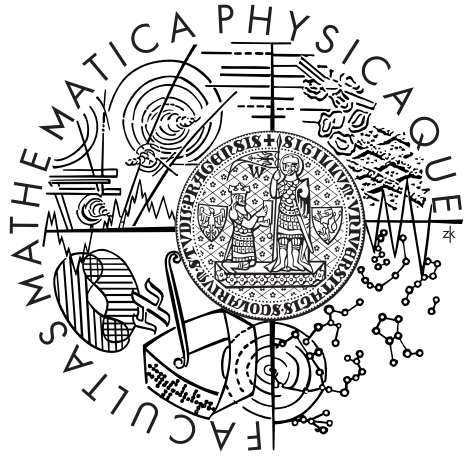
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Vojtěch Šaroch

Úvod do lineárních smíšených modelů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Mgr. Michal Kulich, Ph.D.

Studijní program: matematika

Studijní obor: finanční matematika

Praha 2011

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V `Praze` dne ..... Podpis autora

Poděkování

Děkuji tímto doc. Mgr. Michalu Kulichovi, Ph.D. za odborné vedení bakalářské práce, cenné rady, připomínky a především čas, který mi věnoval. Také děkuji společnosti SPSS CR za zapůjčení softwaru IBM SPSS Statistica 19.0. Dále bych poděkoval rodině za materiální a morální podporu a za pracovnu, kde jsem mohl nerušeně kompletovat svojí práci.

Název práce: Úvod do lineárních smíšených modelů
Autor: Vojtěch Šaroch
Katedra: Katedra pravděpodobnosti a matematické statistiky MFF UK
Vedoucí diplomové práce: doc. Mgr. Michal Kulich Ph.D.

Abstrakt: Práce popisuje obvyklé postupy odhadování a testování hypotéz pro lineární statistické modely. Cílem těchto modelů je porovnávání skupin podle předem určených závislých proměnných. Analýza rozptylu a lineární smíšené modely jsou hojně využívány ve většině vědních oborů jako je farmakologie, biochemie, ekonomie a další. Práce je vhodná pro širokou veřejnost, neboť nevyžaduje pokročilé znalosti z pravděpodobnosti či statistiky, jednotlivé metody jsou zaváděny pozvolna a obsahuje praktické příklady pro usnadnění pochopení látky.

Klíčová slova: Analýza rozptylu (ANOVA), pevný a náhodný efekt, lineární smíšený model

Title: Introduction to Linear Mixed Models
Author: Vojtěch Šaroch
Department: Department of Probability and Mathematical Statistics, MFF UK
Supervisor: doc. Mgr. Michal Kulich Ph.D.

Abstract: The thesis describes general procedures of estimation and hypothesis testing for linear statistical models. The models compare groups of observation due to dependent variable. Analysis of variance and linar mixed models are commonly used in the major science like pharmacology, biochemistry, economy and others. The thesis is appropriate for general public, because no advanced knowledge of probability and statistics are required. Particular methods are introduced gently and contain some practical examples for easier understanding of theory.

Keywords: Analysis of variance (ANOVA), fixed and random effect, linear mixed model

Obsah

Úvod

Cílem mé bakalářské práce na téma "Lineární smíšené modely" je seznámit čtenáře s metodami analýzy rozptylu. Seznámíme se nejprve s jednoduchým a dvojným tříděním. Potom přejdeme k jejich zobecněnému modelu, který vznikne tak, že některé parametry modelu začneme považovat za náhodné veličiny generované z normálního rozdělení. Toto zobecnění použijeme například, pokud chceme vyjádřit nejistotu ve výběru zkoumaných jednotek. V práci jsou uvedeny předpoklady jednotlivých modelů, jejich principy odhadovaní a testování hypotéz v těchto modelech. Testy jsou demonstrovány za použití statistického softwaru IBM SPSS Statistics 19.0 na reálných datech. Jedná se o mezinárodní projekt PISA (Programme for International Student Assessment) z roku 2009, ve kterém vybraní 15-ti letí žáci anonymně odpovídali na otázky především z jejich rodiného a školního života (více na stránkách ústavu pro informace ve vzdělávání: http://www.uiv.cz/). IBM SPSS Statistics 19.0 představuje komplexní a snadno použitelnou sadu nástrojů pro analýzu dat a prediktivní analýzu s velice intuitivním ovládáním.

První část této práce se věnuje jednoduchým a dvojným tříděním, kde jsou definovány základní značení a principy, jejichž znalost je potřebná pro modely v druhé části. Ta se už věnuje pevným i náhodným efektům, které jsou součástí lineárních smíšených modelů. Na konci kapitol jsou příklady, jejichž cílem bude zpracování analýzy známky z matematiky (naší závislé proměnné) v závislosti na různých faktorech (škola, pohlaví).

Část I

Analýza rozptylu

Je dáno několik nezávislých výběrů (např. studenti z jednotlivých škol) a chceme srovnávat jejich průměrné znalosti dané látky (testovat, zda-li se rovnají střední hodnoty). Pokud bychom měli školy pouze dvě, mohli bychom bez obav použít dvouvýběrový T–test s předem zvolenou hladinou testu α. Často však potřebujeme mezi sebou porovnat více skupin určitého faktoru. K tomu můžeme použít analýzu rozptylu.

1 Jednoduché třídění

Zaveďme si základní značení. Máme nezávislé výběry z normálního rozdělení se stejným, ale neznámým rozptylem. Označme pozorování náhodného výběru z rozdělení $N(\mu_i, \sigma^2)$ $Y_{i1}, ..., Y_{in_i}$ pro $i = 1, ..., k$. Celkem tedy máme $n = n_1 + ... + n_k$ pozorování. Dále označíme

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij} \ , \qquad Y_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \ .$$

První nazveme součet i-tého výběru a druhý celkový součet. Zadefinujme aritmetický průměr, tedy

$$\overline{Y}_{i\cdot} = \frac{Y_{i\cdot}}{n_i} \ , \qquad \overline{Y}_{\cdot\cdot} = \frac{Y_{\cdot\cdot}}{n} \ .$$

Zavedeme model:

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., k \ ,$$

kde e_{ij} jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$ a μ, α_i a σ^2 jsou neznámé parametry. Pro větší přehlednost jsme místo μ_i zavedli $\mu+\alpha_i$. Tedy střední hodnotu každého výběru jsme rozdělili na jednu společnou pro celý model a odchylku i-té skupiny od společné střední hodnoty, matematicky $E(Y_{ij}) = \mu + \alpha_i$ pro $i = 1, ..., k$ a $j = 1, ..., n_i$. Ovšem takto jsme si zavedli o jeden parametr více, proto odhady parametrů nejsou při daných datech

jednoznačně určené. Řešíme buď přidáním podmínky $\sum_{i=1}^{k} \alpha_i = 0$ nebo za μ položíme nějakou konstantu (nejčastěji 0).

Parametry μ a α_i odhadneme pomocí *metody nejmenších čtverců*. Proto hledáme minimum funkce

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

v závislosti na proměnných μ a α_i. Postupně zderivujeme danou funkci podle μ a α_i a výsledky položíme rovny 0. Dostaneme *soustavu normálních rovnic*

$$n\mu + \sum_{i=1}^{k} n_i \alpha_i = Y_{..} \ ,$$

$$n_i \mu + n_i \alpha_i = Y_{i.} \ , \quad i = 1, ..., k \ ,$$

k této soustavě připojíme podmínku kvůli jednoznačnosti

$$\sum_{t=1}^{k} \alpha_t = 0 \ .$$

Upravíme rovnice na $\mu + \alpha_i = \overline{Y}_{i.}$ a za α_k dosadíme $-\sum_{t=1}^{k-1} \alpha_t$. Tyto rovnice sečteme a dostaneme

$$k\mu = \sum_{t=1}^{k} \overline{Y}_{t.} \ .$$

Za μ dosadíme a dostaneme

$$\frac{1}{k} \sum_{t=1}^{k} \overline{Y}_{t.} + \alpha_i = \overline{Y}_{i.} \ .$$

Označíme-li řešení soustavy jako μ^0 a α_i^0 , dostaneme

$$\mu^0 = \frac{1}{k} \sum_{t=1}^{k} \overline{Y}_{t.} \ , \qquad \alpha_i^0 = \overline{Y}_{i.} - \frac{1}{k} \sum_{t=1}^{k} \overline{Y}_{t.} \ , \ i = 1, ..., k$$

Střední hodnota jednotlivých měření je $EY_{ij} = \mu + \alpha_i$. Po dosazení μ^0 a α_i^0 dostaneme odhad $\overline{Y}_{i.}$ pro všechna i. Podle vět, které mužeme najít např. v knize[2] (str.193-200), platí, že se jedná o *nejlepší nestranný lineární odhad* (NNLO). *Pozn.* : Hypotézu H_0 lze nyní vyjádřit ve tvaru $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$. Tím se nám původní model zredukuje na submodel $Y_{ij} = \mu + e_{ij}$.

1.1 Odvození testové statistiky

Vyjdeme z *celkového součtu čtverců* $SS_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$, který vyjadřuje kvadratické chyby jednotlivých měření od jejich celkového průměru a postupnými úpravami dostaneme

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.} + \overline{Y}_{i.} - \overline{Y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 +$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{n_i}2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

Poslední člen vypadl úplně, protože $\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.}) = 0$.

$SS_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$ nazveme *reziduální součet čtverců*

$SS_A = \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$ nazveme *skupinový součet čtverců*

(SS od anglického "Sum of squares"). Kdybychom nyní testovali hypotézu H_0 o rovnosti středních hodnot, tak při její platnosti bychom očekávali, že SS_A bude blízký nule a SS_e blízké SS_C.

1.2 Rozdělení součtů čtverců

K určení rozdělení SS_C, SS_e a SS_A použijeme následující věty

Věta (o χ^2 rozdělení)**.** *Nechť $X \sim N_n(0, \Sigma)$ a matice A taková, že $A\Sigma$ je idempotentní, potom $Y = X^T AX \sim \chi^2_{tr(A\Sigma)}$*

Důkaz. *viz [2] (str. 67-69).* □

Řekneme, že matice A je *idempotentní*, pokud je čtvercová a platí: $AA = A$

Definujeme *stopu matice A* ($tr(A)$) jako součet diagonálních prvků matice.

Věta. *Nechť $Y_i \sim N(\mu, \sigma^2)$ jsou nezávislé pro $i = 1, ..., n$. Potom*

$$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1}\,.$$

Důkaz. *náznak : výraz můžeme přepsat jako*

$$\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sigma^2} = X^T A X,\ kde\ X = \left(\frac{Y_1-\mu}{\sigma},...,\frac{Y_n-\mu}{\sigma}\right)^T,\ X \sim N_n(0,I_n),$$

Nyní stačí zjistit, zdali matice $A = I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$, kde $\mathbb{1}_n$ je jednotkový vektor, je idempotentní. Což platí $\Rightarrow \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sigma^2} \sim \chi^2_{tr(AI)}$ podle předchozí věty a

$tr(AI_n) = \sum_{i=1}^{n}(1-1/n) = n-1\ .$ $\square$

Rozdělení součtů čtverců máme následující:

SS_C : Za platnosti H_0 tj. $\alpha_i = 0$ pro $i = 1,...,k$ máme Y_{ij} nezávislé, stejně rozdělené a odhad rozptylu σ^2 je

$$\frac{1}{n-1}\sum_{i,j}(Y_{ij}-\overline{Y}_{..})^2 = \frac{SS_C}{n-1}\ ,\qquad SS_C/\sigma^2 \sim \chi^2_{n-1}\ .$$

SS_e : Platí, že $\sum_j(Y_{ij}-\overline{Y}_{..})^2/\sigma^2 \sim \chi^2_{n_i}$ a jsou nezávislé pro $i = 1,...,k$.

$$\frac{SS_e}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{\sum_j(Y_{ij}-\overline{Y}_{..})^2}{\sigma^2}\right) \sim \chi^2_{\sum_{i=1}^{k}(n_i-1)} = \chi^2_{n-k}$$

$$\Rightarrow E\frac{SS_e}{\sigma^2} = n-k \Rightarrow E\frac{SS_e}{n-k} = \sigma^2$$

Z toho plyne, že $E\frac{SS_e}{n-k}$ je nestranný odhad σ^2.

SS_A : Za platnosti H_0 můžeme $\frac{SS_A}{\sigma^2}$ upravit na X^TAX, kde $X \sim N_k(0,\Sigma)$,

$$A = \frac{1}{\sigma^2}\left(diag(n_1,...,n_k) - \frac{1}{n}\mathbf{n}\mathbf{n}^T\right),\ \mathbf{n} = (n_1,...,n_k)^T \text{ a } \Sigma = \sigma^2 diag\left(\frac{1}{n_1},...,\frac{1}{n_k}\right)$$

Matice $A\Sigma$ je idempotentní a $tr(A\Sigma) = k-1$, proto $\frac{SS_A}{\sigma^2} \sim \chi^2_{k-1}$.

Cílem jednoduchého třídění je především testování, zda jednotlivé skupiny daného faktoru jsou statisticky nevýznamně odlišné. Tedy testujeme hypotézu H_0, při které využíváme faktu, že veličina $\frac{SS_A}{\sigma^2}$ má χ^2 rozdělení s $k-1$ stupni volnosti a veličina $\frac{SS_e}{\sigma^2}$ má nezávislé χ^2 rozdělení s $n-k$ stupni volnosti. Jejich podíl má pak F–rozdělení s $(k-1)$ a $(n-k)$ stupni volnosti. Testová statistika $F_A = \frac{SS_A(n-k)}{SS_e(k-1)}$.

Vyjde-li F_A větší než kvantil $F_{1-\alpha,k-1,n-k}$, zamítáme nulovou hypotézu H_0 na hladině významnosti α zamítnout a efekty α_i považovat za nenulové, čili statisticky významné.

Obrázek 1: informace o datech

	Počet pozorování	chybějící data	průměr známky z mat
Celkem	228	4	2,84
Chlapců	125	3	2,94
Dívek	103	1	2,73

1.3 Příklad

Všechny modely budeme aplikovat, jak jsme již zmínili v úvodu, na data z mezinárodního projektu PISA. V této studii byly nejdříve vybrány školy a následně v nich žáci. V příkladu prozkoumáme, zda známka z matematiky se liší dle pohlaví. Použijeme k tomu software IBM SPSS Statistics 19.0. Původní data byla značně rozsáhlá, protože obsahovala 6064 pozorování z celé ČR. Budeme proto porovnávat pouze žáky pražských základních škol, zbyde nám celkově 10 škol s 228 žáky. Po načtení souboru[1] se nám veškerá naše data zobrazí do přehledné tabulky, ve které lze velmi snadno dělat dílčí úpravy. Data už jsou upravená jen na pražské základní školy a obsahují také známky z českého jazyka, na kterých bychom mohli zkoumat stejné závislosti nebo je porovnávat s matematikou. Metodu pro jednoduché třídění (anglicky one-way Anova) vyvoláme, když do příkazového řádku napíšeme ONEWAY "název závislé proměnné"BY "třídící faktor". Tedy v našem případě napíšeme ONEWAY matematika BY pohlavi. Můžeme přidat další příkazy, například popisnou statistiku dat vyvoláme /STATISTICS DESCRIPTIVES[2].

Na výstupu obdržíme tabulku analýzy rozptylu, kde p-hodnota je 0,122, tudíž při klasické hladině významnosti 0,05 nelze zamítnout hypotézu, že průměrná známka chlapců je stejná jako dívek, i když průměr známek u dívek je 2,73 a u chlapců 2,94.

Pozn.: v datech máme několik žáků, u kterých nám chybí známka (ať už z jakéhokoli důvodu). Tito žaci jsou při výpočtech vynecháni.

[1] Na přiloženém CD uložené pod názvem bprace.sav

[2] Příkaz jde také spustit přes menu v záložce Analyze

Obrázek 2: Výstupní tabulka analýzy rozptylu

Známka - matematika

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,619	1	2,619	2,414	,122
Within Groups	240,912	222	1,085		
Total	243,531	223			

2 Dvojné třídění bez interakcí

Pokud se chystáme porovnávat měření v závislosti na dvou třídících znacích (např. z jaké školy studenti jsou a jaké mají pohlaví), použijeme dvojné třídění.

Zavedeme si model, který je intuitivním rozšířením předchozího. Máme náhodné veličiny Y_{ijp} znamenající p-té měření za podmínek i a j. Přidáme si jednu omezující podmínku, že počet pozorování je pro všechny dvojice (i, j) stejný a je roven P. Tato podmínka nám výrazně ulehčí zápis. Pokud jí náš výběr splňuje, pak říkáme, že máme *vyvážená data*. Předpokládejme, že se náhodné veličiny Y_{ijp} řídí modelem

$$M: \quad Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}$$

pro $i = 1, ..., I,\ j = 1, ..., J,\ p = 1, ..., P$, e_{ijp} jsou náhodné veličiny s rozdělením $N(0, \sigma^2)$ a $\mu,\ \alpha_i,\ \beta_j$ a σ^2 jsou neznámé parametry. Přitom α_i jsou tzv. *řádkové efekty* a β_j jsou *sloupcové efekty*. dále $n = IJP$ je počet veličin Y_{ijp}. Jako v jednoduchém třídění zavedeme součty jednotlivých výběrů

$$Y_{.j.} = \sum_{i=1}^{I} \sum_{p=1}^{P} Y_{ijp}\ , \qquad \overline{Y}_{.j.} = \frac{Y_{.j.}}{IP}\ ,$$

ostatní analogicky.

Opět pokračujeme nadále stejně, tudíž výraz

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{p=1}^{P} (Y_{ijp} - \mu - \alpha_i - \beta_j)^2$$

postupně parciálně zderivujeme podle proměnných μ, α_i a β_j a derivace položíme rovny nule. Máme soustavu

$$n\mu + JP\sum_{i=1}^{I}\alpha_i + IP\sum_{j=1}^{J}\beta_j = Y_{...}\ ,$$

$$JP\mu + JP\alpha_i + P\sum_{j=1}^{J}\beta_j = Y_{i..}, \quad i = 1, ..., I,$$

$$IP\mu + P\sum_{i=1}^{I}\alpha_i + IP\beta_j = Y_{.j.}, \quad j = 1, ..., J,$$

Pro jednoznačnost řešení přidáme *reparametrizační rovnice* $\sum_{i=1}^{I}\alpha_i = 0$ a $\sum_{j=1}^{J}\beta_j = 0$.

Získáme řešení: $\mu^0 = \overline{Y}_{...}$, $\alpha_i^0 = \overline{Y}_{i..} - \overline{Y}_{...}$, $\beta_j^0 = \overline{Y}_{.j.} - \overline{Y}_{...}$. Po dosazení do modelu M máme NNLO pro EY_{ijp} a to

$$\widehat{\mu}_{ijp} = \mu^0 + \alpha_i^0 + \beta_j^0 = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...}$$

2.1 Odvození testové statistiky

Pokud bychom v našem modelu M položili $\beta_1 = ... = \beta_J = 0$, tedy pokud by nezáleželo na sloupcových efektech, dostaneme model $M_1 : Y_{ijp} = \mu + \alpha_i + e_{ijp}$, který odpovídá jednoduchému třídění s JP pozorováními v každé skupině. NNLO EY_{ijp} je v modelu M_1 $\widehat{\nu}_{ijp} = \overline{Y}_{i..}$, podobně jako u jednoduchého třídění. Všimněme si, že NNLO je jiný než pro model M.

Položíme-li v modelu M_1 všechny α_i rovny nule, dostaneme model M_2: $Y_{ijp} = \mu + e_{ijp}$. NNLO EY_{ijp} je v tomto případě $\widehat{\tau}_{ijp} = \overline{Y}_{...}$.

Jsme připraveni na vypočtení všech součtů čtverců, které potřebujeme. Celkový součet

$$SS_C = \sum_i\sum_j\sum_p (Y_{ijp} - \overline{Y}_{...})^2 = \sum_{i,j,p} Y_{ijp}^2 - n\overline{Y}_{...}^2$$

Označme ještě $\widehat{\boldsymbol{\mu}}$ jako vektor se složkami $\widehat{\mu}_{ijp}$. Obdobě také označme $\widehat{\boldsymbol{\nu}}_{\boldsymbol{ijp}}$, $\widehat{\boldsymbol{\tau}}_{\boldsymbol{ijp}}$ jako vektory se složkami $\widehat{\nu}_{ijp}$, resp. $\widehat{\tau}_{ijp}$. Protože $\widehat{\boldsymbol{\mu}}^T\widehat{\boldsymbol{\mu}} = \sum_i\sum_j\sum_p \widehat{\mu}_{ijp}^2$, máme po úpravě

$$\widehat{\boldsymbol{\mu}}^T\widehat{\boldsymbol{\mu}} = JP\sum_{i=1}^{I}\overline{Y}_{i..}^2 + IP\sum_{j=1}^{J}\overline{Y}_{.j.}^2 - n\overline{Y}_{...}^2$$

$$\widehat{\boldsymbol{\nu}}^T\widehat{\boldsymbol{\nu}} = JP\sum_{i=1}^{I}\overline{Y}_{i..}^2\,, \quad \widehat{\boldsymbol{\tau}}^T\widehat{\boldsymbol{\tau}} = n\overline{Y}_{...}^2\,,$$

řádkové (SS_A) a sloupcové (SS_B) součty čtverců lze zapsat

$$SS_A = (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\tau}})^T(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\tau}}) = JP\sum_{i=1}^{I}\overline{Y}_{i..}^2 - n\overline{Y}_{...}^2$$

$$SS_B = (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\nu}})^T(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\nu}}) = JP\sum_{j=1}^{J}\overline{Y}_{.j.}^2 - n\overline{Y}_{...}^2$$

a reziduální součet čtverců jako

$$SS_e = \sum_{i,j,p} Y_{ijp}^2 - \widehat{\boldsymbol{\mu}}^T\widehat{\boldsymbol{\mu}} = \sum_{i,j,p} Y_{ijp}^2 - JP\sum_{i=1}^{I}\overline{Y}_{i..}^2 - IP\sum_{j=1}^{J}\overline{Y}_{.j.}^2 + n\overline{Y}_{...}^2$$

pozn.: platí rovnost $SS_e = SS_C - SS_A - SS_B$

Při tomto postupu odhadujeme řádkové efekty ze submodelu M_1 kde už nemáme zastoupené sloupce. Nabízí se otázka, jestli by se výsledky změnily, kdybychom napřed položili $\alpha_i = 0$. Postup by byl analogický, avšak my výpočty nemusíme provádět. Stačí si povšimnout, že ve výsledných vzorcích budou zaměněny jenom symboly I a J. Tedy při předpokladu vyvážených dat jsou výsledky stejné.

2.2 Testování hypotéz

Ve dvojném třídění můžeme testovat 2 hypotézy:

- $H_0^A : \alpha_1 = \alpha_2 = ... = \alpha_I = 0$, rovnost řádkových středních hodnot
- $H_0^B : \beta_1 = \beta_2 = ... = \beta_J = 0$, rovnost sloupcových středních hodnot

Obě hypotézy mají tedy svou testovou statistiku F_A a F_B. Tyto statistiky jsou obdobné jako u jednoduchého třídění. Za platnosti H_0^A má SS_A χ^2 rozdělení s $I-1$ stupni volnosti (s.v.). Za platnosti H_0^B má SS_B χ^2 rozdělení s $J-1$ s.v. Součty čtverců SS_e a SS_C mají χ^2 rozdělení s $n-I-J+1$ a $n-1$ s.v.

Za platnosti H_0^A má $F_A = \left(\frac{SS_A}{I-1}\right)\left(\frac{SS_e}{n-I-J+1}\right)^{-1}$ Fisherovo rozdělení s $I-1$ a $n-I-J+1$ stupni volnosti a za platnosti H_0^B má $F_B = \left(\frac{SS_B}{J-1}\right)\left(\frac{SS_e}{n-I-J+1}\right)^{-1}$ Fisherovo rozdělení s $J-1$

a $n-I-J+1$ s.v. Výsledky z analýzy dvojného třídění zapisujeme nejčastěji do následující tabulky:

výběr	**Součet čtverců** SS	**počet stupňů volnosti** df	**Podíl** $MS = SS/df$	**testová statistika** $F = MS/s^2$
řádky	SS_A	$f_A = I - 1$	SS_A/f_A	F_A
sloupce	SS_B	$f_B = J - 1$	SS_B/f_B	F_B
rezidua	SS_e	$f_e = n - I - J + 1$	$s^2 = SS_e/f_e$	-
celkem	SS_C	$f_C = n - 1$	-	-

2.3 Nevyvážená data

Nyní odvodíme zcela obecný model, kdy máme v každé skupině jiný počet měření. Označme n_{ij} jako počet dat v i-tém řádku a j-tém sloupci. Označme $\mathbf{b}^0 = (\mu, \alpha_1, ..., \alpha_I, \beta_1, ..., \beta_J)^T$. Použijeme MNČ a reparametrizační rovnice zvolíme kvůli zjednodušení výpočtu jako $\mu = 0$ a $\beta_J = 0$. Dostaneme

$$\sum_{i=1}^{I} n_{i.}\alpha_i + \sum_{j=1}^{J-1} n_{.j}\beta_j = Y_{...} \ ,$$

$$\alpha_i = \frac{Y_{i..}}{n_{i.}} - \frac{1}{n_{i.}} \sum_{j=1}^{J-1} n_{ij}\beta_j \ , \quad \forall \ i = 1, ..., I \ ,$$

$$\sum_{i=1}^{I} n_{ij}\alpha_i + n_{.j}\beta_j = Y_{.j.} \ , \quad \forall \ j = 1, ..., J-1 \ ,$$

dosadíme první do třetí rovnice

$$\left(n_{.j} - \sum_{i=1}^{I} \frac{n_{ij}^2}{n_{i.}} \right) \beta_j - \sum_{j \neq j^1}^{J-1} \left(\sum_{i=1}^{I} \frac{n_{ij} n_{ij^1}}{n_{i.}} \right) \beta_{j^1} = Y_{.j.} - \sum_{i=1}^{I} n_{ij} \overline{Y}_{i..} \ , \ \ j, j^1 = 1, ..., J-1 \ .$$

Tuto soustavu můžeme maticově přepsat na $\mathbf{C}\boldsymbol{\beta}_{J-1} = \mathbf{r}$.

Jednotlivé symboly jsou $\boldsymbol{\beta}_{J-1} = (\beta_1, .., \beta_{J-1})$, $\mathbf{r}$ je pravá strana a $\mathbf{C}$ má tvar $c_{jj} = n_{.j} - \sum_{i=1}^{I} \frac{n_{ij}^2}{n_{i.}}$ a $c_{jl} = -\sum_{i=1}^{I} \frac{n_{ij} n_{il}}{n_{i.}}$ pro $l \neq j$.

Označme $\mathbf{M}_{I,J-1}$ matici takovou, že $\mathbf{m}_{ij} = n_{ij}/n_{i.}$. Výsledné parametry můžeme zapsat jako

$$\mathbf{b}^0 = \left(0, \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0_{j-1}, 0\right)^T = \left(0, \overline{\mathbf{Y}}_a - \mathbf{M}\mathbf{C}^{-1}\mathbf{r}, \mathbf{C}^{-1}\mathbf{r}, 0\right)$$

kde $\overline{\mathbf{Y}}_a = (\overline{Y}_{1..}, .., \overline{Y}_{I..})^T$.

2.4 Typ I a III součtu čtverců

Označme $R(\alpha|\mu, \beta)$ jako redukci modelu $M: \ Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}$ položením $\alpha_i = 0$ a $R(\alpha|\mu)$ jako redukci modelu M_1: $Y_{ijp} = \mu + \alpha_i + e_{ijp}$ o parametry α_i pro $i = 1, ..., I$. V $R(\alpha|\mu)$ tedy nebereme v úvahu sloupcové efekty. Při postupu s vyváženými daty jsme došli k závěru, že $R(\alpha|\mu) = R(\alpha|\mu, \beta)$. S nevyváženými daty ovšem

$$R(\alpha|\mu, \beta) = \sum_{i=1}^{I} n_{i.}\overline{Y}^2_{i..} + \boldsymbol{\beta}^{0T}_{j-1}\mathbf{r} - \sum_{j=1}^{J} n_{.j}\overline{Y}^2_{.j.} \ ,$$

$$R(\alpha|\mu) = \sum_{i=1}^{I} n_{i.}\overline{Y}^2_{i..} - n\overline{Y}^2_{...} \ .$$

Druhý způsob je výpočtově jednodušší, protože neuvažujeme β. Pro výpočet používáme 4 typy součtů čtverců podle toho, jakým principem volíme modely pro odhady parametrů. Pokud předpokládáme model dvojného třídění s nevyváženými daty, vždy bychom měli používat $R(\alpha|\mu, \beta)$ a $R(\beta|\mu, \alpha)$.

- **Typ I**: Princip je, že budujeme náš model od začátku. Tudíž zvolíme první faktor (např. α), který testujeme jen vůči modelu M_2, tedy $R(\alpha|\mu)$. Další faktor testujeme vůči předchozímu vzniklému modelu, tedy $R(\beta|\mu, \alpha)$.

- **Typ III**: Zde předpokládáme model a nulovost každého parametru testujeme vůči němu. Tedy v případě dvojného třídění $R(\alpha|\mu, \beta)$ a $R(\beta|\mu, \alpha)$.

Typ III je asi nejužívanější pro nevyvážená data. U vyvážených dat Typ I a Typ III vychází stejně, tudíž můžeme volit ten výpočtově schůdnější. Testování hypotéz už dále probíhá stejně, jen si musíme předem určit, jaký typ budeme používat.

Pozn.: V příkladech v bakalářské práci používáme pro výpočet součtu čtverců Typ III.

3 Dvojné třídění s interakcemi

U dvojného třídění se často stává, že se řádkové a sloupcové efekty jen prostě nesčítají jak to předpokládá model uvedený na začátku 2. kapitoly. V takových situacích zavádíme model

$$M \ : \quad Y_{ijp} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijp} \, .$$

Parametry λ_{ij} nazveme *interakce*. Interakcí rozumíme jev, při kterém kombinace obou faktorů může mít na výslednou hodnotu sledovaného znaku rozdílný účinek než činí součet účinku každého faktoru uvažovaného zvlášť. Značení a předpoklady u tohoto modelu jsou stejné jako u bezinterakčního třídění. λ_{ij} je neznámý parametr. Opět derivujeme výraz (derivujeme navíc i přes λ_{ij}) :

$$\sum_{i=1}^{I}\sum_{j=J}^{I}\sum_{p=1}^{P}(Y_{ijk} - \mu - \alpha_i - \beta_j - \lambda_{ij})^2$$

Dostaneme soustavu normálních rovnic

$$n\mu + JP\sum_{i=1}^{I}\alpha_i + IP\sum_{j=1}^{J}\beta_j + P\sum_{j=1}^{J}\sum_{i=1}^{I}\lambda_{ij} = Y_{...} \, ,$$

$$JP\mu + JP\alpha_i + P\sum_{j=1}^{J}\beta_j + P\sum_{j=1}^{J}\lambda_{ij} = Y_{i..}, \quad i = 1, ..., I,$$

$$IP\mu + P\sum_{i=1}^{I}\alpha_i + IP\beta + P\sum_{i=1}^{I}\lambda_{ij} = Y_{.j.}, \quad j = 1, ..., J,$$

$$P\mu + P\alpha_i + P\beta_j + P\lambda_{ij} = Y_{ij.}, \ i = 1, ..., I, \ \ j = 1, ..., J \, .$$

Rovnic je $IJ + I + J + 1$, avšak hodnost matice soustavy je IJ. Přidáme reparametrizační rovnice:

$$\sum_{i=1}^{I}\alpha_i = 0, \ \sum_{j=1}^{J}\beta_j = 0, \ \sum_{i=1}^{I}\lambda_{ij} = 0, \ j = 1, ..., J, \ \sum_{j=1}^{J}\lambda_{ij} = 0, \ i = 1, ..., I \, .$$

Po jejich přidání dostaneme řešení

$$\mu^0 = \overline{Y}_{...}, \ \ \alpha_i^0 = \overline{Y}_{i..} - \overline{Y}_{...}, \ \ \beta_j^0 = \overline{Y}_{.j.} - \overline{Y}_{...}, \ \ \lambda_{ij}^0 = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$$

NNLO pro EY_{ijp} označme $\widehat{\mu}_{ijp} = \mu^0 + \alpha_i^0 + \beta_j^0 + \lambda_{ij}^0 = \overline{Y}_{ij.}$

3.1 Testování hypotéz

Vytvoříme submodel $M_1 : Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}$ a z něj pokračujeme jako v bezinteračním modelu ze druhé kapitoly. Odtud dostaneme SS_A, SS_B a SS_C stejné jako v předchozím případě. Reziduální SS dostaneme

$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{p=1}^{P} Y_{ijp}^2 - P\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{Y}_{ij.}^2 \, .$$

Nyní ovšem neplatí, že $SS_C = SS_A + SS_B + SS_e$. Proto si definujme SS_{AB} jako reziduální součet čtverců pro případ bez interakcí, tedy

$$SS_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij.} - \mu^0 - \alpha_i^0 - \beta_j^0\right)^2 = P\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{Y}_{ij.}^2 - JP\sum_{i=1}^{I}\overline{Y}_{i..}^2 - IP\sum_{j=1}^{J}\overline{Y}_{.j.}^2 + n\overline{Y}_{...}^2$$

Výsledky můžeme opět zapsat do tabulky analýzy rozptylu, ve které přibyl jeden řádek.

výběr	**Součet čtverců** SS	**počet stupňů volnosti** df	**Podíl** $MS = SS/df$	**testová statistika** $F = MS/s^2$
řádky	SS_A	$f_A = I - 1$	SS_A/f_A	F_A
sloupce	SS_B	$f_B = J - 1$	SS_B/f_B	F_B
interakce	SS_{AB}	$f_{AB} = (I-1)(J-1)$	SS_{AB}/f_{AB}	F_{AB}
rezidua	SS_e	$f_e = n - IJ$	$s^2 = SS_e/f_e$	-
celkem	SS_C	$f_C = n - 1$	-	-

Je vhodné si povšimnout, že $SS_e + SS_{AB}$ dá hodnotu SS_e u dvojného třídění bez interakcí. Podrobnější model s interakcemi vede k rozštěpení reziduálního řádku v tabulce analýzy rozptylu dvojného třídění bez interakcí. Na třetím řádku dostaneme novou testovou statistiku F_{AB}, ze které můžeme testovat hypotézu o nulovosti interakce, přesněji $H_0^{AB} : \lambda_{ij} = 0 \ , \quad i = 1, ..., I$ a $j = 1, ..., J$. Za přítomnosti interakcí je nutná opatrnost při interpetaci rozdílů mezi efekty A a B. Může se totiž stát, že některé interakce jsou mnohem výraznější než příšlušné efekty.

3.2 Příklad

Zkoumáme závislost známky z matematiky na dvou faktorech, na pohlaví žáků a na jednotlivých základních školách. Uvažujme i možnou interakci mezi nimi. Procedura se v softwaru IBM SPSS Statistic 19.0 vyvolá pomocí příkazu: UNIANOVA matematika BY pohlavi škola. Tento příkaz zařadí oba třídící znaky automaticky do pevných efektů.

Obrázek 3: Výstupní tabulka dvojného třídění

Tests of Between-Subjects Effects

Dependent Variable:Známka - matematika

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
pohlaví	1,825	1	1,825	1,769	,185
škola	24,118	9	2,680	2,598	,007
pohlaví * škola	6,513	9	,724	,702	,707
Error	210,430	204	1,032		
Total	2055,000	224			

Interpretace tabulky je následující (pokud si stanovíme $\alpha = 0,05$)

- První hypotézu nezamítáme, p-hodnota je 0,185. Průměrná známka z matematiky se tedy u chlapců a dívek neliší (stejný výsledek nám dalo i jednoduché třídění).
- Sloupcové efekty (neboli jednotlivé školy) mají p-hodnotu nižší než je hladina významnosti α, proto hypotézu zamítáme. Průměrné známky z matematiky na pražských základních školách považujeme za různé.
- U testu na interakce je p-hodnota velmi vysoká, proto interakce není statisticky významná.

Z výsledku můžeme tvrdit, že faktor pohlaví neovlivňuje známku z matematiky na hladině významnosti α, jelikož nezamítáme ani jednu z hypotéz H_0^A a H_0^{AB}.

Část II

Lineární smíšené modely

4 Pevné a náhodné efekty

U každého efektu v předchozí části jsme odhadovali jeho střední hodnotu a rozptyl jsme předpokládali nulový. Jednalo se o pevné efekty. Jsou však situace, kdy náš hlavní zájem spočívá v odchylkách uvnitř jednotlivých skupin, tedy odhadujeme rozptyl efektu. Tyto efekty nazveme náhodné. Podrobněji

- **Pevné efekty**: Působí dlouhodobě a neměnně na skupiny pozorování. Faktor obsahuje úrovně, které nás konkrétně zajímají (například srovnání účinosti některých léčiv). Úrovní faktoru se zde rozumí určitá hodnota kvantitativního znaku. Jednotlivé úrovně efektu tedy nejsou stanoveny náhodně a budou stejné i při opakování pokusu s jinými daty. Další příklad je pohlaví, kde máme jen dvě úrovně, muže a ženu.

- **Náhodné efekty**: Typickým příkladem náhodného efektu je jedinec z nějaké populace. Vyberu k stejně starých chlapců a porovnávám jejich výšku. Nejde nám tak o to, který z chlapců je nejvyšší, ale zda se jejich výška statisticky významně liší. Jednotlivé úrovně efektu tu jsou stanoveny náhodně (pokud bychom např. při opakování testu vybrali jiné jedince ze stejné skupiny, můžeme dostat výrazně odlišné výsledky).

Modely z předchozích kapitol byli určeny hlavně pro pevné (fixní) efekty. Model nazveme *smíšeným*, pokud obsahuje jak pevné, tak i náhodné efekty. Náhodné efekty také používáme, pokud chceme do modelu zahrnout korelace mezi pozorováními.

4.1 Struktury rozptylových matic

Nyní budeme pracovat s modelovou rovnicí $Y_{ij} = \mu + \alpha_i + e_{ij}$, kde μ a e_{ij} jsou stejné jako v jednoduchém třídění, tedy $e_{ij} \sim N(0, \sigma_e^2)$ pro všechna $i = 1..I, j = 1..J$ a σ_e^2 s μ jsou neznámé parametry. Avšak nyní nově α_i je náhodný efekt s nulovou střední hodnotou, rozptylem σ_α^2 stejným pro všechna α_i a nezávislým na e_{ij} a ostatních α_j. Tedy platí $E(\alpha_i \alpha_j) = 0$ pro $i \neq j$ a $E(\alpha_{i_1} e_{ij}) = 0$, $\forall i_1, i, j$.

Při využití předpokladu nulové střední hodnoty platí, že

$$Var(\alpha_i) = E(\alpha_i^2) - (E\alpha_i)^2 = E(\alpha_i^2) - 0 = E(\alpha_i^2) = \sigma_\alpha^2 \ .$$

Pro kovarianci mezi prvky modelu Y_{ij} a $Y_{i_1 j_1}$ dostaneme

$$Cov(Y_{i_1 j_1}, Y_{ij}) = \begin{cases} \sigma_\alpha^2 + \sigma_e^2 & pro \quad i = i_1, \quad j = j_1 \\ \sigma_\alpha^2 & pro \quad i = i_1, \quad j \neq j_1 \\ 0 & pro \quad i \neq i_1 \end{cases}$$

Podle toho vypadá rozptylová matice z i-té skupiny

$$Var(Y_{i1}, .., Y_{in_i}) = \sigma_e^2 \mathbf{I} + \sigma_\alpha^2 \mathbf{J} \ ,$$

kde $\mathbf{I}$ je diagonální matice a $\mathbf{J}$ čtvercová matice mající na všech místech jednotku. Nyní nás zajímá rozptylová matice pro všechna pozorování ze všech výběrů. Pro větší přehlednost předpokládejme, že máme 3 výběry a každý má stejně pozorování. Matice rozptylu vypadá následovně:

$$Var(\mathbf{Y}) = \begin{bmatrix} \sigma_e^2 \mathbf{I} + \sigma_\alpha^2 \mathbf{J} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \mathbf{I} + \sigma_\alpha^2 \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \mathbf{I} + \sigma_\alpha^2 \mathbf{J} \end{bmatrix}$$

Pro obecný případ, si musíme nejdříve definovat *přímý součet matic* $\oplus \sum$ jako

$$\oplus \sum_{i=1}^{n} \mathbf{A_i} = \begin{bmatrix} \mathbf{A_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A_n} \end{bmatrix}$$

kde $\mathbf{A_i}$ jsou čtvercové matice o rozměrech $n_i \times n_i$.

Máme k výběrů s n_i porozováními, rozptylová matice má tvar

$$\oplus \sum_{i=1}^{k} (\sigma_e^2 \mathbf{I_i} + \sigma_\alpha^2 \mathbf{J_i}) \ ,$$

kde $\mathbf{J_i}$ a $\mathbf{I_i}$ jsou čtvercové matice o rozměrech n_i, $\mathbf{I_i}$ je diagonální matice a $\mathbf{J_i}$ matice mající na všech místech jednotky pro $\forall\, i = 1, ..., k$ Tato rozptylová matice se liší od matic modelů třídění, které byly čistě diagonální, tedy $\sigma_e^2 \mathbf{I_n}$, kde n je počet všech pozorování. Při sestavování tabulky analýzy rozptylu pro pevný i náhodný efekt dostáváme základní kostru stejnou. Máme vyvážená data, I skupin, každá s P pozorováními.

	součet čtverců	$d.f.$	střední čtverce
střední hodn.	$SS_M = IP\overline{Y}_{..}^2$	$f_m = 1$	$MS_M = SS_M$
skupiny	$SS_A = \sum_{i=1}^{I} P(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$f_A = I - 1$	$MS_A = SS_A/f_A$
rezidua	$SS_e = \sum_{i=1}^{I} \sum_{j=1}^{P} (Y_{ij} - \overline{Y}_{i.})^2$	$f_e = I(P-1)$	$MS_e = SS_e/f_e$
celkem	$SS_C = \sum_{i=1}^{I} \sum_{j=1}^{P} Y_{ij}^2$	$f_C = n - 1$	-

avšak rozdíl máme ve středních hodnotách součtů čtverců a středních čtverců (anglicky mean square). Způsobuje to rozdílná definice α_i. Pro pevný efekt máme $E\alpha_i = \alpha_i$ a nulový rozptyl, naproti tomu pro náhodný efekt máme nulovou střední hodnotu a rozptyl je σ_α^2.

Vypočítáme střední hodnoty středních čtverců

$$MS_M = \frac{1}{IP}\left(\sum_{i=1}^{I}\sum_{j=1}^{P} Y_{ij}\right)^2 = \frac{1}{IP}\left(\sum_{i=1}^{I}\sum_{j=1}^{P}(\mu + \alpha_i + e_{ij})\right)^2 =$$

$$= \frac{1}{IP}\left(IP\mu + P\sum_{i=1}^{I}\alpha_i + \sum_i\sum_j e_{ij}\right)^2$$

pokud α_i jsou pevné efekty, střední hodnota MS_M je

$$E(MS_M) = \frac{1}{IP} E\left(IP\mu + P\sum_{i=1}^{I}\alpha_i + \sum_{i=1}^{I}\sum_{j=1}^{P} e_{ij}\right)^2$$

$$IPE\left(\mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i\right)^2 + 2E\left(\mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i\right) E\left(\sum_{i=1}^{I}\sum_{j=1}^{P} e_{ij}\right) + \frac{1}{IP} E\left(\sum_i\sum_j e_{ij}\right)^2$$

jelikož platí, že $E(e_{ij}e_{i_1j_1}) = 0$ pro $i \neq i_1$ nebo pro $i = i_1$, $j \neq j_1$, vyjde

$$IP\left(\mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i\right)^2 + 0 + \frac{1}{IP}\sum_{i=1}^{I}\sum_{j=1}^{P}E(e_{ij}^2) = IP\left(\mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i\right)^2 + \sigma_e^2$$

Nyní pokud uvažujeme α_i jako náhodné efekty, střední hodnota je

$$E(MS_M) = IPE\mu^2 + \frac{P}{I}E\left(\sum_{i=1}^{I}\alpha_i\right)^2 + \frac{1}{IP}E\left(\sum_{i=1}^{I}\sum_{j=1}^{P}e_{ij}\right)^2 = IP\mu^2 + P\sigma_\alpha^2 + \sigma_e^2 \ .$$

Podobně lze odvodit i další střední hodnoty čtverců. Zde jsou některé další uspořádány v tabulce.

pevný efekt		E(MS)	náhodný efekt		
$IP(\mu + 1/I\sum_{i=1}^{I}\alpha_i)^2$	$+\sigma_e^2$	$= E(MS_M) =$	$IP\mu^2$	$+P\sigma_\alpha^2$	$+\sigma_e^2$
$\frac{P}{I-1}\sum_{i=1}^{I}(\alpha_i - 1/I\sum_{i=1}^{I}\alpha_i)^2$	$+\sigma_e^2$	$= E(MS_A) =$		$+P\sigma_\alpha^2$	$+\sigma_e^2$
	$+\sigma_e^2$	$= E(MS_e) =$			$+\sigma_e^2$

.

V modelu s náhodným efektem při odhadování rozptylů $\widehat{\sigma}_e^2$ a $\widehat{\sigma}_\alpha^2$ využíváme střední hodnoty z předcházející tabulky. Tedy $\widehat{\sigma}_e^2 = MS_e$ a $P\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_e^2 = MS_A$.

Tedy $\widehat{\sigma}_\alpha^2 = (MS_A - MS_e)/P$ je odhad rozptylu σ_α^2.

5 Smíšený model

Pro zavedení smíšeného modelu potřebujeme minimálně dva efekty. Upozorněme nejprve, že dvojné třídění je speciální případ tohoto modelu, kde máme 2 pevné efekty. Rovnou budeme předpokládat i interakci mezi nimi.

$$\text{model } M: \quad Y_{ijp} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijp}$$

kde α_i jsou pevné efekty, β_j a γ_{ij} jsou náhodné efekty s rozptyly σ_β^2 a σ_γ^2 a nulovou střední hodnotou pro všechna i, j. Připomeňme, že máme vyvážená data a že I je počet i-tých řádkových skupin, J počet sloupcových skupin, P počet pozorování ve skupině a $n = IJP$. Pokud máme v interakci mezi efekty alespoň jeden náhodný efekt, tak považujeme interakci také za náhodný efekt. Při sestavování součtů čtverců a středních čtverců platí stejná

pravidla, jako byla u dvojného třídění s interakcemi. Definujeme si $\overline{\alpha}_{\cdot}$, $\overline{\beta}_{\cdot}$, $\overline{\gamma}_{i\cdot}$, $\overline{\gamma}_{\cdot j}$, $\overline{\gamma}_{\cdot\cdot}$ a $\overline{e}_{\cdots}$ obdobně jako $\overline{Y}_{\cdots}$ v předchozí části.

Takže například platí, že $\overline{\gamma}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J}\gamma_{ij}$. Přepíšeme aritmetické průměry pomocí nově zavedených značení:

$$\overline{Y}_{\cdots} = \mu + \overline{\alpha}_{\cdot} + \overline{\beta}_{\cdot} + \overline{\gamma}_{\cdot\cdot} + \overline{e}_{\cdots}$$

$$\overline{Y}_{i\cdot\cdot} = \mu + \alpha_i + \overline{\beta}_{\cdot} + \overline{\gamma}_{i\cdot} + \overline{e}_{i\cdot\cdot}$$

$$\overline{Y}_{\cdot j\cdot} = \mu + \overline{\alpha}_{\cdot} + \beta_j + \overline{\gamma}_{\cdot j} + \overline{e}_{\cdot j\cdot}$$

$$\overline{Y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \overline{e}_{ij\cdot}$$

a vypočteme střední hodnoty středních čtverců našeho smíšeného modelu

$$E(MS_M) = nE\overline{Y}_{\cdots}^2 = nE(\mu + \overline{\alpha}_{\cdot} + \overline{\beta}_{\cdot} + \overline{\gamma}_{\cdot\cdot} + \overline{e}_{\cdots})^2 = nE(\mu + \overline{\alpha}_{\cdot})^2 + nE(\overline{\beta}_{\cdot} + \overline{\gamma}_{\cdot\cdot} + \overline{e}_{\cdots})^2 =$$

$$= n(\mu+\overline{\alpha}_{\cdot})^2 + nE\left(\frac{1}{J}\sum_{j=1}^{J}\beta_j\right)^2 + \frac{P}{IJ}E\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}\right)^2 + \sigma_e^2 = n(\mu+\overline{\alpha}_{\cdot})^2 + IP\sigma_\beta^2 + P\sigma_\gamma^2 + \sigma_e^2$$

$$E(MS_A) = \frac{JP}{I-1}\sum_{i}^{I}E(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2 = \frac{JP}{I-1}\sum_{i}^{I}E(\alpha_i + \overline{\gamma}_{i\cdot} + \overline{e}_{i\cdot\cdot} - \overline{\alpha}_{\cdot} - \overline{\gamma}_{\cdot\cdot} - \overline{e}_{\cdots})^2 =$$

$$= \frac{JP}{I-1}\sum_{i}^{I}E(\alpha_i - \overline{\alpha}_{\cdot})^2 + \frac{JP}{I-1}\sum_{i}^{I}\left[E(\overline{\gamma}_{i\cdot}^2) - 2E(\overline{\gamma}_{i\cdot}\overline{\gamma}_{\cdot\cdot}) + E(\overline{\gamma}_{\cdot\cdot}^2)\right] + \sigma_e^2 =$$

$$= \frac{JP}{I-1}\sum_{i}^{I}E(\alpha_i - \overline{\alpha}_{\cdot})^2 + \frac{P}{I-1}\sum_{i}^{I}\left[\sigma_\gamma^2 - 1/I\sigma_\gamma^2\right] + \sigma_e^2 = \frac{JP}{I-1}\sum_{i}^{I}E(\alpha_i - \overline{\alpha}_{\cdot})^2 + P\sigma_\gamma^2 + \sigma_e^2$$

$$E(MS_B) = \frac{IP}{J-1}\sum_{j}^{J}E(\beta_j - \overline{\beta}_{\cdot} + \overline{\gamma}_{\cdot j} - \overline{\gamma}_{\cdot\cdot})^2 + \sigma_e^2 = IP\sigma_\beta^2 + P\sigma_\gamma^2 + \sigma_e^2$$

$$E(MS_{AB}) = \frac{P}{(I-1)(J-1)}\sum_{i}^{I}\sum_{j}^{J}E(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 =$$

$$= \frac{P}{(I-1)(J-1)} \sum_{i}^{I} \sum_{j}^{J} E(\gamma_{ij} - \overline{\gamma}_{i.} - \overline{\gamma}_{.j} + \overline{\gamma}_{..} + \overline{e}_{ij.} - \overline{e}_{i..} - \overline{e}_{.j.} + \overline{e}_{...})^2$$

Upravíme mezivýsledek $E(\gamma_{ij} - \overline{\gamma}_{i.} - \overline{\gamma}_{.j} + \overline{\gamma}_{..})^2$

$$E(\gamma_{ij}^2) + 2(-\frac{1}{I} - \frac{1}{J} + \frac{1}{IJ})E\gamma_{ij}^2 + \left[E(\overline{\gamma}_{i.}^2) + E(\overline{\gamma}_{i.}\overline{\gamma}_{.j}) - \frac{1}{I}E(\overline{\gamma}_{i.}^2) + E(\overline{\gamma}_{..} - \overline{\gamma}_{.j})^2\right] =$$

$$= \sigma_\gamma^2 \left(1 + \frac{2(1-I-J)}{IJ} + \frac{1}{J} + \frac{J-1}{IJ}\right) = \sigma_\gamma^2 \left(\frac{(I-1)(J-1)}{IJ}\right)$$

Pro výpočet členů $\overline{e}$ použijeme podobného principu, dopočítáme

$E(MS_{AB}) = P\sigma_\gamma^2 + \sigma_e^2$

$$E(MS_E) = \frac{1}{IJ(P-1)} \sum_{i}^{I} \sum_{j}^{J} \sum_{p}^{P} (Y_{ijp} - \overline{Y}_{ij.})^2 = \frac{1}{IJ(P-1)} \sum_{i}^{I} \sum_{j}^{J} \sum_{p}^{P} (e_{ijp} - \overline{e}_{ij.})^2 = \sigma_e^2 \ .$$

Při výpočtech jsme hlavně využívali vlastnosti střední hodnoty a nezávislosti náhodných efektů. Naše výsledky přepíšeme pro přehlednost do tabulky

střední hodnoty čtverců smíšeného modelu,

$E(MS_M) =$	$IJP(\mu + \overline{\alpha}_{.})^2$	$+IP\sigma_\beta^2$	$+P\sigma_\gamma^2$	$+\sigma_e^2$
$E(MS_A) =$	$\frac{JP}{I-1}\sum_i^I(\alpha_i - \overline{\alpha}_{.})^2$		$+P\sigma_\gamma^2$	$+\sigma_e^2$
$E(MS_B) =$		$+IP\sigma_\beta^2$	$+P\sigma_\gamma^2$	$+\sigma_e^2$
$E(MS_{AB}) =$			$+P\sigma_\gamma^2$	$+\sigma_e^2$
$E(MS_E) =$				$+\sigma_e^2$

5.1 Problém záporného rozptylu

Rozptyl je podle definice vždy kladný. Už samotná interpretace záporného rozptylu nedává žádný smysl. Bohužel i přesto rozptyl odhadnutý ze středních hodnot středních čtverců v našich metodách může k tomuto výsledku vést, viz tento jednoduchý příklad: Máme model s náhodným efektem ve kterém porovnáváme 2 skupiny. První skupina má naměřené hodnoty hodnoty 19, 17, 15 a druhá 25, 5, 15. Po dosazení máme tabulku:

výběr	součty čtverců	MS	očekávaný MS
skupinový	$3((17-16)^2+(15-16)^2)=6$	6	$3\sigma_\alpha^2+\sigma_e^2$
reziduální	$2^2+2^2+2*10^2=208$	52	σ_e^2
celkový	$3^2+1+1+9^2+11^2+1=214$		

Odtud vychází, že $\widehat{\sigma}_e^2 = 52$ a $\widehat{\sigma}_\alpha^2 = -15\frac{1}{3}$. Problém nastává i při vyšším počtu efektů v modelu jak pro vyvážená, tak i pro nevyvážená data. V tomto modelu neexistuje žádný přirozený nástroj, který by tento problém eliminoval, avšak jsou zde nějaké postupy, které se snaží tuto nepříjemnou věc odbourat, často na úkor něčeho jiného.

1. Nahradíme záporný rozptyl nulou, avšak zde je problém, že nyní neodpovídají střední hodnoty čtverců.

2. Záporný rozptyl značí, že tento efekt bude z modelu odstraněn. Při přechodu na submodel může problém záporného rozptylu přetrvat.

3. Řekneme, že máme špatně celý model, pokusíme se překontrolovat data a podíváme se po novém (například změníme náhodný efekt na pevný apod.).

5.2 Předpoklady normality

Až dosud jsme nemuseli předpokládat žádné rozdělení všech náhodných efektů. Všechny výsledky platí pro jakékoli rozdělení. Avšak až dosud jsme se záměrně vyhýbali testu hypotéz ve smíšeném modelu, protože ze stávajících předpokladů nejde zatím vyvodit. Přidejme tedy předpoklad, že všechny náhodné efekty (včetně e) mají normální rozdělení s nulovou střední hodnotou. Potom podle knihy[1] (str.408-410) platí:

Věta. *Nechť $MS = SS/f$,kde f jsou stupně volnosti a SS součty čtverců, potom platí, že $SS/E(MS) \sim \chi_f^2$ a součty čtverců jsou vzájemně nezávislé.*

Důkaz. *Výraz $SS/E(MS)$ si rozepíšeme jako kvadratickou formu $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ a aplikujeme větu o χ^2 rozdělení. I když nemáme varianční matici tvaru $\sigma_e^2\mathbf{I}$, přesto vždy můžeme najít matici $\boldsymbol{A}$ takovou, že matice $\boldsymbol{AV}$ je idempotentní.*

Ukážeme si to na příkladě modelu s jedním náhodným efektem. $\boldsymbol{V} = \sigma_e^2\mathbf{I} + \sigma_\alpha^2 \oplus \sum_{i=1}^{k} \mathbf{J}$ *, kde* $\boldsymbol{I}$ *a* $\boldsymbol{J}$ *jsou jako v kapitole 3.1 a* $N = kn$ *.*

$$SS_A = \boldsymbol{y}^T \left(n^{-1} \oplus \sum_{i=1}^{k} \mathbf{J} - N^{-1}\mathbb{1}_N^T\mathbb{1}_N \right) \boldsymbol{y} \quad , \qquad E(MS_A) = n\sigma_\alpha^2 + \sigma_e^2$$

Tedy pro $SS_A/E(MS_A)$ *máme matici*

$$\boldsymbol{A} = \frac{1}{n\sigma_\alpha^2 + \sigma_e^2} \left(n^{-1} \oplus \sum_{i=1}^{k} \mathbf{J} - N^{-1}\mathbb{1}_N^T\mathbb{1}_N \right)$$

$$\boldsymbol{AV} = \left[\sigma_e^2 \left(n^{-1} \oplus \sum_{i=1}^{k} \mathbf{J} - N^{-1}\mathbb{1}_N^T\mathbb{1}_N \right) + \sigma_\alpha^2 \left(\oplus \sum_{i=1}^{k} \mathbf{J} - nN^{-1}\mathbb{1}_N^T\mathbb{1}_N \right) \right] / \left(n\sigma_\alpha^2 + \sigma_e^2 \right) =$$

$$=^{\oplus} \sum_{i=1}^{k} \mathbf{J} - nN^{-1}\mathbb{1}_N^T\mathbb{1}_N$$

odtud už je zkoumaná vlastnost vidět. □

Je dobré si uvědomit, že ve jmenovateli se tedy mohou vyskytnout kromě σ_e^2 také složitější vzorce v závisloti na $E(MS)$. Například pro model s jedním náhodným efektem máme

$$\frac{SS_A}{n\sigma_\alpha^2 + \sigma_e^2} \sim \chi^2_{k-1} \quad , \quad \frac{SS_e}{\sigma_e^2} \sim \chi^2_{k(n-1)}$$

odtud a za použití předpokladu nezávislosti můžeme odvodit, že

$$\widehat{\sigma}_\alpha^2 = \frac{MS_A - MS_e}{n} \text{ má přibližně rozdělení } \frac{n\sigma_\alpha^2 + \sigma_e^2}{n(k-1)}\chi^2_{k-1} - \frac{\sigma_e^2}{kn(n-1)}\chi^2_{kn-k} \, .$$

Avšak skutečný tvar distribuční funkce nemůžeme takto získat, protože se ve výrazu objevují obě σ a jsou neznámé. Navíc druhý člen se odečítá, tudíž konečné číslo může být záporné. To vysvětluje, proč nám občas odhad rozptylu náhodného efektu v této metodě vychází záporný. Na druhou stranu nám však odhad σ_e^2 členů e_{ijp} tak vyjít nikdy nemůže, protože

$$\widehat{\sigma}_e^2 = MS_e \sim \frac{\sigma_e^2}{f_e}\chi^2_{f_e} \quad ,$$

kde f_e jsou příslušné stupně volnosti.

5.3 Testování hypotéz

Opět se pokusme vytvořit testovou statistiku založenou na F-rozdělení, která nám bude testovat nulové střední hodnoty u pevných efektů a nulovost rozptylů u náhodných efektů. Střední hodnoty středních čtverců naznačí, které střední čtverce jsou vhodné do naší testové statistiky. Tedy pokud máme smíšený model zavedený na začátku 5. kapitoly, bude podíl

- MS_{AB}/MS_e vhodný k testování hypotézy $H_\gamma : \quad \sigma^2_\gamma = 0$, neboť $\frac{E(MS_{AB})}{E(MS_e)} = = \frac{P\sigma^2_\gamma+\sigma^2_e}{\sigma^2_e} = 1$, pokud $\sigma^2_\gamma = 0$
- MS_B/MS_{AB} vhodný k testování hypotézy $H_\beta : \quad \sigma^2_\beta = 0$, neboť $\frac{E(MS_B)}{E(MS_{AB})} = = \frac{IP\sigma^2_\beta+P\sigma^2_\gamma+\sigma^2_e}{P\sigma^2_\gamma+\sigma^2_e} = 1$ pro $\sigma^2_\beta = 0$

V některých případech tabulka středních hodnot čtverců nemusí mít u testů hypotéz vždy tak jednoduše určený jmenovatel. Pokud bychom předpokládali situaci, že máme určené střední hodnoty středních čtverců následujícím způsobem

$$\begin{aligned}
E(M_1) = \quad & k_1\sigma^2_b \quad +k_2\sigma^2_{c:b} \quad +k_3\sigma^2_{ab} \quad +k_4\sigma^2_{ac:b} + \sigma^2_e \\
E(M_2) = \quad & \qquad +k_2\sigma^2_{c:b} \qquad\quad\; +k_4\sigma^2_{ac:b} + \sigma^2_e \\
E(M_3) = \quad & \qquad\qquad\quad +k_3\sigma^2_{ab} \quad +k_4\sigma^2_{ac:b} + \sigma^2_e \\
E(M_4) = \quad & \qquad\qquad\qquad\quad\; +k_4\sigma^2_{ac:b} + \sigma^2_e
\end{aligned}$$

kde M_1 až M_4 jsou postupně MS_B, $MS_{C:B}$, MS_{AB} a $MS_{AC:B}$. Taková tabulka vznikne, pokud bychom považovali faktor C za *vnořený* do B (C:B). Faktor C je tedy vnořený do faktoru B, pokud skupiny faktoru C jsou různé pro jednotlivé skupiny faktoru B. Například pokud máme 2 učební kurzy, matematiku a zeměpis a každý kurz navíc vyučují 3 cvičící. Takže máme v podstatě 6 kurzů, kde hlavní faktor je matematika a zeměpis a vnořený faktor je číslo skupiny. Avšak matematika a zeměpis z pohledu vnořeného faktoru nemají nic společného. Vnořené efekty nikdy nevystupují jako hlavní efekty.

Nebudeme se vnořenými faktory nadále zabývat (více v knize [1], str.156-169), sloužily k ilustraci problému se složeným jmenovatelem při testování hypotéz. Pokud teď chceme testovat hypotézu, zda $\sigma^2_b = 0$, nenajdeme žádnou střední hodnotu středního čtverce, která

má vynechaný člen pouze σ_b^2.

$$E(M_1) - k_1\sigma_b^2 = \sigma_{c:b}^2 + k_3\sigma_{ab}^2 + k_4\sigma_{ac:b}^2 + \sigma_e^2 \ .$$

Pravou stranu můžeme přepsat do tvaru $E(M_2)+E(M_3)-E(M_4)$. Záporné členy převedeme na druhou stranu rovnice

$$E(M_1) + E(M_4) = k_1\sigma_b^2 + E(M_2) + E(M_3) \ .$$

Další výpočty provedeme v obecném případě.

$$E(M_m) + ... + E(M_n) = k\sigma_\alpha^2 + E(M_r) + ... + E(M_s)$$

a uvažujme hypotézu $H_0 : \quad \sigma_\alpha^2 = 0$, kde σ_α^2 je jakýkoli faktor v našem modelu. Testová statistika, kterou navrhl Satterthwaite v roce 1946 má tvar

$$F = \frac{M_H}{M_D} = \frac{M_m + ... + M_n}{M_r + ... + M_s} \quad , \qquad \text{kde F } \sim F(p, q) \ .$$

Pro odhad stupňů volnosti použijeme Satterthwaitovu aproximaci.

$$p = \frac{M_H^2}{M_m^2/f_m + ... + M_n^2/f_n} \text{ a obdobně i } \quad q = \frac{M_D^2}{M_r^2/f_r + ... + M_s^2/f_s}$$

f_i jsou stupně volnosti příslušných středních čtverců M_i. Je zřejmé, že tato aproximace p a q nemusí být nutně celočíselná. Jádrem tohoto testu je, že jak čitatel, tak i jmenovatel mají přibližně χ^2 rozdělení s p (resp. q) stupni volnosti. Více o aproximaci a jejím odvozením například v článku [4].

5.4 Příklad

Pro srovnání použijeme stejná data jako při dvojném třídění. Budeme uvažovat ID školy jako náhodný efekt. Našim hlavním cílem bude porovnat tyto výsledky s příkladem u dvojného třídění, kde naopak máme školy jako pevný efekt. Načteme data [3] a do příkazového

[3]Naše data jsou nevyvážená, avšak v naší práci jsme odvozovali lineární modely jen pro vyvážená data. Principiálně se moc od sebe neliší, ale zápis s nevyváženými daty je mnohem obecnější a zdlouhavý. O nezbytných základech jsme pojednali na konci 2. kapitoly

řádku napíšeme: UNIANOVA matematika BY pohlavi škola. Připíšeme /RANDOM=škola, aby ID školy vystupovalo jako náhodný efekt. Nemusíme spouštět test na nulovost střední hodnoty (připsáním /INTERCEPT=EXCLUDE), protože víme, že by byl zamítnut. Dostaneme

Obrázek 4: Testy hypotéz smíšeného modelu

Tests of Between-Subjects Effects

Dependent Variable:Známka - matematika

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
pohlaví	Hypothesis	1,825	1	1,825	2,352	,146
	Error	11,596	14,946	,776[a]		
škola	Hypothesis	24,118	9	2,680	3,703	,032
	Error	6,513	9	,724[b]		
pohlaví * škola	Hypothesis	6,513	9	,724	,702	,707
	Error	210,430	204	1,032[c]		

a. ,830 MS(pohlaví * škola) + ,170 MS(Error)
b. MS(pohlaví * škola)
c. MS(Error)

Sloupek Type III Sum of Squares obsahuje číselné hodnoty součtů čtverců vypočtené přes Typ III. Sloupek df jsou stupně volnosti jednotlivých součtů čtverců. Mean Square jsou střední čtverce, neboli SS/df. Sloupek F obsahuje hodnoty našich testových statistik. U smíšených modelů nemají všechny testové statistiky ve jmenovateli reziduální součet čtverců jako ve dvojném třídění, proto u každého testu máme přidaný řádek error. V něm se nachazí součet čtverců, který odpovídá jmenovateli naší testové statistiky. V poznámkách pod tabulkou můžeme vyčíst, jaké střední čtverce to jsou. Poslední sloupek nám udává p-hodnotu F–testu Pokusme se interpretovat výsledky z tabulky.

- Začneme odspoda, interakční řádek pohlaví:škola má p-hodnotu 0,707, proto zde nezamítáme hypotézu o nulovosti σ^2_γ. To znamená, že můžeme tuto interakci zanedbat a přejít k modelu bez γ_{ij}.

- Zvolme si hladinu testu $\alpha = 0,05$. Naše p-hodnota na řádku škola je menší než α, proto tvrdíme, že jednotlivé školy mají různý vliv na známku z matematiky.

- Co se týče pohlaví, tam je p-hodnota kolem 0,146, takže tentokráte hypotézu nezamítáme. Tvrdíme, že na známku z matematiky nemá vliv, zda jste chlapec nebo dívka.

Na odhad počtů stupní volnosti u chyby testu hypotézy na pohlaví byla použita výše zmíněná Satterthwaitova aproximace. Provedeme kontrolu

$$d.f. \doteq \frac{(0,83 * 0,724 + 0,17 * 1,032)^2}{(0,83 * 0,724)^2/9 + (0,17 * 1,032)^2/204} = \frac{0,602}{0,0402} \doteq 14,9751$$

Při porovnání s výsledky z dvojného třídění vidíme, že interakce je v obou případech výpočtově úplně stejná. Avšak je zde rozdíl v inerpretaci. Ve dvojném třídění jí máme jako pevný efekt, tudíž tam jsme testovali nulovost střední hodnoty parametrů γ_{ij}. Zatímco u smíšeného modelu máme interakci jako náhodný efekt, tudíž jsme testovali nulovost rozptylu náhodného efektu γ_{ij}. Obdobný je i rozdíl v interpretaci hypotézy u jednotlivých škol, kde navíc máme rozdíl ve výpočtech. Sice jsme v obou případech hypotézy zamítli, avšak smíšený model měl vyšší p-hodnotu.

Závěr

Cílem této práce bylo přiblížit čtenáři metody analýzy rozptylu a jejich zobecnění na lineární smíšené modely. Jednotlivé kapitoly jsou vedeny od nastínění problému, přes definice důležitých pojmů a odvození jednotlivých metod až do odvození základních testů. Pro názornost jsou přidány příklady. Na všechny metody byla použita stejná data z projektu PISA, aby názorně ukázala rozdílnosti mezi jednotlivými metodami.

Reference

[1] *S. R. Searle:* **Linear Models**
John Wiley & Sons, 1971

[2] *J. Anděl:* **Základy Matematické Statistiky**
MATFYZPRESS, 2007

[3] *M. Meloun, J. Militký:* **Kompendium statistického zpracování dat**
Academia, 2002

[4] *Univeristy of Florida:*
Satterthwaite's Approximation for Degrees of Freedom
dostupné na [www.stat.ufl.edu/∼winner/cases/satter.ppt]

[5] *A. Khuri:* **Linear Model Methodology**
CRC Press, 2010