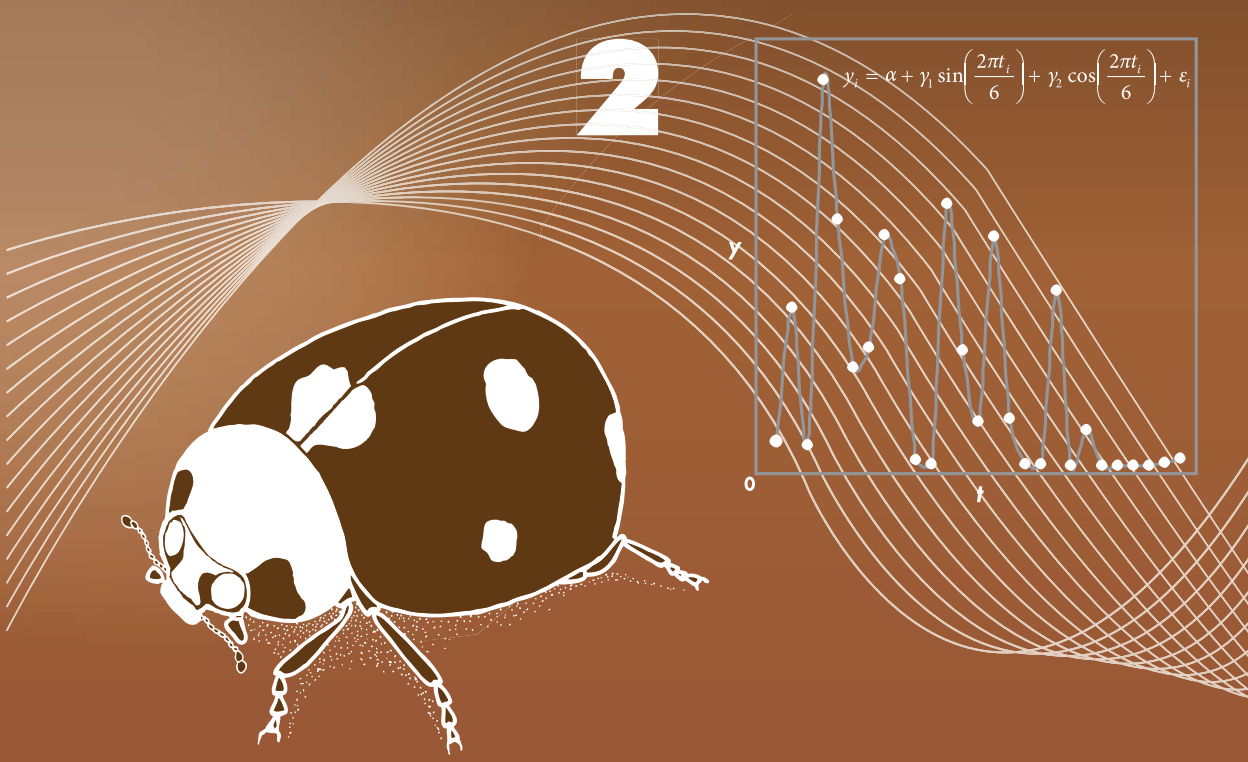
MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

LINEÁRNÍ MODELY S KORELACEMI V PROSTŘEDÍ R

2



STANO PEKÁR
MAREK BRABEC

muni PRESS

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT
LINEÁRNÍ MODELY S KORELACEMI V PROSTŘEDÍ R
2. díl

STANO PEKÁR, MAREK BRABEC

muni
PRESS

Recenzoval: Prof. V. Jarošík

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

LINEÁRNÍ MODELY S KORELACEMI V PROSTŘEDÍ R

2. díl

STANO PEKÁR
MAREK BRABEC

Masarykova univerzita, Brno 2012

http://www.muni.cz/press/books/pekar

Pekár S. & Brabec M. 2012. Modern Analysis of Biological Data. 2.
Linear Models with Correlations in R. Masaryk University Press, Brno.

Realizace knihy byla podpořena výzkumným záměrem MŠMT (MSM 0021622416).

© 2012 Stano Pekár, Marek Brabec
Illustrations © 2012 Stano Pekár
Design © 2012 Ivo Pecl, Stano Pekár, Grafique
© 2012 Masarykova univerzita

ISBN 978-80-210-7719-5 (online : pdf)

ISBN 978-80-210-5812-5 (brožovaná vazba)

OBSAH

PŘEDMLUVA

Po třech letech přicházíme s dalším dílem Moderní analýzy biologických dat. Tento díl představuje modely a metody, které jsou dalším rozšířením obecného lineárního modelu a které umožňují efektivně modelovat korelovaná (statisticky závislá) data. Dle našich zkušeností jsou korelovaná data v biologických, medicínských, ale i jiných oborech velmi častá. Proto je velká i motivace pro použití výše zmíněných modelů a na nich založených metod statistické analýzy. Někdy bohužel až přespříliš. Jejich korektní použití vyžaduje přece jenom určité úsilí a alespoň základní přehled o tom, jak fungují, jaké jsou jejich předpoklady, co přesně říkají a na co odpověď naopak nedávají. Pokoušet se pracovat „naslepo", bez takovýchto znalostí, tedy jen metodou „pokusů a omylů" je samozřejmě zcela nevhodné. Zejména proto, že prostor pro chyby v analýze i nekorektní prezentaci výsledků je zde velký – podstatně větší než u metod jednodušších. Touto knihou vám chceme pomoci alespoň při základní orientaci. Její četba by měla být jakýmsi odrazovým můstkem k vlastním analýzám experimentálních či observačních dat, ale také ke kritickým úvahám o jejich výsledcích. Měla by také pomoci rozpoznat limity standardního či „učebnicového" přístupu ke komplikovaným a/nebo nestandardním problémům a motivovat k hlubšímu studiu specializované literatury a/nebo vyhledání pomoci u profesionálního statistika.

Podobně jako první díl, je i tento zaměřen na širokou biologickou problematiku, a to z oblastí ekologie, zoologie, botaniky, genetiky, zemědělství i medicíny. Používáme stejnou koncepci jako v prvním dílu, postavenou na řešení konkrétních příkladů od specifikace problému až po jeho závěr. Celkově je kniha zaměřena na praktickou analýzu reálných dat, a proto je v ní jen skutečně nezbytné množství teorie a maximum praktických příkladů. Přesto je zde teorie o něco více než v prvním dílu – vyžaduje to náročnější látka, o které zde pojednáváme.

K pochopení modelů z tohoto dílu je nezbytné znát leccos z dílu prvního. Na mnohých místech na první díl dokonce přímo odkazujeme. Proto radíme těm z vás, kteří mají s (zobecněnými) lineárními modely zkušenosti malé či žádné, aby si nejprve prostudovali díl první. A to i v případě, že k analýze chtějí použít pouze metody z tohoto dílu. Číst a pracovat takříkajíc od konce se v tomto případě určitě nevyplatí!

Protože se programové prostředí R stává velice populární v mnoha aplikovaných oborech (zejména např. v biologii, medicíně, chemii, environmentalistice apod.), postavili jsme, stejně jako v prvním dílu, řešení příkladů opět na něm. To však neznamená, že by tato kniha byla jen jakýmsi manuálem tohoto softwaru. Právě naopak, snažili jsme se rozebrat příklady analýz takovým způsobem, aby mohly být snadno provedeny i v jiných statistických softwarových balících. Snažíme se klást důraz na formulaci a pochopení modelu samotného, tedy

na to, aby čtenář pochopil jeho praktické vlastnosti, jakož i věcnou (např. biologickou) interpretaci jeho jednotlivých složek. Právě proto dbáme na formalizovaný zápis modelu. Tento přístup čtenáře nutí ujasnit si model až do detailů tak, aby ho byl schopen popsat i ostatním (např. v publikaci). Praktickou výhodou je pak i to, že převedení formalizovaného zápisu do syntaxe konkrétního softwaru je pak už jen otázkou poměrně jednoduchého „překladu".

V naší pedagogické praxi jsme zaznamenali, že někteří studenti, kteří se podle prvního dílu připravovali na zkoušku, měli problémy s interpretací výsledného modelu. A to jak s jeho matematickým zápisem, tak s jeho grafickou prezentací. Proto v tomto dílu klademe ještě větší důraz na pochopení struktury probíraných modelů a na interpretaci jejich výsledků.

Podobně jako v prvním dílu, jsou i zde použita reálná data (pocházející z různých, zejména biologických projektů), upravená tak, aby lépe vyhovovala v pedagogickém smyslu. Při výběru dat jsme tedy zohledňovali zejména následující kritéria: zajímavá problematika, modelování různých problémů věcných i formálních, ilustrace úskalí při specifikaci i odhadu různých modelů. Všechny použité příklady jsou relativně jednoduché a velikost dat je nevelká zvláště proto, aby se v nich čtenář mohl snadno orientovat (i tak v některých příkladech abstrahujeme od některých detailů, jež by bylo možné/vhodné studovat pokročilejšími metodami, než jsou ty probírané v této a předchozí knize – vaše názory na to, které by to mohly být, rádi uvítáme na našich emailových adresách: pekar@sci.muni.cz či mbrabec@cs.cas.cz).

Na závěr bychom chtěli poděkovat kolegům, kteří nám pro účely ilustrací v tomto dílu poskytli svoje data. Jsou to T. Bilde, A. Honěk, I. Jankovská, S. Koprdová, M. Omesová a P. Šmarda. Také ale studentům kurzů, které se konaly před napsáním této knihy, za cennou zpětnou vazbu. Zejména připomínky a poznatky o tom, co se studuje snadněji, co méně snadno, kde je vhodné výklad zrychlit, kde zpomalit apod. Ty nám pomohly modifikovat naše původní nápady tak, abychom lépe vyšli vstříc při vysvětlování prezentované látky. A konečně pak doc. RNDr. D. Fryntovi, CSc., a prof. RNDr. V. Jarošíkovi, CSc., za cenné připomínky, které naše dílo dozajista vylepšily.

Prosinec 2011

Stano Pekár
Marek Brabec

1 ÚVOD

Již nejednomu vědci (studentovi) se stalo, že mu oponent zpochybnil analýzu dat kvůli závislosti pozorování v jeho studii (diplomové práci). Klidně přitom mohl použít nehezkého slova „pseudoreplikace" nebo nějakého podobného pejorativa. V jakém smyslu se slovo pseudoreplikace v biologických i jiných přírodních vědách používá, si podrobněji objasníme v kapitole 2.1. Nyní je podstatné to, že takové zpochybnění není zanedbatelné a může klidně vést k zamítnutí celé práce. Podobnému problému se samozřejmě lze vyhnout – a to se znalostí statistických metod, které s korelovanými daty umí korektně pracovat. Některé z takových metod probereme v této knize.

Budeme se zde zabývat daty, která nejsou nezávislá ve statistickém či pravděpodobnostním smyslu. Tedy takovými, která vykazují nějakou formu **statistické závislosti**, např. korelace mezi opakovanými pozorováními. Velmi často se setkáváme s tím, že analyzovaná data mají v jistém smyslu jemnější strukturu, než jakou předpokládají základní učebnicové regresní modely (např. ty, kterým jsme se věnovali v prvním dílu, Pekár & Brabec 2009). Třeba proto, že data byla pořízena na různých subjektech s tím, že každý subjekt byl měřen opakovaně. Tuto formulaci je zapotřebí vnímat abstraktně, protože „subjektem" může být leccos: člověk, zvíře nebo kus materiálu, který byl náhodně odebrán ze vzorkovaného prostoru. Měření pak může být opakováno v čase a/nebo v prostoru. Zatímco bývá přirozené a rozumné pozorování na různých subjektech považovat za vzájemně nezávislá, pozorování uvnitř téhož subjektu mají často tendenci být si podobnější. Formalizací tohoto intuitivního a praktického pohledu může být předpoklad o statistické nezávislosti mezi subjekty a připuštění jistého stupně statistické závislosti mezi opakovanými pozorováními téhož subjektu. Tedy model s komplikovanější strukturou oproti standardním (zobecněným) lineárním regresním modelům, kde jsme předpokládali kompletní nezávislost mezi všemi pozorováními.

Velmi často (v této knize téměř výlučně) máme na mysli speciální případ (lineární) statistické závislosti měřený **korelací**. Obecně je pojem statistické závislosti však mnohem širší a komplikovanější. V případě vícerozměrných normálních dat však oba pojmy splývají – tak tomu bude ve většině textu tohoto dílu. Přítomnost korelace či obecně závislosti v datech představuje komplikaci. Připomeňte si, že v prvním dílu (Pekár & Brabec 2009) jsme řešili pouze příklady, ve kterých byla všechna pozorování na sobě vzájemně nezávislá. Použití takovýchto modelů na korelovaná data je však nerealistické. Inference na nich založené jsou pak bohužel často zcela nesprávné. Nejde o drobnosti. Chyby, které přitom vznikají, mohou být zcela fundamentální povahy, například falešné významnosti efektů, jež by při korektním ošetření závislostí významné vůbec nebyly.

Poznamenejme, že terminologie označující opakovaná pozorování je velmi pestrá. V různých oborech je zvykem používat mírně odlišné názvy s obdobným významem. Tak se v anglické terminologii setkáme s pojmem „clustered data", „panel data", „longitudinal data" či „repeated measures", pokud jde o opakování v čase. Obecně můžeme (a v této knize i budeme) kromě subjektů mluvit ve stejném významu také o navzájem nekorelovaných skupinách (shlucích, angl. clusters) pozorování a korelovaných pozorováních uvnitř skupiny. To je totiž názvosloví blízké klíčovým slovům používaným v balíčcích, které k řešení příkladů a analýzám dat budeme používat.

Jak jsme již naznačili výše, pozorovanými subjekty mohou být živé organismy (jedinci) i neživé jednotky (misky, boxy nebo různá území apod.). Vzájemné závislosti mezi pozorováními uvnitř stejné plochy či subjektu mohou vznikat proto, že pozorování jsou lokalizována prostorově „blízko". Například na téže rostlině, ve stejném hnízdě, v tomtéž klasu či na jiném vhodném místě výskytu. Vezměme si případ zrna a klasu. Subjektem či shlukem by mohl být klas, opakováním uvnitř subjektu pak jednotlivé zrno. V důsledku vzájemné blízkosti opakovaných pozorování uvnitř subjektu (a velice podobných podmínek prostředí) lze očekávat, že (sub)jednotky uvnitř shluku jsou spolu nějak provázány – vzájemně se ovlivňují nebo jsou si prostě jenom „nějak" podobnější. Často (ne však vždy) je pak nárůst závislosti vzhledem k vzdálenosti monotónní: čím blíže k sobě subjednotky jsou, tím je jejich podobnost větší a naopak čím dále od sebe jsou, tím jsou jejich pozorování méně závislá.

Uvědomme si také velký význam měřítka experimentu, ze kterého data pocházejí. Jestliže může být vcelku realistické považovat různé klasy na poli za nekorelované (a tedy za „subjekty") v mikrostudii, která sleduje např. velikost jednotlivých zrn, tak považovat různé klasy rostoucí vedle sebe za nezávislé nelze. Nelze tedy postupovat mechanicky a zapamatovat si poučky typu: klas = subjekt! Záleží na okolnostech a cílech studie, ze které data pocházejí. O nich bychom měli vždy podrobně a kriticky uvažovat. Mějme na paměti, že volba toho, co se považuje za **subjekt** (a tedy co je mezi sebou nezávislé), není automatická a je již prvním (a velmi významným) krokem při specifikaci modelu pro naše další analýzy.

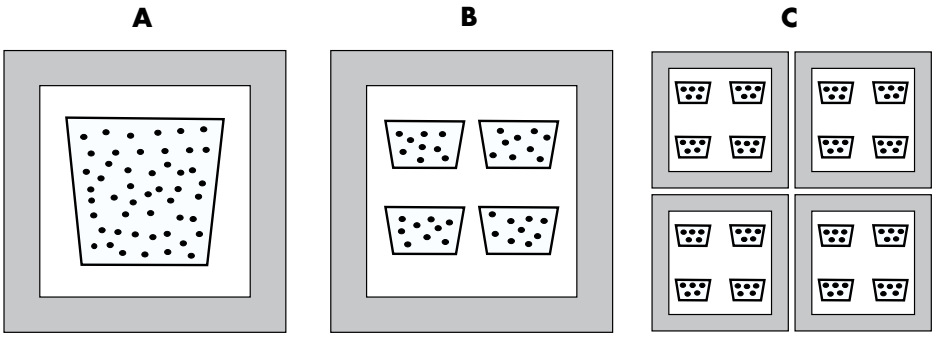
Data se závislostmi vznikají často neúmyslně v důsledku špatného naplánování experimentu či observační studie. To je v praxi sice celkem častá situace, zde ji ale dále komentovat nebudeme. Nebo vznikají v důsledku omezení prostorových, logistických a jiných, kdy se studie jiným způsobem provést prostě nedá. Přítomná závislost (korelace) je jen rušivým parametrem, který náš statistický model zavádí, a je nutné ji odhadovat jen proto, že korektním způsobem zohledňuje množství informací v datech obsažené. Korelovaná data mohou ale také vznikat zcela úmyslně, s cílem odhadnout různé charakteristiky plynoucí se závislostí mezi pozorováními. Závislosti zde nejsou jen rušivými parametry, ale jsou přímo předmětem studia. Například proto, že mají důležitou biologickou interpretaci (kladné korelace mohou souviset s agregací, s různými mechanismy disperze apod.). Dalším (úmyslným) důvodem přítomnosti korelací v datech pak může být záměrné speciální uspořádání experimentálního schématu s cílem zefektivnit odhady či srovnání sledovaných parametrů.

1.1 Konkrétní situace

Uveďme si teď konkrétní příklad dat s korelacemi mezi opakovanými pozorováními uvnitř subjektu a nezávislostí mezi subjekty. Představte si, že k uskutečnění jednoduchého experimentu odeberete semena daného druhu rostliny rostoucí na jednom poli, vysejete je do jednoho květináče, ten umístíte do jednoho klimaboxu (obr. 1-1A) a sledujete jejich růst. V době květu pak odhadnete průměr (třeba výšky rostlin) a k tomu hodnotu směrodatné odchylky. Tyto dva parametry vám řeknou, jaký je průměrný vzrůst rostlin a jaká je variabilita růstu v populaci, ze které jste odebírali – tedy z toho konkrétního pole, z něhož semena pocházejí.

A teď si představte, že stejné množství odebraných semen vysejete nikoli do jednoho, ale hned do několika květináčů, které umístíte do jednoho klimaboxu (obr. 1-1B). Třeba proto, že nemáte tak veliký květináč, aby se tam všechna semena pohodlně vešla a mohla bezproblémově růst. Budete se sice snažit starat se o každý květináč stejně – ale i tak se prostě stane, že některý zalijete o trochu více, jiný o trochu méně apod. Důsledkem pak bude, že vzrůst rostlin bude v jednom květináči větší než v jiném. A navíc v důsledku náhodných rozdílů podmínek (způsobených kromě naší péče také drobnými odchylkami ve složení půdy, v množství světla pro jednotlivé květináče atd.) bude vzrůst rostlin ze stejného květináče podobnější než vzrůst rostlin z květináčů různých. Velikost rostlin pocházejících ze stejného květináče tudíž nebude nezávislá, ale závislá (korelovaná). Oproti tomu, velikost rostlin pocházejících z různých květináčů bude v podstatě nezávislá. Rozdělením semen do různých květináčů je do naší studie vnesen zdroj heterogenity mezi květináči. Má pak smysl mluvit o dvou složkách celkové variability výšky rostlin: o variabilitě *mezi* květináči (mezi průměry rostlin pocházejících z jednoho květináče) a variabilitě mezi jednotlivými rostlinami *uvnitř* květináčů.

A nyní si představte mnohem složitější situaci, kdy semena vysejete do několika květináčů, jež pak umístíte do několika klimaboxů. Například proto, že se všechny do jednoho klimaboxu prostě nevejdou (obr. 1-1C). Klimaboxy by sice měly fungovat stejně (všechny



Obr. 1-1 Schematické znázornění umístění semen v květináčích a v klimaboxech (šedý rám). **A**. Všechna semena jsou umístěna v jednom květináči a jednom klimaboxu. **B**. Semena umístěna ve čtyřech květináčích v jednom klimaboxu. **C**. Semena umístěna v šestnácti květináčích ve čtyřech klimaboxech.

mají certifikát od stejného renomovaného výrobce), ale v některém může být třeba v důsledku náhodných poruch regulace (a v důsledku mnoha dalších drobných vlivů, o jejichž povaze nemusíme nic ani tušit) o něco vyšší teplota než v ostatních, takže růst rostlin bude v tomto termostatu lepší. Výška rostlin v květináčích uvnitř jednoho klimaboxu si bude podobnější než v květináčích umístěných v různých klimaboxech. Správně tušíte, že rozdělením květináčů do klimaboxů vnášíte do příkladu další druh heterogenity. Máme zde náhodné odchylky mezi klimaboxy (mezi průměry rostlin pocházejících z jednoho klimaboxu) a tím i závislost mezi květináči uvnitř klimaboxu (mezi průměry jednotlivých květináčů umístěných ve stejném klimaboxu). Celkovou variabilitu výšky rostlin pak můžeme (hierarchicky) rozložit do tří složek:

- *mezi* klimaboxy
- *mezi* květináči *uvnitř* daného klimaboxu
- *mezi* rostlinami *uvnitř* daného květináče a klimaboxu

Nebo si představte, že v experimentu pokračujete a opakovaně měříte výšku každé rostliny v nějakém intervalu, dejme tomu jednou týdně po dobu několika měsíců. Z naměřených hodnot výšek pak odhadnete nějaký model růstu. Pro snadnější pochopení si můžeme představit, že růst rostlin je přibližně lineární, a tedy charakterizován dvěma parametry: absolutním členem a směrnicí. Různá individua budou vykazovat odlišné hodnoty těchto růstových parametrů. Budeme-li interindividuální odchylky parametrů považovat za náhodné – tedy budeme-li pracovat s náhodnými absolutními členy a směrnicemi – povede to k modelu, který má korelace mezi opakovanými měřeními téhož individua (a nekorelovaná měření provedená na individuích různých). Celkovou variabilitu zde můžeme oproti předchozí situaci s květináči rozkládat ještě podrobněji (byť už ne hierarchicky) na složky:

- variabilita *mezi* absolutními členy jednotlivých rostlin
- variabilita *mezi* směrnicemi jednotlivých rostlin
- variabilita reziduí po lineárním trendu růstu téže rostliny

K analýze dat se závislostmi se dnes hodně (byť určitě ne výlučně) používá model se smíšenými efekty. Jeho použitím můžeme z dat odhadnout jednak složky variability (tzv. komponenty rozptylu), o nichž jsme se právě zmínili, i korelace mezi jednotlivými pozorováními téhož subjektu. Relativní velikost jednotlivých komponent rozptylu nám prozradí, na které hierarchické úrovni je variabilita největší. To může být zajímavé samo o sobě, ale též z různých praktických důvodů. Pokud např. zjistíme velkou variabilitu mezi klimaboxy, může být užitečné jednotlivé klimaboxy prověřit a identifikovat ty, které nejvíce vybočují z průměru (a případně je nechat opravit). Podobně, pokud bychom zjistili velkou variabilitu mezi květináči, mohli bychom před příštím experimentem lépe instruovat technika, aby si dával větší pozor na dodržování stejné zálivky. Konečně, velká variabilita mezi semeny uvnitř květináčů by mohla naznačovat nestálost tohoto znaku v populaci rostlin, ze kterých jsme semena na poli odebrali, a tudíž nevhodnost této odrůdy v zemědělské výrobě.

Zjištěné korelace jsou také velmi užitečné při zakládání pokusů tzv. „poučeným způsobem". Tedy pokusů s co největší efektivitou, šetřících zdroje experimentálního materiálu

apod. O plánování experimentů existuje nemalé množství velmi užitečné statistické teorie, je ale jasné, že výhodné bude investovat do velkého počtu opakování na úrovních hierarchie, jejichž komponenty rozptylu jsou velké.

Model se smíšenými efekty je také vhodný, pokud nás odhady variability a korelací nezajímají, ale chceme získat vyhlazený (tj. výpočetním způsobem šumu zbavený) odhad jednotlivých pozorování (užitečný např. při odhadu plemenné hodnoty v zemědělství). Nebo pokud chceme něco vědět o středních hodnotách sledované charakteristiky a jejich změnách s různými vysvětlujícími proměnnými. V takovém případě je potřeba zohlednit korelaci kvůli korektnímu výpočtu testů hypotéz o odlišnostech středních hodnot (či o efektech různých experimentálních ošetření). Pokud bychom závislost mezi opakovanými měřeními nevzali v úvahu a sestavili model jako jeden z „obyčejných" regresních modelů předpokládajících nezávislost, odhady parametrů by typicky nebyly až tak moc nesprávné. Mnohem horší to však bývá s odhady středních chyb (SEM) koeficientů. Ty bychom zpravidla hrubě podcenili (méně často přecenili), což by v konečném důsledku vedlo k nižším (resp. vyšším) p-hodnotám, než jaké bychom dostat měli. Korektní modely pro korelovaná data obecně jsou tedy (mimo jiné) *nástrojem korekce naivních p-hodnot* a z nich plynoucích nekorektních závěrů – falešně pozitivních výsledků testů, tedy nálezu signifikantních výsledků tam, kde by být neměly!

Odhady rozptylů a korelací jsou parametry, které dostaneme z analýzy jaksi navíc. Práce s nimi je obtížnější. Jedním z důsledků je i to, že k jejich spolehlivému odhadu potřebujeme mít více měření než pro modely bez korelací. To „více" není nějaké obecně platné číslo (protože jeho hodnota závisí na mnoha okolnostech spojených s vlastnostmi konkrétního použitého modelu i modelovaných dat, jako například stupněm korelací, i cílem studie). Z hlubších statistických i věcných důvodů bývá většinou cennější mít v experimentu/studii spíše více subjektů než mnoho opakování několika málo subjektů. Někdy se uvádí, že počet subjektů/bloků by měl být alespoň 30 (Hardin & Hilbe 2003). Doopravdy by to ale mělo být *o dost* více. Uvědomme si, že odhad směrodatné odchylky je náročnější oproti odhadu průměru i v tom nejjednodušším (jednovýběrovém, jednorozměrném) případě (χ^2 – rozdělení výběrového rozptylu je mnohem „méně pohodlné" než rozdělení výběrového průměru). A zde máme situaci mnohem komplikovanější. Pro malá nebo jen „nedobře poskládaná" (např. silně nevyvážená) data se pak může lehce stát, že komponenty rozptylu nebude možné odhadnout vůbec. Nebo, a to je ve svém důsledku ještě horší, sice odhadnout „nějak" půjdou, ale odhady budou velmi špatně podmíněné (nestabilní). Takový problém může v průběhu výpočtu automaticky detekovat k odhadu použitá optimalizační procedura a zahlásit problémy s konvergencí. Bohužel však automatická procedura problém rozpoznat také nemusí – v horším případě prostě „jen" poskytne špatné/nesmyslné výsledky. Odpovědnost za formulaci modelu vhodného pro analyzovaná data spočívá vždy na nás (nikoli na softwaru).

1.2 Jak číst tuto knihu

Kniha je uspořádána tak, že v její první části (kap. 1–8) převažuje spíše teorie (skutečně jde o naprosté minimum teorie, která je nezbytná k pochopení principů probíraných me-

tod potřebných ke správné interpretaci výsledků i korektní strategii analýzy) a druhá část (kap. 9–11) se zabývá praktickými ukázkami a zaměřuje se na řešení konkrétních příkladů. Doporučujeme číst knihu popořádku. Čtenářovi bez znalosti matematiky doporučujeme číst nejprve kapitoly 1–3 a 9–11. V nich najde odkazy na teorii z kapitol 4–8. Poté, co se při řešení příkladů seznámí s praktickým významem těchto teoretických kapitol, patrně si je rád sám prostuduje a pokusí se je pochopit.

V první kapitole je velmi důležitá hned následující část, ve které budeme mluvit o pomocné proměnné, se kterou jsme se v prvním dílu nesetkali. Ta nám pomůže definovat (navzájem nekorelované) skupiny korelovaných pozorování. Tato proměnná je zcela zásadní pro všechny analýzy diskutované v této knize.

Ve druhé kapitole zmiňujeme několik různých způsobů vzniku a sběru korelovaných/závislých dat. Cílem zde není podat rozsáhlý přehled existujících experimentálních designů, uvedeme jenom několik z nich (resp. ty, jež s korelacemi nějak souvisí). Dále pak probereme některá základní pravidla a definice různých pojmů běžných v oblasti navrhování experimentů. Mimo jiné se dozvíte, co to jsou *pseudoreplikace* a jaké jsou výhody designů postavených na závislých pozorováních.

Třetí kapitola obsahuje motivační příklad. Jde o příklad velmi jednoduchý – ukázkový, který představuje základní přístup a filozofii modelů se smíšenými efekty (modelů s náhodnými parametry). Jeho cílem je ukázat, že metody analýzy korelovaných dat nejsou důležité proto, že jsou to metody „moderní" a často citované, ale proto, že poskytují korektní řešení reálných problémů.

V kapitole 4 si zopakujeme pár elementárních definic a faktů z lineární algebry, hlavně ty, které se týkají práce s vektory a maticemi. A to proto, že při definici modelů, hlavně jejich korelačních a kovariančních struktur, budeme matice používat. Maticový zápis je nyní zcela standardní v mnoha učebnicích zejména proto, že umožňuje kompaktní a přehledné vyjádření toho, co budeme později v souvislosti se smíšenými modely probírat. Pokud tuto látku máte v živé paměti, můžete čtvrtou kapitolu přeskočit.

V kap. 5 definujeme základní typy modelů, se kterými budeme dále pracovat. Přesněji řečeno, budeme mluvit o jedné jejich skupině, která je založena na vícerozměrném normálním rozdělení. Jeden z (jednodušších a pragmatičtějších) přístupů použitelných pro širší třídu rozdělení pak probereme v kap. 8.

Kap. 6 je nejrozsáhlejší „teoretickou" kapitolou. Podrobně popisuje různé varianční a korelační struktury, které se nám při praktickém modelování budou hodit jako základní stavební kameny. Její zvládnutí a pochopení je důležité. Před každým modelováním se totiž musíme rozhodnout, jak budeme vlastnosti náhodných chyb v našem modelu popisovat, a tudíž kterou korelační a/nebo varianční strukturu použijeme. Je potřeba se naučit znát jejich vlastnosti a pochopit důsledky, které pro model mají. Kvalita výsledného modelu bude dána i správnou volbou kovarianční (korelační a varianční) struktury. Jak uvidíte, je jich k dispozici docela dost.

V kap. 7 jsou pak podrobně rozebrány různé kovarianční struktury náhodných efektů. Ty jsou (samozřejmě) nedílnou součástí specifikace modelů s náhodnými efekty. Také zde je potřeba se naučit znát jejich vlastnosti.

Kapitoly 9 až 11 poskytují podrobná řešení několika konkrétních příkladů. Jsou rozděleny dle typu modelu vzhledem k přítomnosti náhodných proměnných a typu závisle proměnné. Jednotlivé příklady jsou prezentovány od popisu problému přes zdůvodnění jednotlivých částí modelu až k formulaci závěru o zkoumaném problému. Různé příklady se snaží ukázat vždy něco nového (jiný typ modelu, jinou korelační nebo kovarianční strukturu apod.). Postupy, které by se opakovaly (takových je docela hodně), jsme ve výkladu vynechali. Vy je naopak při praktických analýzách s výhodou použijete.

Na konci jsou zařazeny dva rejstříky: obecných termínů, funkcí a argumentů jazyka R. Indexy jsou podrobné, aby umožnily efektivní dohledání požadovaných informací. Třeba příkladů podle použitých organismů.

Datové soubory a syntaxe příkazů použitých v této knize si můžete stáhnout z adresy http://www.muni.cz/press/books/pekar.

1.3 Definice proměnných

V modelech, kterým se v této knize budeme věnovat (a zejména pak při jejich softwarové formulaci), vystupuje oproti modelům z prvního dílu navíc proměnná se zcela specifickou funkcí. Budeme jí říkat **skupinová proměnná** (grouping variable). Jde o proměnnou zadávanou jako speciální argument funkcí pro práci s modely pro korelovaná data. Tato proměnná definuje subjekty, shluky, panely či prostě skupiny závislých (korelovaných) pozorování. Skupinová proměnná (mimo jiné) určuje separaci variability do složek „mezi" a „uvnitř", o nichž jsme mluvili již dříve. Je to proměnná klasifikační (nominální) povahy, protože určuje zařazení pozorování do nezávislých skupin. Tak je automaticky interpretována uvnitř funkcí pro modelování korelovaných dat, ať už je původně zadána jako proměnná znaková nebo numerická. Obsahuje prostě jedinečný kód pro každý subjekt, jedince, květináč, klimabox, blok pozorování, experimentální či observační jednotku apod. Přestože skupinové proměnné jsou zpravidla numerické, budeme je značit velkými písmeny (stejně jako faktory). Tyto proměnné jsou velice důležité. Bez nich nelze provést žádnou z analýz probíraných v této knize. Obecně je jasné, že při sběru dat je potřeba opravdu pečlivě zaznamenat podstatné části struktury, kterou data mají. V našem kontextu to bude mimo jiné znamenat, že budeme muset registrovat a uchovat příslušnost pozorování (např. semen ve výše zmíněném příkladu) k jednotlivým subjektům či prostě experimentálním jednotkám (květináčům a klimaboxům). To proto, aby se nám na konci studie nestalo, že si nebudeme moci vzpomenout, „kam které pozorování patřilo". Pokud jsme si v průběhu studie hodnoty skupinové proměnné nezaznamenali (i to se bohužel stává), nelze pak (korelovaná) data správně analyzovat. A to by byla (nenapravitelná) škoda!

Kromě skupinové proměnné budeme v modelech používat všechny ostatní typy proměnných, kategorické a kontinuální (to jest spojité numerické), podobně jako v prvním dílu. Také zde budeme mluvit o modelech regresního typu, tedy takových, které popisují vztah závislé proměnné k jedné či vícero proměnným vysvětlujícím. Ve většině textu této knihy půjde o závislou proměnnou spojitou, normálně rozdělenou. Závislá proměnná bude vždy jen jedna. Formálně tedy bude model postaven velmi podobně jako **model jednorozměrný** (univariate). Ve skutečnosti však budeme mít co do činění se specifickou třídou **modelů vícerozměrných**. Kvůli korelaci mezi opakovanými pozorováními téhož subjektu jsme totiž nuceni modelovat vícerozměrná data (zajímat se i o souvztažnosti mezi několika pozorováními, nikoli jen o marginální rozdělení pozorování jednotlivého). A to pomocí relativně jednoduchých nástrojů. Takové modely mají oproti klasickým vícerozměrným modelům řadu výhod. Umožňují elegantní a pohodlnou práci s chybějícími daty (pro jednodušší typy mechanismu, kterým chybějící data vznikají). Jde o modely strukturované a díky tomu poměrně parsimonní (s relativně malým počtem parametrů). Umožňují snadno testovat různé hypotézy včetně těch, jež jsou motivovány nějakou (např. biologickou) teorií. Jsou tedy velmi vhodné zejména v situaci, kdy je o studovaném problému již „něco známo", a to něco chceme buď testovat, zpřesnit, nebo použít k efektivnější, ale korektní analýze. Naopak modely se smíšenými efekty se příliš nehodí k explorativním průzkumům „na zelené louce", kdy o datech jako by nic nevíme a chceme se nechat inspirovat daty. Tam může být výhodnější použití jiných (některé z vícerozměrných exploratorních) metod (viz Lepš & Šmilauer 2003).

V prvním dílu knihy jsme probírali (zobecněné) lineární regresní modely, o jejichž parametrech (koeficientech, kterými se násobí jednotlivé vysvětlující proměnné v lineárním prediktoru) jsme uvažovali vždy jako o neznámých, ale pevných (nenáhodných) hodnotách. Zde budeme často mluvit jak o koeficientech **pevných** (fixed, obdobných těm, jež vystupovaly v modelech z prvního dílu), tak **náhodných** (random, tedy koeficientech nejen neznámých, ale i náhodných, generovaných z nějakého pravděpodobnostního rozdělení). Bude-li model obsahovat v lineárním prediktoru jak pevné, tak náhodné efekty, budeme mluvit o **modelu smíšeném** (Mixed Effects Model či zkráceně Mixed Model). V žargonu smíšených modelů se typicky mluví o pevných a náhodných *efektech*. Efektem se přitom v zásadě rozumí koeficient v lineárním prediktoru (pevná/náhodná terminologie se nevztahuje k jiným v modelu se vyskytujícím parametrům, např. těm, jež popisují rozptyl či korelaci – ty jsou totiž vždy pevné). Můžeme tedy mluvit jak o efektu kategorické proměnné (odpovídá střední úrovni odpovědi pro nějakou úroveň faktoru – zcela v souladu se zažitou terminologií v modelech ANOVA), tak i o efektu spojité proměnné (směrnici). I tento efekt může být jak pevný, tak náhodný.

Ve zkratkovitém vyjádření někdy slýcháme, že daná vysvětlující proměnná ve smíšeném modelu je pevná či náhodná. Vzato doslova, je to vyjádření naprosto nekorektní. Ve standardních regresních modelech jsou vysvětlující proměnné *vždy* pevné (nenáhodné). Pokud jsou vysvětlující proměnné generovány nějakým náhodným mechanismem, modelujeme vše za podmínky, že jsou nastaveny na hodnotách zjištěných v dané studii (tedy obdobně, jako kdyby vysvětlující proměnné byly pevné). Je třeba si uvědomit, že adjektivum pevný/náhodný ve výše zmíněné terminologii se vztahuje k *něčemu jinému* než charakteru vy-

světlujících proměnných. Vztahuje se ke klasifikaci *koeficientů*, kterými se daná proměnná násobí.

Specifikace efektu dané proměnné jako efektu pevného, či náhodného, není dána „matematicky", ale povahou analýzy a tím, k čemu chceme (nebo nechceme) náš model použít. Naprosto nejde o volbu automatickou či mechanickou. Formálně můžeme pro každou vysvětlující proměnnou (ať už spojitou či faktor) uvažovat v zásadě jak náhodný, tak pevný efekt. Konkrétní situace může jednu z možností vylučovat (malá či špatně strukturovaná data např. neumožní odhadnout některé efekty jako náhodné). Záleží hlavně na záměrech naší analýzy. Dokonce až do té míry, že dva lidé mohou stejná data analyzovat různě: co pro jednoho bude efektem pevným, to pro jiného může být efektem náhodným, v závislosti na tom, oč jim v analýze jde. Pokud úrovně vybíráme náhodně (např. pomocí nějakého způsobu randomizace našeho experimentu) z celé populace úrovní, budeme efekty takového faktoru typicky specifikovat jako náhodné, zejména tehdy pokud nám půjde o zobecnění poznatků o pozorovaných efektech na celou populaci (efektů), z níž byl výběr pořízen. Nemusí tomu tak být vždy. Pokud je úrovní málo, můžeme být nuceni od (poněkud ambicióznější) formulace náhodného efektu ustoupit a spokojit se s analýzou efektů pevných. Uvědomme si, že v modelu, ve kterém efekt (koeficient) dané proměnné považujeme za náhodný, je třeba odhadovat rozptyl mezi průměrnou odpovědí pro různé úrovně této proměnné. K tomu, aby to byl odhad kvalitní (nebo alespoň snesitelný), potřebujeme v datech takových úrovní mít pokud možno „hodně". Například nebude příliš rozumné odhadovat rozptyl z 3 úrovní (pro odhad rozptylu je totiž relevantní počet opakování dán počtem úrovní daného faktoru, nikoli celkovým počtem pozorování).

Motivací pro volbu náhodného efektu může být modelování korelací v datech (ty jsou přítomností náhodných efektů implicitně indukovány). Dalším důvodem pro to považovat daný efekt za náhodný může být i snaha zobecnit inference na celou populaci efektů, z nichž ty v datech pozorované jsou jen vzorkem. Hranice tu však není tak ostrá, aby nedovolovala různé náhledy. V příkladu o klinické studii z prvního dílu se čtyřmi druhy bakterií (kap. 8.5) nám šlo právě o odpověď těchto čtyř konkrétních druhů – tedy o odhad pevných efektů (jež nám pak umožnily dané druhy srovnat, otestovat jejich ekvivalenci apod.). Kdybychom ale sledovali mnoho druhů, které bychom náhodně vybrali dejme tomu z mnoha druhů bakterií jedné taxonomicky (nebo ekologicky) definované skupiny, a šlo by nám o zobecnění našich výsledků na celou skupinu (např. by nás zajímal odhad mezidruhové variability v celé skupině), pak bychom efekt druhu bakterie asi považovali spíše za náhodný.

Pro pevné efekty jsou výsledkem analýzy **odhady středních hodnot** pro jednotlivé úrovně vysvětlující proměnné (nebo vhodně zvolený rozdíl středních hodnot v závislosti na zvolené parametrizaci), obecně odhady (nenáhodných, ale neznámých) koeficientů z lineárního prediktoru. Pracujeme-li s náhodnými efekty, získáme **odhady jejich rozptylů** (eventuálně též korelace/kovariance mezi různými náhodnými efekty, pokud je model dovoluje). Odhady (podmíněných) středních hodnot zahrnujících náhodné efekty lze také pořídit. To je výhodné pro některé aplikace (např. genetické či plemenářské či pro sestavování různých žebříčků dle výkonnosti), není to ale nutné. Pro mnoho aplikací jde buď jen o komponenty

rozptylů, odhad korelací, nebo dokonce pouze o korektní výpočet testu v modelu přítomných pevných efektů se zohledněním korelací, jež v datech jsou.

1.4 Konvence

Používáme stejných obecných konvencí pro zápis modelů, proměnných a R funkcí jako v prvním dílu. To znamená, že v textu využíváme dva základní typy fontů: Courier New pro příkazy v prostředí R a Times New Roman pro ostatní text. Courier New tučný označuje uživatelem zadávané příkazy a jejich argumenty, Courier New obyčejný pak názvy objektů a odpovědi programu. Pro úsporu místa jsme některé dlouhé výpisy z R programu zkrátili na nezbytně nutné minimum. Názvy proměnných, hodnoty parametrů a matematické formulace jsou psány kurzivou, názvy úrovní faktorů ve strojopisných uvozovkách. Jména balíčků (packages) jsou podtržena.

Grafy v rámci EDA byly vytvořeny s použitím pokud možno co nejmenšího počtu příkazů, proto jim často chybí popisky, legendy apod. Teprve finální grafy obsahují všechny detaily (za cenu delších příkazů). Všechny grafy byly vytvořené v černobílé verzi s použitím argumentu **`col=1`**. Tento argument byl pro úsporu místa z příkazů vynechán. Podobně neuvádíme argumenty pro vypsání názvů grafů (**`main`**) a pro dělení grafického pole (**`par(mfrow)`**).

Přirozený logaritmus (se základem *e*) se v prostředí R zapisuje **`log`**. Jako oddělovač desetinných míst je použita tečka, nikoliv čárka. Hodnoty parametrů jsou v textu zaokrouhleny na 2 až 4 cifry.

K výpočtům byla použita verze R 2.13.0 (R Development Core Team 2012).

2 O DESIGNECH

V prvním dílu (Pekár & Brabec 2009) měla řada příkladů v podstatě stejný jednoduchý design: plně znáhodněný výběr jedinců, tj. **Completely Randomised Design**, ve zkratce CRD. To znamená, že pokud srovnáváme několik experimentálních variant (např. způsobů „ošetření"), přiřazujeme je zcela náhodným způsobem vybraným experimentálním jednotkám (např. losováním či generováním náhodných čísel). CRD má sice také svoje nevýhody (viz Sokal & Rohlf 2000), ale v mnohých běžných situacích je nejpragmatičtějším řešením.

CRD je charakteristický také tím, že pro každou kombinaci úrovní vysvětlujících proměnných máme několik opakování neboli **replikací**, které byly pořízeny za stejných (či podobných) podmínek. V klasickém pojetí vyhodnocování dat replikace dále zahrnuje nezávislost mezi opakováními. Ve stejném kontextu se pak leckdy užívá pejorativního termínu **pseudoreplikace** pro označení situace, ve které opakovaná pozorování nejsou nezávislá.

2.1 Replikace *versus* pseudoreplikace

Závislost mezi měřeními může vzniknout buď v důsledku prostorové, časové, genetické nebo fylogenetické příbuznosti, či jiné „souvislosti" mezi měřeními a/nebo subjekty. **Prostorové pseudoreplikace** (spatial pseudoreplications) vznikají, když všechna měření pořídíme blízko sebe – na jedné lokalitě nebo v jejím blízkém okolí. Například pH podél toku stejné řeky, klíčení semen rostlin umístěných v jedné misce, nebo výška rostlin na stejném pokusném políčku. Co je blízko a co daleko, však není ani zdaleka absolutní, hodně záleží na vlastnostech studovaného procesu. Měření provedená blízko sebe si samozřejmě budou mnohem podobnější než měření, mezi nimiž je vzdálenost větší. Stejně tak si budou podobnější i měření provedená na stejném zvířeti, na jedincích ze stejného vrhu, na stejném kusu materiálu apod. **Časové pseudoreplikace** (temporal pseudoreplications) dostaneme, když měření na stejných subjektech následují relativně krátce po sobě. Jako například odběry krve u krys v hodinovém intervalu, vážení stejných jedinců ploštic jednou týdně, výběr úlovku pastí v měsíčním intervalu atp. Za předpokladu, že mezi subjekty je nemálo variability, zatímco měřicí chyba je relativně malá, budou si měření stejného subjektu (krysy, ploštice, pasti) mnohem podobnější, než kdyby byla provedena vždy na jiném subjektu.

Termín pseudoreplikace v ekologii zpopularizoval Hurlbert (1984). Jeho publikace způsobila mezi biology opravdový poprask. Hurlbert totiž na závěr svého článku doporučuje editorům časopisů, aby studie, které obsahují nesprávně analyzované pseudoreplikace, zamítli. Mnozí editoři ho poslechli, bohužel ne všichni si jeho článek přečetli dostatečně pozorně. Studie, jež obsahovaly pozorování nějakým způsobem závislá, byly pak až nadměrně