad je vhodnější po určité diagnostice vytvořit nový faktor, který popisuje vzájemné postavení faktorů. Nebo lze spoléhat na funkci, která faktory rozpozná.

Tvorba modelu pro vnořené náhodné efekty je stejná jako tvorba modelu s křížený mi náhodnými efekty.

Model pro vnořené náhodné efekty:

```
fm3 <- lmer(strength  ~  1 + (1|sample) + (1|batch), Pastes, REML=0)
```

```
Linear mixed model fit by maximum likelihood
Formula: strength  ~  1 + (1 | sample) + (1 | batch)
     Data: Pastes
AIC BIC logLik deviance
256 264.4 -124 248
```

```
Random effects:
Groups Name       Variance Std.Dev.
sample (Intercept) 8.4337 2.9041
batch (Intercept) 1.1992 1.0951
Residual           0.6780 0.8234
Number of obs: 60, groups: sample, 30; batch, 10
```

```
Fixed effects:
     Estimate Std. Error t value
(Intercept) 60.0533 0.6421 93.52
```

Model pro křížené náhodné efekty:

```
fm2 <- lmer(diameter  ~  1 + (1|plate) + (1|sample), Penicillin)

Linear mixed model fit by REML
Formula: diameter  ~  1 + (1 | plate) + (1 | sample)
Data: Penicillin
 REML
330.9

Random effects: Groups Name     Variance Std.Dev.
plate (Intercept) 0.71691 0.84671
sample (Intercept) 3.73097 1.93157
Residual           0.30241 0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
   Estimate Std. Error t value
(Intercept) 22.9722 0.8086 28.41
```

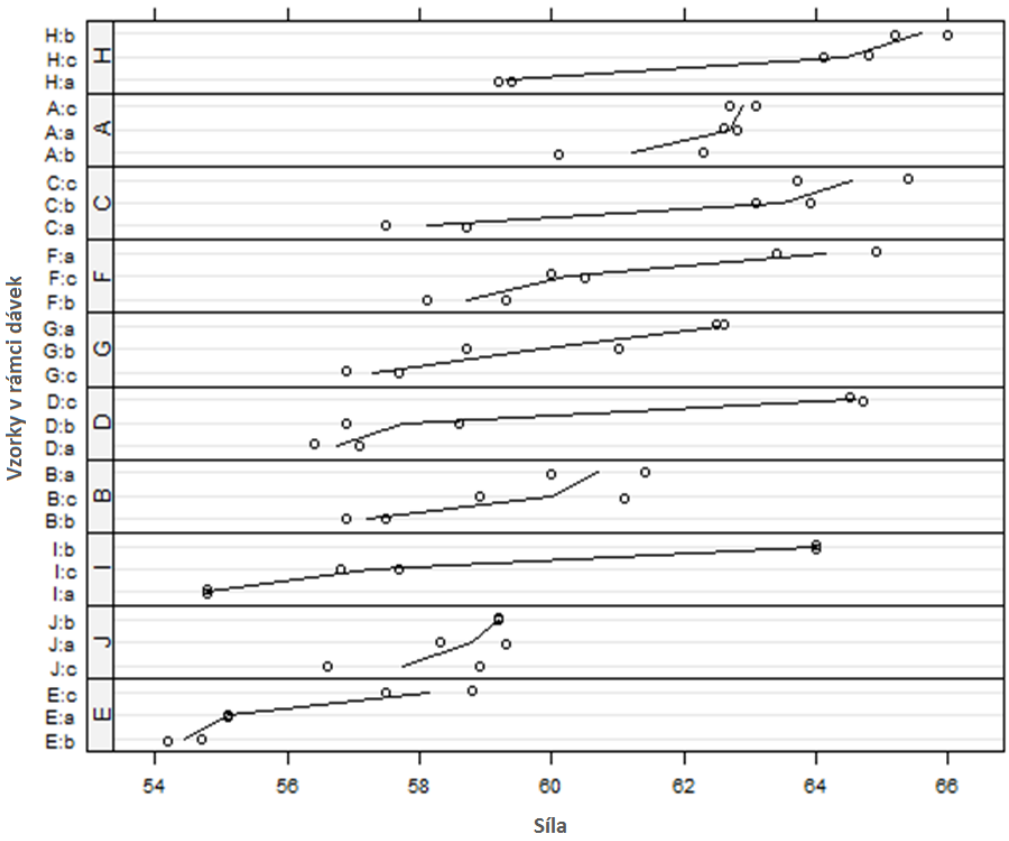
Výsledek

Z výstupu modelu fm2 lze pozorovat vysokou variabilitu (0.71691) u Vorku ku Vzorku. Naopak variabilita Dávky je daleko menší (3.73097). Residual je malý, tedy zbyla jen malá část variability, která nemůže být vysvětlena. Počáteční průměrná hodnota neměnného efektu (Síly) je 60.0533.

5.3. Testování statistické hypotézy

H_0 : $\sigma_2 = 0 \ldots$ nulová hypotéza
H_a : $\sigma_2 > 0 \ldots$ alternativní hypotéza

Testování statistické hypotézy se využívá, při zjišťování kvality umístění dvou modelů, které mají parametry $\sigma_2 = 0$ a $\sigma_2 > 0$. Tuto kvalitu lze zjistit z tzv. „p-value“ (p-hodnota), což je pravděpodobnost rozdílu v uchycení modelů. Pokud je hodnota p-value nízká až blízká nule, tak lze zamítnout nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_a a preferuje se model H_a. Naopak pokud je hodnota p-value vysoká, tak nelze zamítnout nulovou hypotézu H_0 a preferuje se model H_0.



Graf 10.: Průměrná hodnota pro Sílu na každém Vzorku v rámci každé Dávky.

H_0 : σ_2 = 0 ...vypadne hodnota (1|batch)

```
fm3a <- lmer(strength  ~  1 + (1|sample), Pastes, REML=0))

Linear mixed model fit by maximum likelihood
Formula: strength  ~  1 + (1 | sample)

 Data: Pastes
AIC BIC logLik deviance
254.4 260.7 -124.2 248.4

Random effects:
 Groups Name    Variance Std.Dev.
 sample (Intercept) 9.6328 3.1037
 Residual           0.6780 0.8234
Number of obs: 60, groups: sample, 30

Fixed effects:
    Estimate Std. Error t value
(Intercept) 60.0533 0.5765 104.2
```

H_a : $\sigma_2 > 0$ ...zachováme všechny hodnoty

```
fm3 <- lmer(strength ~ 1 + (1|sample) + (1|batch), Pastes, REML=0)

Linear mixed model fit by maximum likelihood
Formula: strength ~ 1 + (1 | sample) + (1 | batch)
 Data: Pastes
AIC BIC logLik deviance
256 264.4 -124 248

Random effects:
 Groups Name Variance Std.Dev.
 sample (Intercept) 8.4337 2.9041
 batch (Intercept) 1.1992 1.0951
 Residual          0.6780 0.8234
Number of obs: 60, groups: sample, 30; batch, 10

Fixed effects:
     Estimate Std. Error t value
(Intercept) 60.0533 0.6421 93.52
```

Pro porovnání umístěných modelů fm3 a fm3a se použije statistická metoda s názvem anova. Tou se zjistí, který model byl lépe uchycen. Model fm3a zde přestavuje nulovou hypotézu H_0 a model fm3 alternativní hypotézu H_a.

```
anova(fm3a, fm3)
Data: Pastes
Models:
fm3a: strength ~ 1 + (1 | sample)
fm3: strength ~ 1 + (1 | sample) + (1 | batch)
Df AIC BIC logLik Chisq Chi Df Pr(>Chisq)
fm3a 3 254.40 260.69 -124.20
fm3 4 255.99 264.37 -124.00 0.4072 1 0.5234
```

Protože je hodnota p-value velká, 0.5234, tak nelze zamítnout nulovou hypotézu H_0 a preferuje se model s H_0, tedy model fm3a. Aby se preferoval model fm3, tedy alternativní hypotéza, musela by být hodnota p-value menší než 0.05 či 0.01, podle toho zda počítáme s 5% chybou či jen s 1%.

5.4. Modely s částečně křížených náhodnými efekty

V reálném světě jsou běžnější datové sady s mnoha seskupujícími faktory, které nejsou zcela vnořené nebo zcela křížené. Takové to datové sady se označují jako datové sady s částečně křížených seskupujícími faktory náhodných efektů. Příkladem může být studie, která zachycuje výsledky studentů za určitý čas. Studenti jsou spojováni s učiteli a školami. Pokud měli studenti v průběhu školní docházky různé učitele, nemůže být faktor studentů vnořený v rámci učitelů

a také se kompletně nekříží faktor učitelů a studentů. Proto se pro tuto analýzu použije model s částečně křížených náhodnými efekty.

Studie, která se zabývá analýzou různých faktorů za určitý časový interval, se nazývá longitudinální analýza. V tomto případě, longitudinální analýza zkoumá tisíce faktorů v podobě studentů a stovky faktorů v podobě učitelů za daný čas s náhodnými efekty, které jsou jen částečně křížené. Na nadcházejícím příkladu je vysvětlena teorie. Ukázková studie ukáže, jak žáci ohodnotili své učitele.

Bude použita datová sada s názvem InstEval, balík lme4. Tyto data popisují evaluaci učitelů, za účelem udělení ceny nejlepšímu učiteli. Studie byla vytvořena Švýcarskou federální institucí pro technologie. Data obsahují 73 421 pozorování a 7 proměnných: s – Student, d – Instruktor, studage – Věk studenta, lectage – Věk lektora, dept – Oddělení kurzu, service – Zastoupení oddělením (zda neučilo studenty jiné oddělení), y – Známka.

```
str(InstEval)
'data.frame': 73421 obs. of 7 variables:
$ s : Factor w/ 2972 levels "1","2","3","4",..: 1 1 1 1 2 2 3 3 3 ..
$ d : Factor w/ 1128 levels "1","6","7","8",..: 525 560 832 1068 6..
$ studage: Ord.factor w/ 4 levels "2"<"4"<"6"<"8": 1 1 1 1 1 1 1 1 1 1 ..
$ lectage: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 2 1 2 2 1 1 1 1 1..
$ service: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
$ dept : Factor w/ 14 levels "15","5","10",..: 14 5 14 12 2 2 13 3 3 ..
$ y : int 5 2 5 3 2 4 4 5 5 4 ...
```

Model se uchytí pomocí tří náhodných efektů, to jsou Studenti, Oddělení a dept:service, což znamená, že pro Oddělení a Zastoupení oddělení byl vytvořen nový křížený faktor. To již bylo zmíněno u vnořených efektů.

```
fm4 <- lmer(y  ~  1 + (1|s) + (1|d)+(1|dept:service), InstEval, REML=0)

Linear mixed model fit by maximum likelihood
Formula: y  ~  1 + (1 | s) + (1 | d) + (1 | dept:service)
 Data: InstEval
AIC BIC logLik deviance
237663 237709 -118827 237653

Random effects:
Groups Name Variance Std.Dev.
s (Intercept) 0.105404 0.32466
d (Intercept) 0.262563 0.51241
dept:service (Intercept) 0.012126 0.11012
Residual       1.384953 1.17684
Number of obs: 73421, groups: s, 2972; d, 1128; dept:service, 28
```

```
Fixed effects:
     Estimate Std. Error t value
(Intercept) 3.25521 0.02824 115.3
```

Na otázku, zda lze použít jednodušší model, odpoví anova.

```
fm4a <- lmer(y  ~  1 + (1|s) + (1|d), InstEval, REML=0)

anova(fm4a,fm4)
Data: InstEval
Models:
fm4a: y  ~  1 + (1 | s) + (1 | d)
fm4: y  ~  1 + (1 | s) + (1 | d) + (1 | dept:service)
Df AIC BIC logLik Chisq Chi Df Pr(>Chisq)
fm4a 4 237786 237823 -118889
fm4 5 237663 237709 -118827 124.43 1 < 2.2e-16
```

Model fm4a je nulová hypotéza H_0: dept:service=0 a model fm4 je alternativní hypotéza H_a: dept:service>0. Podle velmi malého výsledného p – value lze zamítnou H_0 ve prospěch H_a, tedy se preferuje model fm4.[19][12]

5.5. Shrnutí

Ve formulích pro sestavení modelu s lineárními mixovanými efekty, byla vždy použita formulace pro vyjádření seskupujícího parametru pro náhodné efekty v podobě (1|faktor). Faktor je hodnota náhodného efektu, která bude použita pro seskupení. Třeba jako v modelu fm4a, kde se vybral faktor „s“ jako seskupující faktor Studentů. Faktor může být uveden také v podobě (1|faktor1:faktor2), jak bylo uvedeno v modelu fm4, kde se vytvořil nový křížený faktor, (1|dept:service).

Pro vytvoření modelu lze ve formuli uvést daleko větší počet náhodných faktorů, které budou určeny jako seskupující např. `a<- lmer(faktor1 ~ 1 + (1|faktor2) + (1|faktor3)+....+( 1|faktorN)`, kde N je počet faktorů v datasetu. Pro neměnné faktory by byla formule vytvořena takto: `b<- lmer(faktor1 ~ faktor2+ faktor3+....+faktorN)`. Lze využijí oba zápisy současně. Také lze vytvořit faktor pomocí (faktor1|faktor2).

V modelu se také nespecifikuje co je vnořený faktor či částečně nebo plně křížený faktor, protože tato informace je obsažena v datech samotných.

Po vytvoření modelu následuje testovaní hypotéz pomocí anovy, kdy proběhne srovnání modelu se složitějším nebo jednodušším modelem. Pomocí hodnoty p-value se vybere ten vhodnější.

Prostřednictvím dat InstEval je představena reálná datová sada s mnoha proměnnými a s mnoha mixovanými efekty, které byly částečně křížené. Poté byl model uchycen a porovnán pomocí anovy.

6. Modely pro longitudinální data

Zde je představena longitudinální analýza s korelovanými a nekorelovanými náhodnými efekty. Pro ukázkový příklad jsou vybrána data s časovou složkou a s hodnotami v podobě opakovaného měření na stejných subjektech. Využijí se modely LMM. Pro použití modelu GLMM je podmínkou ne-normální rozdělení dat.

Longitudinální data v sobě obsahují informaci o časové následnosti. Pozorované proměnné jsou naměřeny v průběhu určitého časového intervalu, tedy jsou opakovatelně měřeny za daný čas. Pozorované proměnné jsou vždy tytéž nebo jsou z velmi podobné skupiny subjektů (př. zkoumání 10 humrů, kteří jsou v akváriu, jak rostou za 5 let, či zkoumání růstu populace humrů za 5 let ve volné přírodě, když se pokaždé vyloví jiný jedinec).

Longitudinální analýza dokáže charakterizovat časový trend v rámci subjektů a mezi subjekty. V datech se vždy bude nacházet odezva, jakou měl subjekt v době pozorování určitého měřeného faktoru. Je možno také analyzovat vztah mezi testovacími a kontrolními subjekty, vztah v rámci subjektů či mezi subjekty. To už bylo nastíněno u datové sady InstEval, kde studenti ohodnocovali své učitele v průběhu let své školní docházky.

Jako ukázkový příklad byla vybrána studie o spánkových návycích řidičů kamiónů jménem „SleepStudy“, data jsou obsažena v balíku lme4. Subjekty, tedy vybraní řidiči, byli rozděleni do skupin, ve kterých řidiči mohli spát jen po omezenou dobu. Zde je analyzována skupina 18 subjektů, kterým bylo dovoleno spát jen tři hodiny denně a pak se jim několikrát změřil reakční čas. Data obsahují 180 pozorování a 3 proměnné: Reakce (Reaction, změřená doba reakce, v každém z deseti dnů), Dny (Days, počet po sobě následujících dnů, tedy 10 dnů) a Subjekt (Subject, 18 zkoumaných řidičů, kteří byli 10krát měřeni, tedy každý den jednou).

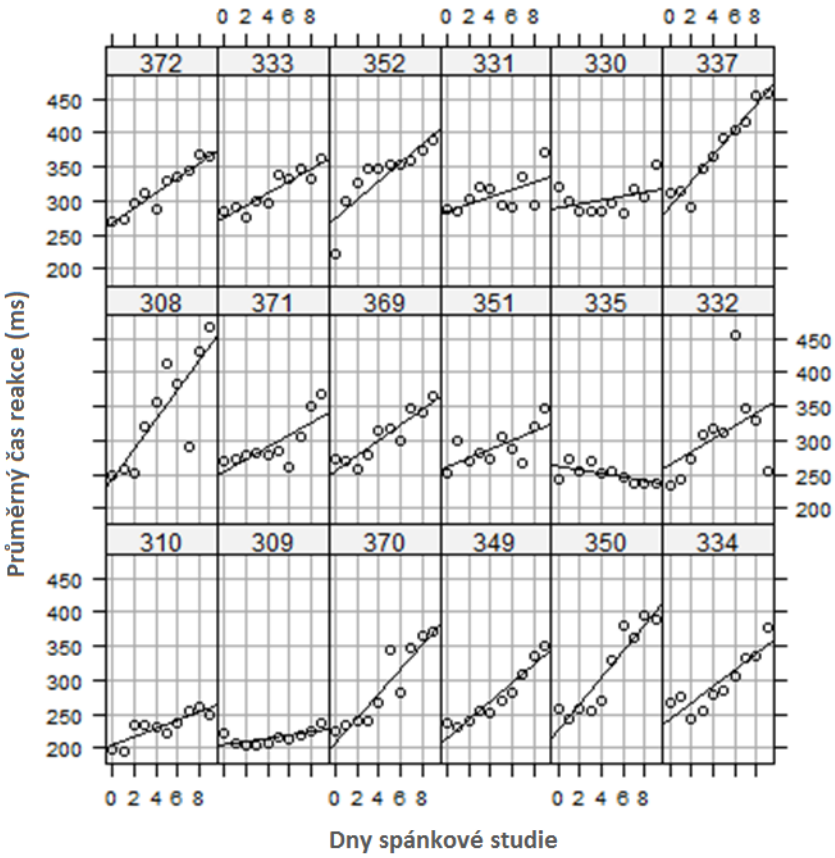
```
str(sleepstudy)
'data.frame': 180 obs. of 3 variables:
$ Reaction: num 250 259 251 321 357 ...
$ Days : num 0 1 2 3 4 5 6 7 8 9 ...
$ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1
```

Data jsou v tomto případě vyvážená, každý Subjekt je změřen v každém dnu.

```
xtabs( ~  Subject + Days, sleepstudy)
Days
Subject 0 1 2 3 4 5 6 7 8 9
308 1 1 1 1 1 1 1 1 1 1
309 1 1 1 1 1 1 1 1 1 1
310 1 1 1 1 1 1 1 1 1 1
330 1 1 1 1 1 1 1 1 1 1 atd.
```

Na 11. grafu lze pozorovat průměrnou reakční dobu řidičů (subjektů) v daných po sobě následujících dnech. Čísla subjektů jsou např. 372, 333 a pro každý subjekt je vytvořen vlastní panel (osa x jsou dny a osa y naměřená reakce). Výsledná přímka prostupuje diskrétními naměřenými hodnotami Reakcí řidičů. Čísla subjektů nejsou seřazeny chaoticky, ale jsou seřazeny podle vzrůstajících Interceptů, což usnadňuje srovnání subjektů.

```
print(xyplot(Reaction ~ Days | Subject, sleepstudy, aspect = "xy",layout
= c(6,3), type = c("g", "p", "r"),index.cond = function(x,y) coef(lm(y
x))[1],xlab = "Dny spánkové studie",ylab = "Průměrný čas reakce (ms)"))
```



Graf 11.: Průměrná doba reakce řidičů v průběhu spánkové studie.

K uchycení modelu jsou použity dva neměnné parametry (Intercept a sklon regresní přímky pro časový trend) a dva náhodné parametry pro Subjekt. Náhodné efekty pro daný Subjekt jsou rozdíly v Interceptu a sklonu časového trendu pro populaci. Sklon je typická změna Reakce Subjektu v čase spánkové studie a je vyjádřena regresní přímkou. Intercept je průměrná doba Reakce Subjektu v jednotlivých dnech. Náhodný efekt pro daný Subjekt je odchylka v Interceptu a sklonu časového trendu pro Subjekt.

6.1. Model s korelovanými náhodnými efekty

V prvním modelu je přípustná korelace náhodných efektů pro stejný Subjekt (nepodmíněná distribuce). To například znamená, že Subjekty s delší počáteční reakční dobou mohou být silněji ovlivněni nedostatkem spánku. Druhý model s nekorelovatelnými náhodnými efekty zaručí nezávislost náhodných efektů pro Intercept a sklon každého Subjektu.

Korelace vyjadřuje vzájemný vztah dvou proměnných. Pokud se jedna proměnná mění, mění se i jejich vztah. Jestli vykazují dvě proměnné vzájemnou korelaci, tak je jedna proměnná označována jako příčina a druhá jako následek. Výsledná míra korelace je v intervalu od $\langle -1; 1 \rangle$. [3]

```
fm8 <- lmer(Reaction  ~  1 + Days + (1 + Days|Subject), sleepstudy,
 REML = 0)

Linear mixed model fit by maximum likelihood
Formula: Reaction  ~  1 + Days + (1 + Days | Subject)
Data: sleepstudy
 AIC BIC logLik deviance
1764 1783 -876 1752

Random effects:
Groups Name Variance Std.Dev. Corr
Subject (Intercept) 565.516 23.7806
Days                32.682 5.7168 0.081
Residual            654.941 25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
Estimate Std. Error t value
(Intercept) 251.405 6.632 37.91
Days        10.467 1.502 6.97

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

Výsledek

Odhady parametrů pro neměnné efekty jsou 251.405 (Intercept) a 10.467 (Dny). Tyto hodnoty představují typickou počáteční reakční dobu řidičů bez spánkového deficitu (cca 250 milisekund) a typický vzrůst (sklon) reakční doby řidičů (tedy cca 10,5 milisekund za každý den spánkového deficitu). V náhodných efektech je sklon vyjádřen číslem 5.7168, což je asi 6 ms.

Odhadnutá směrodatná odchylka pro Subjekt ku Subjektu v Interceptu je 23.7806, to koresponduje se směrodatnou odchylkou Residual, 25.5918 ms. Tyto hodnoty vyjadřují, jak se od sebe jednotlivé Subjekty liší. Odhad 95% intervalu

by měly být okolo ± 50 ms. Po kombinaci těchto rozsahů s celkovou průměrnou reakční dobou (250 ms) jsou hodnoty Interceptu v intervalu od 200 ms do 300 ms. To lze odpozorovat z dat `str(sleepstudy)` i z 11. grafu.

Ze sklonů náhodných (5.7168) a neměnných (10.467) efektů pro Dny lze také vypočítat hraniční intervaly sklonu. Výpočet: $10,5 - 2 * 5,7 = -0,9$ msDny (spodní hranice) a $10,5 + 2 * 5,7 = 21,9$ msDny (horní hranice).

Hranice jsou patrné i z 11. grafu. V tomto grafu lze pozorovat, že Subjekty 309, 372 a 337 maní nižší rozsah intervalu než ± 50 ms, ale u Subjektů 308, 332 a 331 lze pozorovat rozsah vyšší. To je dobře, protože se počítalo od počáteční průměrné hodnoty.

Korelace v rámci Subjektů pro náhodné efekty je velmi nízká, 0.081. Z toho vyplývá, že nedostatek spánku a délka reakční doby jsou v určitém vztahu. Nelze říci, zda nedostatek spánku opravdu tak silně ovlivňuje pozorované počáteční reakční doby Subjektů. Takový malý korelační koeficient také naznačuje, že se může získat dobré umístění modelu i bez korelace.

6.2. Model s nezávislými náhodnými efekty

Pro modely s jednoduchými skalárními náhodnými efekty, kde bylo třeba vyjádřit jejich seskupení, je použita formule (1|faktor). Ve spánkové studii je jeden jednoduchý náhodný efekt pro Subjekt ve tvaru (1|Subject).

Protože Subjekty závisí na Dnech, tak je třeba vytvořit další seskupující faktor. Ten lze zapsat jako (0+Days|Subject). V tomto zápisu budou náhodné efekty sami produkovat korelovaný vektor hodnot náhodných efektů. Nula v zápisu znamená žádný Intercept a zápis Days|Subject značí seskupování Subjektů po Dnech. Alternativou je zápis v podobě (Days - 1 |Subject)., kde jsou Dny také bez Interceptu.

```
fm9 <- lmer(Reaction ~ 1 + Days + (1|Subject) + (0+Days|Subject),
sleepstudy, REML = 0)
```

```
Linear mixed model fit by maximum likelihood
Formula: Reaction ~ 1 + Days + (1 | Subject) + (0 + Days | Subject)
Data: sleepstudy
AIC BIC logLik deviance
1762 1778 -876 1752

Random effects: Groups Name Variance Std.Dev.
Subject (Intercept) 584.249 24.1713
Subject Days 33.633 5.7994
Residual 653.116 25.5561
Number of obs: 180, groups: Subject, 18
```

```
Fixed effects: Estimate Std. Error t value
(Intercept) 251.405 6.708 37.48
Days         10.467 1.519 6.89

Correlation of Fixed Effects:
    (Intr)
Days -0.194
```

V modelu fm9, je již korelace vypočtena jen pro neměnné efekty. V sekci náhodných efektů modelu fm9 nyní přibyl parametr Subjekt Dny.

Model fm9 není dále podrobněji komentován, protože je velmi podobný fm8 a lze si nepatrné odlišnosti odvodit. Modely se porovnají a vybere se ten lepší. K tomu využije opět metoda anovy.

Srovnání modelů fm8 a fm9:

```
anova(fm9, fm8)
Data: sleepstudy
Models:
fm9: Reaction  ~  1 + Days + (1 | Subject) + (0 + Days | Subject)
fm8: Reaction  ~  1 + Days + (1 + Days | Subject)
Df AIC BIC logLik Chisq Chi Df Pr(>Chisq)
fm9 5 1762.0 1778.0 -876.00
fm8 6 1763.9 1783.1 -875.97 0.0639 1 0.8004
```

Chisq (0,0639) velmi malý a p-value velké (0,8004), nelze zamítnout H_0 ve prospěch alternativní hypotézy H_a, preferuje se model fm9. Toto tvrzení vyplývá i z 11. grafu. Počáteční reakční doba nemá silný vztah na to, jak subjekt reaguje v průběhu spánkové studie s nedostatkem spánku. [18][13]

6.3. Shrnutí

Longitudinální data jsou data s časovou složkou, díky níž lze modelovat vývoj v čase. Tyto data v sobě ukrývají určité vztahy, například závislost mezi měřenými subjekty nebo závislost počáteční doby. I tyto vztahy se dají z modelu poznat. Pokud je třeba zahrnout závislost mezi daty, použije se model s korelovanými náhodnými efekty. Pokud je lepší vztah nezahrnout, použije se model pro nezávislé náhodné efekty.

Výstup modelu fm8 je také podrobně okomentován. Pomocí časové složky lze namodelovat i průběh sklonu v datech.

7. Případová studie 1: Využití modelu GLMM pro hodnocení navigace OLINA

Dada pro tuto případovou studii pochází z dotazníků, kterými bylo zjišťováno, jak vybraní lidé vnímají turistický multimediální průvodce Olomoucí - OLINA. Dotazníky byly vyplňovány v roce 2011 a 2012 stejnými subjekty v rámci projektu OLINA.

7.1. Data a jejich zpracování

Data se sbírala za účelem zhodnocení turistického multimediálního průvodce Olomoucí - OLINA. Tento průvodce umožňuje přístup k tisícům fotografií, hodinám audiovizuálních záznamů a nesčetně psaných komentářů o městě Olomouc. Na zpracování projektu OLINA se podílela česká firma Digital Urban Legends, město Olomouc a Katedra geoinformatiky, Univerzity Palackého v Olomouci. Více informací na http://geoinformatics.upol.cz/olina/.[23]

ID	rok	OT1	OT2	OT3	OT4	OT5
1	2011	2	3	4	1	3
1	2012	2	1	1	1	1
2	2011	3	3	3	1	3
2	2012	2	1	1	1	1
3	2011	2	2	4	2	3
3	2012	1	2	2	1	2
4	2011	2	2	4	1	2
4	2012	1	1	1	1	1
5	2011	1	3	3	2	3
5	2012	2	1	2	1	1

Obrázek 3.: Náhled na datovou sadu OLINA.

Dotazník obsahoval tyto otázky:

- OT1 – Hodnocení ovladatelnosti a srozumitelnosti Turistického multimediálního průvodce Olomoucí - OLINA (1 - Nejlepší, 5 - Nejhorší).
- OT2 – Hodnocení grafického zpracování a doprovodného grafického materiálu Turistického multimediálního průvodce Olomoucí - OLINA (1 - Nejlepší, 5 - Nejhorší).
- OT3 – Hodnocení textového zpracování pro body zájmu a reálie Turistického multimediálního průvodce Olomoucí - OLINA (1 - Nejlepší, 5 - Nejhorší).
- OT4 – Hodnocení atraktivity zvolených tras Turistického multimediálního průvodce Olomoucí - OLINA (1 - Nejlepší, 5 - Nejhorší).

- OT5 – Celkové hodnocení Turistického multimediálního průvodce OLINA v Olomouci (1 - Nejlepší, 5 - Nejhorší).

Otázky byly sestaveny panem Mgr. Pavlem Tučkem Ph. D. Autorka této diplomové práce data sbírala, upravovala a vyhodnocovala.

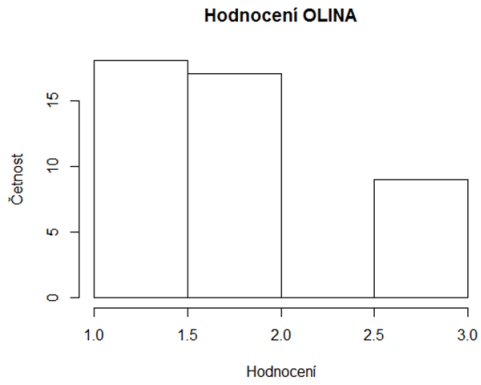
```
## načtení dat do proměnné o jako olina
o <- read.delim("D:/dat_olina.txt")
## zda jsou faktory
str(o)
## převod proměnných na faktory
o$ID<-factor(o$ID)

##snížení počtu proměnných
summary(aov(o$OT5 o$OT1*o$OT2*o$OT3*o$OT4))
                    Df Sum Sq Mean Sq F value Pr(>F)
o$OT1               1 1.856 1.856 14.406 0.000725 ***
o$OT2               1 15.023 15.023 116.597 1.72e-11 ***
o$OT3               1 1.877 1.877 14.569 0.000685 ***
o$OT4               1 0.526 0.526 4.084 0.052953 .
o$OT1:o$OT2         1 0.034 0.034 0.267 0.609520
o$OT1:o$OT3         1 0.000 0.000 0.002 0.964699
o$OT2:o$OT3         1 0.049 0.049 0.382 0.541631
o$OT1:o$OT4         1 0.140 0.140 1.086 0.306316
o$OT2:o$OT4         1 1.138 1.138 8.832 0.006022 **
o$OT3:o$OT4         1 0.077 0.077 0.595 0.446806
o$OT1:o$OT2:o$OT3     1 0.244 0.244 1.895 0.179578
o$OT1:o$OT2:o$OT4     1 0.257 0.257 1.992 0.169139
o$OT1:o$OT3:o$OT4     1 0.133 0.133 1.036 0.317447
o$OT2:o$OT3:o$OT4     1 0.036 0.036 0.277 0.603093
o$OT1:o$OT2:o$OT3:o$OT4 1 0.161 0.161 1.248 0.273414
Residuals              28 3.608 0.129
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pomocí funkce `summary` a funkce `aov` (anova), lze poznat, které z otázek mají největší vliv na celkové hodnocení OLINY. Existuje i sofistikovanější metoda ke zjištění nejvýznamnějších faktorů a tím je stepwise regrese.

```
## histogram
hist(o$OT5, main = paste("Hodnocení OLINA"), xlab="Hodnocení",
ylab="Četnost" )

##jak subjekty odpovídaly
print(dotplot(reorder(o$ID,o$OT5) o$OT5,Rail,xlab="Hodnocení",ylab="Subjekty"))
```



Graf 12.: Četnost hodnocení OLINY. Poissonovo rozdělení.

7.2. Snížení faktorů pomocí GLM a tvorba modelu GLMM

```
## vytvoření modelu GLM, hodnota family se musí nastavit na poissonovo
rozdělení
o.glm1<-glm(formula = OT5 ~ OT1 + OT2 + OT3 + OT4, family =
poisson(identity), data = o)

## výpočet stepwise regrese
stepwise(GLM.4, direction='backward/forward', criterion='BIC')
```

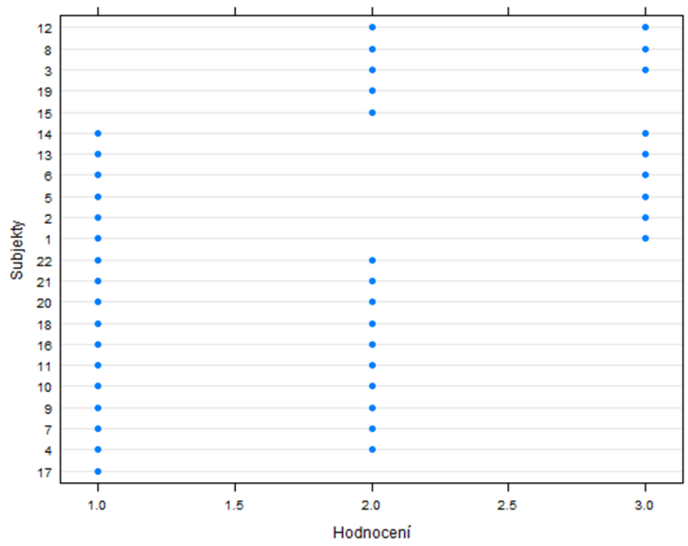
Pomocí stepwise regrese byla vybrána pouze jedna otázka, která nejvíce působí na celkovou hodnotu hodnocení, tou je OT2.

```
##vytvoření GLMM modelu

o.lmer1<-lmer(o$OT5 ~ 1+o$rok+(1+o$rok|o$ID)+(1|o$OT2),data=o)

Linear mixed model fit by maximum likelihood ['merMod']
Formula: o$OT5 ~ 1 + o$rok + (1 + o$rok | o$ID) + (1 | o$OT2)
 Data: o
 AIC BIC logLik deviance
 69.8747 82.3640 -27.9374 55.8747

Random effects:
 Groups Name Variance Std.Dev. Corr
 o$ID (Intercept) 2.393e-02 0.8119333
        o$rok     1.594e-09 0.0004035 -1.000
 o$OT2 (Intercept) 4.145e-01 0.9711412
 Residual          1.529e-01 3.910e-01
Number of obs: 44, groups: o$ID, 22; o$OT2, 4
```



Graf 13.: Známký, kterými subjekty hodnotily průvodce OLINA.

```
Fixed effects:
 Estimate Std. Error t value
(Intercept) 2.3773 328.9616 3.589
o$rok        -0.3638 0.1636 -3.582

Correlation of Fixed Effects:
       (Intr)
o$rok -1.000
```

Zda je vybrán dobrý model, lze zjistit jednoduše. Jednodušší model bude porovnán se složitějším modelem. Z výsledku stepwise regrese lze zjistit, že druhý nejlepší model v sobě počítá i s OT3. Proto se vytvoří model o.lmer2 a porovná se s modelem o.lmer1 pomocí metody anovy.

```
o.lmer2<-lmer(o$OT5 ~ 1+o$rok+(1+o$rok|o$ID)+(1|o$OT2)+(1|o$OT3),
family="poisson",data=o, REML=0)

Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: poisson
Formula: o$OT5  ~  1 + o$rok + (1 + o$rok | o$ID) + (1 | o$OT2) + (1 |
o$OT3)
 Data: o
 AIC BIC logLik deviance
 36.1672 48.6565 -11.0836 22.1672

Random effects:
 Groups Name     Variance Std.Dev. Corr
 o$ID (Intercept) 6.600e-01 0.8124165
```

```
        o$rok     1.631e-07 0.0004039 -1.000
 o$OT3 (Intercept) 1.181e+00 1.0868525
 o$OT2 (Intercept) 4.767e-01 0.6904263
Number of obs: 44, groups: o$ID, 22; o$OT3, 4; o$OT2, 4

Fixed effects:
         Estimate Std. Error z value
(Intercept) 2.3584 844.3197 0.500
o$rok -0.2097 0.4198 -0.499

Correlation of Fixed Effects:
      (Intr)
o$rok -1.000

anova(o.lmer1,o.lmer2)
Data: o
Models:
o.lmer1: o$OT5  ~  1 + o$rok + (1 + o$rok | o$ID) + (1 | o$OT2)
o.lmer2: o$OT5  ~  1 + o$rok + (1 + o$rok | o$ID) + (1 | o$OT2)
 Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
o.lmer1 6 26.928 37.633 -7.4641 14.928
o.lmer2 6 26.928 37.633 -7.4641 14.928 0 0 1
```

Nulovou hypotézu zastupuje model o.lmer1 a alternativní hypotézu zastupuje o.lmer2. Velká p-value hodnota naznačuje, že nelze zamítnout nulovou hypotézu ve prospěch alternativní. Proto se preferuje model o.lmer1.

7.3. Výsledky

Variance jsou velmi malé, v řádech $\langle 10^{-1}; 10^{-9} \rangle$, to ukazuje na podobnost dat. Počáteční průměrná hodnota celkového ohodnocení OLINY je 2.3773 (Intercept). To znamená, že OLINA byla v průměru hodnocena druhou nejlepší známkou.

Vývoj hodnocení OLINY, lze poznat ze sklonu neměnných efektů pro roky (-0.3638). I když je hodnota sklonu záporná, tak dobré mínění o turistickém průvodci roste. To je způsobeno inverzní stupnicí, jak je vidět na 15. grafu. Sklon v letech u náhodných efektů má hodnotu 0.0004035. Rovnice modelu je ve tvaru:

$$\text{OT5}=0.8119333*\text{ID}+0.9711412*\text{OT2}+0.0004035*\text{rok}+2.3773+(-0.3638).$$

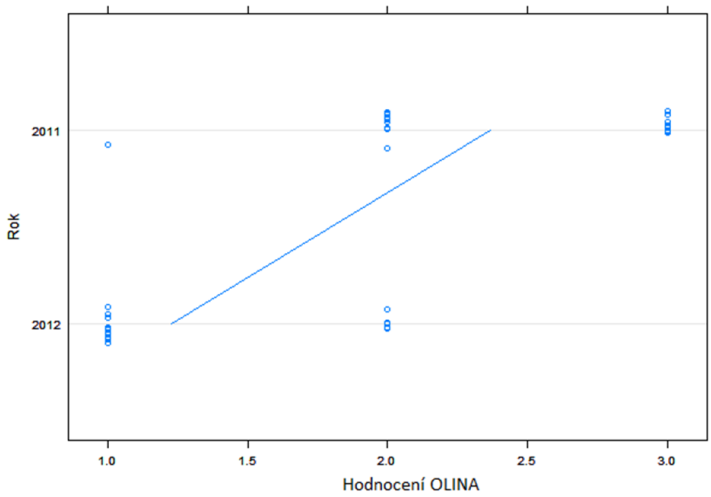
Ze směrodatné odchylky náhodných efektů ID a OT2, které jsou si podobné, lze odhadnout spodní a horní hranici průměrného celkového ohodnocení OLINY. Rozpětí tohoto intervalu je přibližně ± 1. Horní hranice, tedy nejlepší známka, kterou je OLINA ohodnocena je 1. Spodní hranice, tedy nejhorší známka, kterou je OLINA ohodnocena, je 3. To ukazuje 15. graf.

Ze sklonů náhodných (0.0004035) a neměnných (-0.3638) efektů pro roky lze také vypočítat hraniční intervaly sklonu.

Výpočet je následující: $0 - 2 * (-0.36) = 0.72$ ohodnocení OLINY/roky (spodní hranice) a $0 + 2 * (-0.36) = -0,72$ ohodnocení/roky (horní hranice). To znamená, že se sklon pohyboval v rozmezí ± 0,72. Ale na 15. grafu lze vidět, že některé subjekty se vymykají těmto intervalům, to je ovšem správně. Tyto intervaly jsou počítané od průměrné počáteční hodnoty.

Korelace dosahuje extrémní hodnoty -1. To je zapříčiněno tím, že celkové ohodnocení OLINY je povětšinou stejné jako ohodnocení otázky OT2 (Hodnocení grafického zpracování a doprovodného grafického materiálu Turistického multimediálního průvodce Olomoucí - OLINA). Z toho tedy plyne, že celkové ohodnocení, bylo podmíněno grafickým zpracováním OLINY a ostatní atributy jsou pro celkové hodnocení málo podstatné.

```
## hodnocení během let
print(dotplot(reorder(o$rok, o$OT5) ~ o$OT5, a,ylab = "Rok", jitter.y =
TRUE, pch = 21,xlab = "Celkové hodnocení",type = c("p", "a")))
```

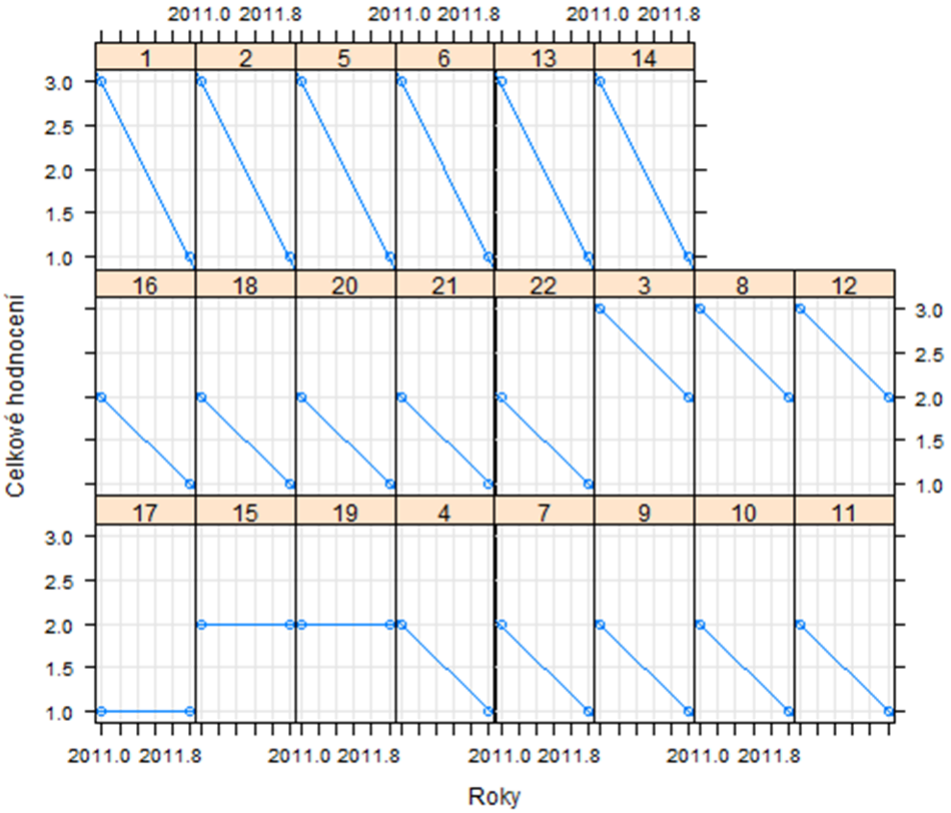


Graf 14.: Hodnocení OLINY za dané roky.

```
## míra a prudkost změny subjektů
print(xyplot(o$OT5 ~ o$rok| o$ID, o, aspect = "xy",layout = c(8,3),
type = c("g", "p", "r"),index.cond = function(x,y) coef(lm(y x))[1],xlab =
"Roky",ylab = "Celkové hodnocení"))
```

7.4. Diskuze

Datová sada je velmi jednoduchá, aby se upevnily nabyté znalosti z teoretického bloku. Použita jsou subjektivní data, nasbírána pomocí dotazníků, aby se



Graf 15.: Změna hodnocení v průběhu let u vybraných subjektů.

dokázala využitelnost GLMM modelů i v této oblasti. Položené otázky v dotazníku jsou dobře sestaveny a i stupnice ohodnocení je přirozená. Jen je naměřeno málo subjektů.

Proto je použita pro odhad parametrů metoda maximální pravděpodobnosti, která je pro počty s nedostatkem dat určena.

Na 15. grafu nejsou jednotlivé subjekty uvedeny nahodile, ale jsou seřazeny podle vzrůstajících Interceptů pro lepší porovnání subjektů. Časová osa neukazuje přímo roky 2011 a 2012, protože použitá implementace jazyka R si přizpůsobila popis osy, aby ji mohla vykreslit.

7.5. Závěr

V předešlých příkladech v teoretickém bloku, jsou data většinou objektivního charakteru. Taková data lze například získat měřením.

Zde jsou však představena data subjektivního charakteru. Každý jedinec odpovídal podle svého uvážení, které bylo podloženo jeho fyziologickými vlastnostmi, jeho vzděláním či jinde nabytými zkušenostmi.

Případové studie 2 a 3 nemohou být zveřejněny.

8. Diskuze

Většina lidí si pod pojmem dotazníky představí papír s vypsanými otázkami, které jsou ohodnoceny či zodpovězeny. V této diplomové práci jsou pod pojmem dotazníky chápany i datové sady, které vznikly pozorováním nebo měřením reálného světa. To je logické, protože dotazníky či datové sady obsahují informace, které jsou určitým způsobem nasbírány, uskladněny a je s nimi nějak nakládáno. Regresní modely a longitudinální analýzu lze použít jak na subjektivní tak i na objektivní data. Diplomová práce má za cíl představit co nejširší okruh použití těchto statistických metod. I když je pravda, že v praktických příkladech je počítáno s „pravými“ dotazníky jen jednou, rešerše uvádí mnoho dalších příkladů.

Praktické ukázky a příklady v teoretickém bloku jsou vybírány tak, aby byly všeobecně známé. To je prospěšné, neboť je velmi snadné k těmto příkladům dohledat mnoho alternativních výukových materiálů, které tyto data také používají. S danými daty se postup práce liší a někdy je dobré se podívat na jejich zpracování z více úhlů pohledu.

Teorie je primárně získána z jednoho druhu zdroje a to z výukových materiálů R-forge. Na tyto materiály se ovšem odkazují téměř všechny ostatní výukové materiály, které k výpočtu modelů používají jazyk R. Protože k těmto jedinečným dokumentům neexistuje český překlad, tak tato diplomová práce může být pokládána za jejich českou alternativu. Do teoretického bloku ovšem nejsou začleněny celé dokumenty, protože popisovaly i složité metody, které se neslučovaly s cílem této diplomové práce.

Přestože je cílem práce hlavně představit metodu GLMM, tak v teoretickém bloku se více pracuje s metodou LMM. Tyto dvě metody se liší jen v rozdělení dat, jak je uvedeno a podrobně vysvětleno. V případových studiích se pracuje jen s GLMM modely a jak je vidět, teorie je ekvivalentně použitelná.

Nad rámec jsou uvedeny intervaly spolehlivosti a odhady parametrů, které se v případových studiích nepoužily. To je zapříčiněno cílem této diplomové práce. Měla se poprvé představit teorie GLMM modelů a longitudinální analýzy. Intervaly spolehlivosti a odhady parametrů jsou vyšší statistikou a jsou uvedeny, jako možnost dalšího rozvoje.

Ohodnotit vhodnost modelu lze pomocí metody anovy, která srovnává jednodušší model se složitějším. Jistě existuje sofistikovanější metoda k odhadu přesnosti modelu, ale pro účel této diplomové práce, která prvotně představuje metodu longitudinální analýzy a regresních modelů, postačuje tato jednoduchá metoda.

Pro případové studie jsou záměrně vybírány netradiční datové sady, protože cílem je prohloubit znalost o představených metodách. Proto se někdy může zdát, že případová studie je nedokončena z hlediska tématu dat, ale téma dat je až druhotné a jako výsledek postačí komentář výsledného modelu.

Prezentování výsledků z modelu lmer, mohlo být zatíženo drobnou chybou. Pro vizualizaci výsledků funkcí z balíku lme4 je nově definovaná funkce `plotLMER.fnc`, která by měla umět výsledky zobrazit. To se nepovedlo, i když se podnikly veškeré kroky, které měly vést ke spuštění této funkce. Proto se tato funkce nahradila jinými ploty a jinými funkcemi.

Každá případová studie se zaměřuje na práci s jinou části uvedené teorie. Například práce s Poissonovým rozdělením, testování hypotéz a použití anovy jsou hlavním bodem první případové studie. Nad daty druhé případové studie se podrobněji představila stepwise regrese a snižování parametrů. Poslední případová studie se zaměřuje se na modely s korelovanými a nezávislými náhodnými efekty.

Využití regresních modelů, zvláště pak GLMM modelu, a longitudinální analýzy v geoinformatice je velmi praktické, protože je to další možnost statistického ohodnocení dat. Díky možnostem GIS se dají výsledky metod lépe zpracovat a prezentovat.

9. Závěr

Tato diplomová práce představila longitudinální analýzu a regresní modely, nejvíce pak LMM a GLMM modely, v základním jednoduchém pojetí prvotního seznámení s těmito metodami. Nejdříve je přestavena teorie, poté je na praktických příkladech vysvětlena a použita na případové studie.

První kapitola rešerše má za cíl přiblížit oblasti využití vybraných regresních modelů a longitudinální analýzy. Tato kapitola slouží také jako motivace pro pochopení složité teorie, která tyto analýzy provází. Studie byly vybírány na základě příbuznosti tématu s oborem geoinformatiky.

Následující kapitola definuje regresní modely matematicky a ke každému modelu je uveden praktický příklad v jazyku R. Díky této sumarizaci všech vybraných regresních modelů a jejich příkladů se ucelila představa o těchto metodách. Také díky rostoucí složitosti modelů je lépe pochopen princip jejich tvorby a prohloubilo se chápání jejich odlišností. Statistické odlišnosti jsou v počtu přípustných proměnných v modelu, rozložení dat a jejich vztah. V implementaci jazyku R se liší funkce pro upevnění modelu na data a způsob tvorby výsledného grafu.

Pokračováním teoretického bloku je detailnější popis modelů LMM a GLMM. Kapitola „Mixovanými náhodnými efekty – jeden efekty“ představuje základní používané termíny jako náhodný a neměnný efekt. Také je vysvětlen význam a tvorba faktoru (štítku) v datech. Na datových sadách DyestuffData a Dyestuff2Data se popsaly sekce výsledku modelu. Podrobněji se objasnily metody odhadů pravděpodobnosti REML a MLE pro regresní modely. Velká část je věnována vysvětlení pojmů Residual, Intercept, Standard deviation, Variance a Standard Error. Tyto pojmy jsou velmi důležité pro okomentování výsledů modelu. Nad rámec jsou uvedeny pojmy interval spolehlivosti a odhady parametrů. V jednoduchých příkladech, které jsou zde představeny, se tato vyšší statistika nepoužívá a slouží jako motivace k dalšímu rozvoji.

Následná kapitola „Modely s mixovanými efekty – více náhodných efektů“ vysvětluje pojmy křížené a vnořené náhodné efekty na datech Penicillin a Pastes. Zde je důležité první podrobné okomentování výsledků modelů. Vzorce přímek uvedeny nebyly, protože to je tématem první části teorie.

Testování statistických hypotéz je nedílnou součástí tvorby modelů, protože pomáhají ohodnotit vhodnost modelu. Toto ohodnocení je provedeno na základě srovnání dvou modelů, jednoho složitějšího a druhého jednoduššího. Pro srovnání se používá metoda anovy, která srovná nulovou hypotézu (jednodušší model) s alternativní hypotézou (složitější model) a podle velikosti hodnoty p-value se jedna z hypotéz preferuje.

Longitudinální analýza je vysvětlena v poslední části teorie. Longitudinální data jsou tedy data, která v sobě nesou časovou posloupnost a na která se dá aplikovat jedna z metod regrese. Longitudinální analýza je potom regresní analýza nad longitudinálními daty s korelovanými nebo nezávislými efekty.

První případová studie pracuje s daty z dotazníků, které hodnotily turistický multimediální průvodce Olomoucí - OLINA. Studie je zaměřena práci s Poissonovým rozdělením a na testování hypotéz pomocí anovy.

Druhá případová studie podrobněji představuje stepwise regresy a snižování parametrů pro tvorbu modelu. Třetí případová studie je vytvořena pro kontrolu druhé případové studie a zaměřuje se na modely s korelovanými a nezávislými náhodnými efekty.

10. Summary

In this diploma thesis was introduced the regression models and the longitudinal analysis. Acquired theory was used for solving practical problems in R language.

The chosen regression models were: the linear model (LM), the mixed linear model (LMM), the generalized linear model (GLM) and the generalized mixed linear model (GLMM). The main difference among these models is in their data distribution and in their data dependency.

Longitudinal data are data with time parameter. These data contain the measurement of subjects, which should have been accomplished on the same or very similar subjects in continuous time period. The longitudinal analysis is then analysis above these longitudinal data.

Keywords: regression, regression analysis, regression models, linear model, LM, mixed linear model, LMM, generalized linear model, GLM, generalized mixed linear model, GLMM, longitudinal data, longitudinal analysis, questionnaires, tree damage.

The regression analysis is method for modeling and analyzing dependent and independent variables. This method is able to estimate an independent variable(s) on the basis of knowledge of a dependent variable(s). This is very beneficial for all kind of data characterization. The data have different type of dependency and distribution, therefore the regression models are divided into more groups. The linear model (LM) has normal distribution and independent data. The mixed linear model (LMM) has normal distribution and dependent data, the generalized mixed model (GLM) has non-normal distribution and independent data. And at last the generalized mixed linear model (GLMM) has non-normal distribution and dependent data. All these models can be used for longitudinal analysis.

The term questionnaires was understood as all types data sets and it was not matter if the data were subjective or objective. The regression and longitudinal analysis work on both types. And moreover the aim of this thesis was to introduce the widest range of useful areas, where these analysis can be performed.

In the second chapter are given examples, where these statistics methods can be used and why are these methods so wholesome for geoinformatics field.

Next chapter is the mathematical summary of these models with practical illustration on the real data sets in R language.

The main theoretical block describes the mixed linear models (LMM) but as we know the findings are equally useful also for generalized mixed linear models (GLMM). Here was explained the random and fixed effect, the restricted maximum likelihood and the maximum likelihood estimation and described the sections of functions outputs. The essential terms were Residual, Intercept, Standard deviation, Variance and Standard Error. The further assessment and parameters

estimation were included as opportunity of following progress in the statistics methods.

It was necessary to explain the crossed and nested random effects, what was done with examples in the next section. The main point of this section was the hypothesis testing, which can evaluate the chosen models.

Last theory block was reserved for longitudinal analysis, where the regression models could be applied. Also the models outcomes were explained in detail.

The case studies were designed for un-typical data, where can be the theory comprehend in other way. In the first case study was shown an example with the Poisson distribution and the hypothesis testing. Next one explained the stepwise regression and the elimination of extra parameters. And in the last one chapter the models with correlated and uncorrelated random effects were processed.

Reference

[1] Anděl, J. *Základy matematické statistiky.* Univerzita KarlovavPraze, Matematicko-fyzikální fakulta. Preprint. Praha 2002

[2] LEVÁKOVÁ, Marie. *Zobecněné lineárnísmíšené modely* PBrno, 2011. Diplomová práce. Masarykova universita.

[3] MELOUN, Milan; MILITKÝ, Jiří. *Statistická analýza experimentálních dat.* Vydání 2., upravené a rozšířené. Praha: Academia (Akademie věd České republiky), 2004. Korelace, s. 737 - 779. ISBN 80 – 200 – 1254 – 0.

[4] ABSHER, J.D., A.G. GRAEFE a R.C. BURNS. Longitudinal Monitoring of Public Reactions to the U.S. *Series of the Institute for Landscape and Open Space* online]. 2008, Issue 2, s. 9-15 [cit. 2012-02-18]. Dostupné z: *http://www.treesearch.fs.fed.us/pubs/36837*

[5] BAKKESTUEN, Vegar, Rune HALVORSEN and Einar HEEGAARD. Disentangling complex fine-scale ecological patterns by path modelling using GLMM and GIS.*Journal of Vegetation Science*[online]. 2009, č. 5, 779–790 [cit. 2012-02-18]. DOI:10.1111/j.1654-1103.2009.01001.x. Dostupné z: *http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2009.01001.x/full*

[6] BOLKER, Benjamin M., Mollie E. BROOKS, Connie J. CLARK, Shane W. GEANGE, John R. POULSEN, Henry H. STEVENS a Jada-Simone S. WHITE. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* online]. 2009, Volume 24, Issue 3, 127–135 [cit. 2012-03-03]. Dostupné z: *http://www.sciencedirect.com/science/article/pii/S0169534709000196*

[7] BRANDÃO, A., D.S. BUTTERWORTH, S.J. JOHNSTO a J.P. GLAZER. Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster, Jasus lalandii. *Fisheries Research*[online]. 2004, roč. 20, 2-3, 339–349 [cit. 2012-02-18]. ISSN 01657836. DOI: 10.1016/j.fishres.2004.08.012. Dostupné z: *http://www.sciencedirect.com/science/article/pii/S0165783604001766*

[8] EASTON, Valerie J. a John H. MCCOLL. STATISTIC GLOSSARY. *Paired data, correlation & regression* [online]. 1.1. 2000 [cit. 2012-03-04]. Dostupné z: *http://www.statsci.org/smyth*

[9] EVERITT, Brian S. a Torsten HOTHORN. A Handbook of Statistical Analyses Using R. Http://cran.r-project.org [online]. 2008, [cit. 2012-02-26]. Dostupné z: *http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_analysing_longitudinal_dataI.pdf*

[10] Forest health hazard zones in The Czech Republic: Based on statistical evaluation of fuzzy model for years 2000-2010. *Journal of Maps, Modelování růstových podmínek lesů v ČR*. 2012, in print. *http://www.statsci.org/smyth*

[11] Ch1.R. In: R-PROJECT, R-forge. *Lme4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *http://lme4.r-forge.r-project.org/book/Ch1.R*

[12] Ch2.R. In: R-PROJECT, R-forge. *Lme4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *http://lme4.r-forge.r-project.org/book/Ch2.R*

[13] Ch4.R. In: R-PROJECT, R-forge. *Lme4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *http://lme4.r-forge.r-project.org/book/Ch4.R*

[14] Linear mixed model implementation in lme4. In: BATES, Douglas. *R-project.org* [online]. 2011 [cit. 2012-03-03]. Dostupné z: *http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf*

[15] Linear Regression in R. COLLEGE OF THE REDWOODS. *Department of Mathematics College of the Redwoods* [online]. 2008 [cit. 2012-03-03]. Dostupné z: *http://msenux.redwoods.edu/math/R/regression.php*

[16] MAJUMDAR Anandamayee, Corinna GRIES a Jason WALKER. A Non-stationary Spatial Generalized Linear Mixed Model Approach for Studying Plant Diversity. *Journal of Applied Statistics*[online]. 2011, Volume 38, Issu. 9, s. 1935-1950 [cit. 2012-03-03]. DOI: 10.1080/02664763.2010.537650. Dostupné z: *http://stat.asu.edu/ ananda/JAppStat-plants.pdf*

[17] Mission. AMERICANS INSTITUTES FOR RESEARCH.*The National Center for Analysis of Longitudinal Data in Education Research*[online]. 2011 [cit. 2012-03-03]. Dostupné z: *http://www.caldercenter.org/mission.cfm*

[18] Models for Longitudinal Data. In: *R-project: lmer4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *: http://lme4.r-forge.r-project.org/book/Ch4.pdf*

[19] MOdel With Multiple Random-effects Terms. In: *lmer4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *http://lme4.r-forge.r-project.org/book/Ch2.pdf*

[20] MYUNG, In Jae. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* [online]. 2003, č. 47, s. 90-100 [cit. 2012-03-04]. Dostupné z: *http://people.physics.anu.edu.au/ tas110/Teaching/Lectures/L3/Material/-Myung03.pdf*

[21] NAGY, Krisztina, Anikó SCHROTT a Péter KABAI. Possible influence of neighbours on stereotypic behaviour in horses. *Applied Animal Behaviour Science* [online]. 2008, Volume 111, Issue 3, s. 321-328 [cit. 2012-03-03]. Dostupné z: *http://www.behav.org/kabai/abstracts/kabai_horse_stereotypy.pdf*

[22] O'BRIEN, M., P. R. NADER, R. M. HOUTS, R. BRADLEY, S. L. FRIEDMAN, J. BELSKY, E. SUSMAN a THE NICHD EARLY CHILD CARE RESEARCH NETWORK. The ecology of childhood overweight: a 12-year longitudinal analysis. *International Journal of Obesity*[online]. 2007, 1469–1478 [cit. 2012-02-18]. DOI: 10.1038/sj.ijo.0803611. Dostupné z: *http://www.nature.com/ijo/journal/v31/n9/abs/0803611a.html#aff6*

[23] OLINA: turistický průvodce Olomoucí. KATEDRA GEOINFORMATIKY.*Geoinformatics.upol.cz* [online]. Olomouc, 2010 [cit. 2012-04-06]. Dostupné z: *http://geoinformatics.upol.cz/olina/*

[24] RODRÍGUEZ, Germán. PRINCETON UNIVERSITY. *Generalized Linear Models* [online]. 2011 [cit. 2012-03-03]. Dostupné z: *http://data.princeton.edu/R/glms.html*

[25] ŘEZANKOVÁ, Hana; MAREK, Luboš Marek; VRABEC, Michal. *IASTAT* [online]. 2001 [cit. 2012-02-16]. INTERAKTIVNÍ UČEBNICE STATISTIKY. Dostupné z WWW: *http://iastat.vse.cz/*

[26] Simple, Linear, Mixed-effects Model. In: *R-project: lmer4.r-forge* [online]. 2010 [cit. 2012-03-04]. Dostupné z: *http://lme4.r-forge.r-project.org/book/Ch1.pdf*

[27] SMYTH, G. K. a A. P. VERBYLA. A conditional approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society* [online]. 1996, č. 58, s. 565-572 [cit. 2012-03-04]. Dostupné z: *http://www.statsci.org/smyth*

[28] STATPAC. *The Statistics Calculator Statistical Analysis Tests At Your Fingertips* online]. 2012 [cit. 2012-03-04]. Dostupné z: *http://www.statpac.com/statistics-calculator/correlation-regression.htm*

[29] SVENNING, J.C., T. FABBRO a S. J. WRIGHT. Seedling interactions in a tropical forest in Panama. *Oecologia*[online]. 2008, č. 1, s. 143-150 [cit. 2012-02-18]. DOI: DOI: 10.1007/s00442-007-0884-y. Dostupné z: *http://www.springerlink.com/content/5122nn2524222863/*

[30] VIDRINE, Jennifer Irvin, Damon J. VIDRINE, Tracy J. COSTELLO, Carlos MAZAS, COFTA-WOERPEL a David W. WETTER. The Smoking Consequences Questionnaire: Factor structure and predictive validity among Spanish-speaking Latino smokers in the United States.*Nicotine & Tobacco Research* [online]. 2009, č. 1 [cit. 2012-03-03]. DOI: 10.1093/ntr/ntp128. Dostupné z: *http://ntr.oxfordjournals.org/content/early/2009/08/20/ntr.ntp128.abstract?etoc*

A. Obsah přiloženého CD

Adresáře:

A) **rimska_pripadstud** (data a kódy k případovým studiím):

1_OLINA:
dotazniky_olina.xls (hodnoty odpovědí z dotazníku)
olina_kod.txt (kód/výpočet pro prostředí R softwaru s výsledky)
dat_olina.txt (data z dotazniky_olina.xls; pro načtení do R softwaru)
dat_olina_kratsi.txt (vybrané data z dat_olina.txt, pro prezentování výsledků)

B) **rimska_www** (internetové stránky k diplomové práci):
index.html (hlavní stránka)
long.html (longitudinální analýza)
modely.html (typy regresních modelů)
popis.html (popis modelu)
practicka.html (případová studie OLINA)
sum.html (Summary)
style.css(styl webových stránek)
images (složka s přiloženými obrázky)