MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

NELINEÁRNÍ MODELY V PROSTŘEDÍ R

STANO PEKÁR
MAREK BRABEC

MASARYKOVA UNIVERZITA

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT
NELINEÁRNÍ MODELY V PROSTŘEDÍ **R**
3. díl

STANO PEKÁR, MAREK BRABEC

MUNI
PRESS

Knihu recenzoval: doc. RNDr. Petr Šmilauer, Ph.D.

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

NELINEÁRNÍ MODELY V PROSTŘEDÍ R

3. díl

STANO PEKÁR
MAREK BRABEC

Masarykova univerzita, Brno 2019

https://www.press.muni.cz/analyza-3-dil

Pekár S. & Brabec M. 2019. Modern Analysis of Biological Data. 3.
Non-Linear Models in R. Masaryk University Press, Brno.

ISBN 978-80-210-9278-5
ISBN 978-80-210-9277-8 (brožováno)
ISBN 978-80-210-9784-1 (vázaná)

OBSAH

PŘEDMLUVA

Třetí díl navazuje na předchozí dva a rozšiřuje tak skupinu dosud prezentovaných regresních modelů o modely nelineární. Ty obecně sice nejsou tak snadné na pochopení jako modely lineární, a tudíž ani přespříliš populární mezi biology, ale v některých oblastech biologie mohou být poměrně často používané. Například jsou časté ve fyziologii – vzpomeňme třeba sledování růstu organismů nebo enzymatickou kinetiku. Jsou totiž často odvozeny z mechanisticky či fenomenologicky formulovaných procesů, které mnohdy generují průběh nelineární ve vysvětlující proměnné (např. v čase). Ale i v ekologii se v poslední době objevují stále častěji a častěji. Například křivky popisující výskyt druhů v závislosti na gradientu environmentální proměnné. Naopak v oblasti behaviorálního výzkumu jsou data i zkoumané procesy mnohdy komplexnější a pozorované jen s nezanedbatelným šumem, který často znemožňuje formulovat detailní modely, takže zde často dominují jednoduché (např. lineární) modely jakožto generické aproximace. Společným rysem modelů obsažených v tomto díle bude nelinearita. Nelinearitou přitom nemyslíme jen nelinearitu vzhledem k vysvětlující proměnné, ale zejména nelinearitu vzhledem k parametrům, která má daleko závažnější důsledky, co se týče komplexity statistických odhadů i jiných inferenčních procedur a metod. Jinak to budou modely různě komplikované – od jednoduchých příkladů regrese až po vícenásobné regresní modely obsahující varianční a korelační struktury, pevné a náhodné efekty. Tedy vše to, co jsme probrali v prvních dvou dílech MABD (Pekár & Brabec 2009, 2012). A proto je nezbytně nutné, abyste si před čtením této knihy nastudovali nebo alespoň zopakovali zobecněné lineární modely a modely se smíšenými efekty. Na mnoha místech v tomto díle pak přímo odkazujeme na některé aspekty teorie či **R** funkce probírané v některém z předchozích dílů.

Svým konceptem je třetí díl podobný svým předchůdcům. Obsahuje minimum nezbytné teorie, která je dále rozebrána na devatenácti praktických příkladech. Ty jsou posbírány z různých odvětví biologie, například biochemie, ekologie, zoologie, botaniky a agroekologie. Tyto příklady pocházejí z různých reálných projektů. Data byla ale upravena tak, aby vyhovovala účelu použití – zjednodušena, zkrácena atd. Věříme, že tato „manipulace" neovlivnila jejich atraktivitu a srozumitelnost.

I v tomto dílu jsme k analýze dat použili software **R**. Samozřejmě v době psaní aktuální verzi. Je možné, že ti z vás, kteří s nelineárními modely pracují, již používají jiný software, speciálně vyvinutý pro nelineární regresi. Existuje jich celá řada. K nejznámějším asi patří CurveExpert, SigmaPlot nebo TableCurve. Výhodou těchto programů je snadné ovládání a někdy možná i lepší detekce a názornější řešení problémů s konvergencí. Konvergence numerického algoritmu pro odhad regresních parametrů

v nelineární regresi představuje, jak uvidíte, leckdy docela problém. A to problém mnohdy opravdu frustrující, problém, se kterým jste se možná doposud vůbec nesetkali.

Na závěr bychom chtěli poděkovat kolegům, kteří nám data propůjčili: P. Ghislandi, A. Honěk, J. Hubert, M. Chytrý, S. Korenko, B. Pekárová, J. Schenková a V. Šustr. Dále bychom rádi poděkovali O. Hájkovi za přípravu koordinát k síťovému mapování a hranic ČR. A také doc. RNDr. P. Šmilauerovi, Ph.D., za řadu cenných připomínek, které tento díl výrazně vylepšily.

Srpen 2019

Stano Pekár
Marek Brabec

1 ÚVOD

Obecně se studenti biologie potkávají poprvé s nelineárním modelem asi na přednáškách z fyziologie, kde se popisuje sigmoidní dynamika růstu organismů. Model růstu s horní i dolní asymptotou se sice na první pohled může zdát složitý, ale ve skutečnosti je velmi přirozený. Každý z nás jej zažil na vlastní kůži (byť o něm mnozí možná doposud nepřemýšleli). Podíváte-li se do vlastních zdravotních záznamů nebo do zdravotního průkazu svých dětí, uvědomíte si, že jednak rychlost růstu těla (v délce) není konstantní, a pak také to, že růst nepokračuje do nekonečna. To způsobuje nelineární, přibližně esovitě prohnutý tvar závislosti velikosti (délky nebo hmotnosti) těla na čase. Přitom jak tvar růstové křivky, tak i její asymptota (která souvisí s celkovým přírůstkem od narození) jsou individuálně specifické. Někteří z nás prostě rostou rychle a dosahují značné výšky, jiní rostou pomalu a jsou menší. Popsat tak komplikovanou změnu lineárním modelem nemusí být snadné. Avšak na základě relativně jednoduchých biologických úvah lze sestavit mechanistický (třeba na diferenciálních rovnicích založený) model růstu. Cílem takového modelu je podchytit nelineární průběh růstu v čase (průběh, ve kterém okamžitá růstová rychlost není konstantní) – tedy dospět k realisticky motivované „nelinearitě". Takovýto nelineární model (nelineární vzhledem k vysvětlující proměnné – času) může a nemusí být nelineární z pohledu běžně používané statistické terminologie. Ta modely třídí dle linearity/nelinearity vzhledem k parametrům, a nikoli vzhledem k vysvětlujícím proměnným. Mnohé nelineární modely je možné převést vhodnou transformací na modely lineární (viz první díl MABD, Pekár & Brabec 2009). To bychom měli udělat vždy, kdy to jde, protože statistické lineární modely jsou mnohem jednodušší na odhad i interpretaci. Bohužel to ale nejde vždy, a tak jsme v praxi leckdy nuceni pracovat i s modely nelineárními (v parametrech). Jak si ukážeme v této knize, k nelineárním modelům bychom měli sahat až po zralé úvaze – když byly možnosti lineárních nebo po-transformaci-lineárních modelů vyčerpány (nefitují data dobře) nebo např. když nám jde o odhad konkrétního parametru z dříve publikovaného modelu s jasnou věcnou (např. biologickou) interpretací.

Doposud (v obou předcházejících dílech) jsme se věnovali modelům lineárním nebo jim příbuzným GLM. Ty zahrnovaly velice pestrou škálu modelů nelineárních vzhledem k vysvětlující proměnné, ale lineárních vzhledem k parametrům (případně lineárních vzhledem k parametrům po vhodné transformaci). Dříve probíraná třída GLM (zobecněných lineárních modelů) je vzhledem k parametrům sice nelineární, ale umožňuje pracovat s nelinearitou elegantním způsobem, založeným na link transformaci a konkrétním rozdělením-implikované varianční funkci. Prediktor je sice stále lineární v parametrech, ale vztahuje se k potenciálně nelineárně transformované střední hodnotě. Tato elegance je však možná jen v relativně omezené třídě modelů (exponenciální třídě rozdělení). GLM třída zahrnuje mnohé důležité a v praxi hojně používané modely – připomeňme si třeba logistickou,

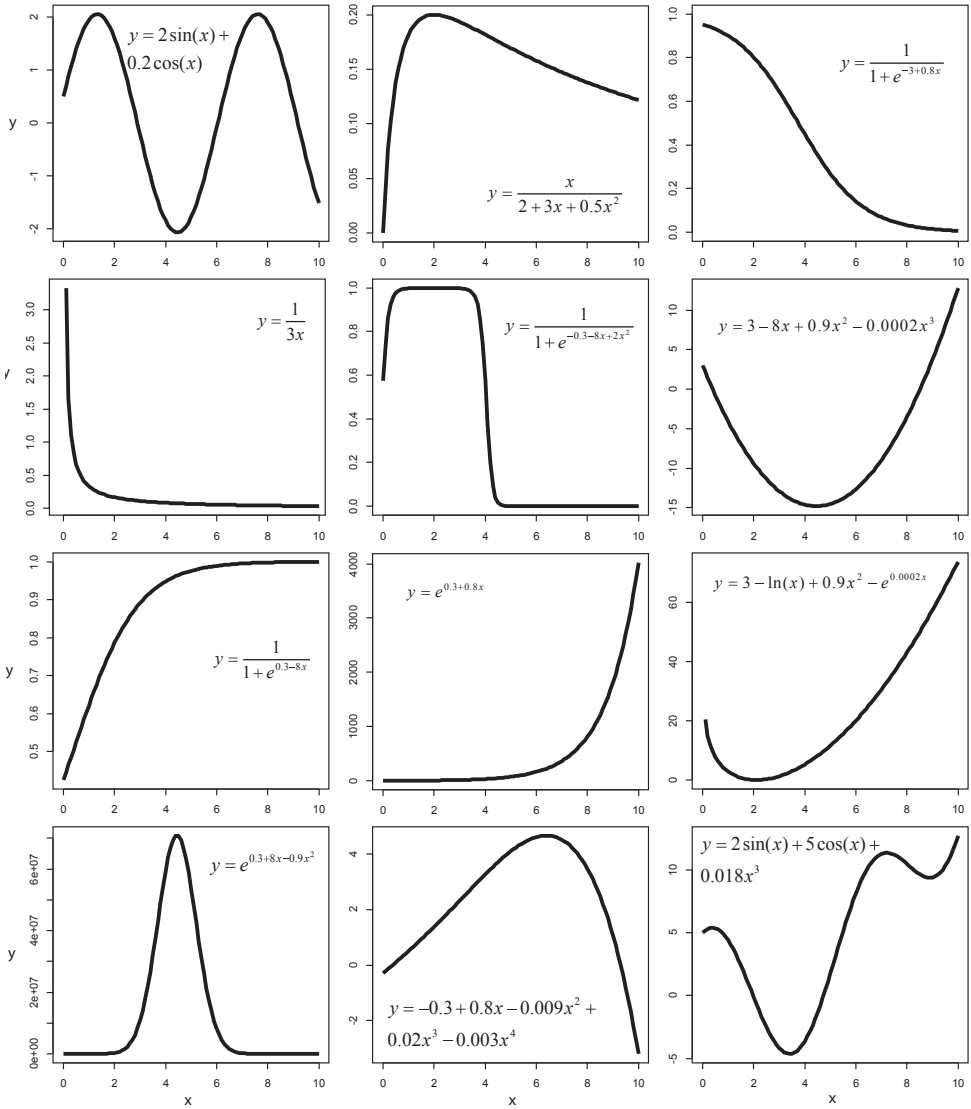
Poissonovskou i normální lineární regresi. Po vhodné (na neznámých parametrech nezávislé) transformaci vysvětlujících proměnných můžeme s pomocí GLM modelovat i poměrně velmi komplexní jevy, silně nelineární ve vysvětlující proměnné. Bohužel však zdaleka ne všechny. Například sigmoidní model růstu zmíněný výše lze sice převést vhodnou transformací na model lineární (lineární vzhledem k vysvětlující proměnné – času), ale tato transformace zpravidla závisí na neznámém parametru či parametrech. Právě jejich odhad již pak nelze zvládnout jako lineární statistický problém.

Z praktického pohledu je pro nás důležité naučit se poznat, kdy lze použít lineární (či po transformaci lineární) model a kdy již je nutné sáhnout k modelu nelineárnímu, pro nějž je statistický odhad parametrů (a odhad jejich nejistot) podstatně komplikovanější záležitostí.

Připomeňme si, že zde budeme – v souladu s předchozími díly MABD – používat koncept lineárního prediktoru zavedený v kontextu GLM třídy. Lineární prediktor je dán jen jako lineární kombinace neznámých parametrů, tedy nikoli složitěji. V lineárním prediktoru mohou být neznámé parametry násobeny hodnotami známých vysvětlujících proměnných a sečteny. Nic komplikovanějšího není povoleno. Za nelineární model budeme považovat ten model, jejž nelze (ani po vhodné link transformaci střední hodnoty a/nebo transformaci vysvětlujících proměnných a/nebo transformaci obou stran modelové rovnice) přepsat tak, aby byl lineární v parametrech lineárního prediktoru.

Jelikož lineární modely (či modely lineární po vhodné transformaci) umožňují popis relativně široké třídy situací, přinejmenším jako aproximace, mnozí uživatelé si s nimi vystačí po celou svou kariéru. Nevěříte? Podívejte se na obr. 1-1, abyste si uvědomili, jak různorodé tvary závislosti mezi modelovanou a vysvětlující proměnnou lze fitovat (po vhodné transformaci) jednoduše pomocí statistického lineárního modelu. A to i přesto, že jsou všechny silně nelineární vzhledem k vysvětlující proměnné.

Proč je linearita v parametrech tak důležitá? Poskytuje velikou výpočetní výhodu – jednoduchost a přímočarost odhadu pomocí metody nejmenších čtverců (v případě gaussovských lineárních modelů) a IRWLS (*iteratively reweighted least squares*) v případě obecných GLM. To je **obrovská** výhoda, kterou jste si doposud asi ještě ani neuvědomili, ale snad poznáte v následujících kapitolách (a oceníte poté v praxi) její přednosti. U nelineární regrese je často frustrující se odhadů vůbec dočkat – konvergence odhadovací procedury může být velmi pomalá. Pro ne zcela vhodně zadané startovací hodnoty nemusí daná odhadovací procedura dokonce konvergovat vůbec. Jak uvidíme později, algoritmy pro odhad parametrů jsou iterativní – startují z nějakých (typicky uživatelem zadaných) počátečních hodnot, které jsou na základě různě sofistikovaných optimalizačních metod zlepšovány. Fitování modelu je totiž podloženo maximalizací či minimalizací vhodného kritéria (např. věrohodnosti či součtu čtverců). Volba počátečních hodnot přitom nemusí být ani zdaleka triviální záležitostí. Kromě toho, na rozdíl od lineárního modelu, pro nějž je (např. za předpokladu normality modelované veličiny) dostupná rozsáhlá exaktní statistická teorie platná pro libovolný rozsah dat, pro nelineární modely takováto teorie typicky není k dispozici a je nutné se spoléhat na asymptotické aproximace. Ty dobře fungují pro „velká" data, ale pro „malá"



Obr. 1-1. Příklady 12 lineárních modelů (lineárních vzhledem k parametrům, jejichž konkrétní hodnoty jsou vypsány v obrázcích, a tedy lineárních ve statistickém smyslu) s nelineárním trendem ve vysvětlující proměnné.

už fungovat nemusí. To může být v praxi značný problém, kdy rozsah dat nemusí být příliš velký (a ani není tak úplně jasné, co „velký“ a „malý“ přesně znamená). Takže není úplnou vzácností, že na asymptotice založené odhady nejistot odhadnutých parametrů získané z fitovací procedury (byť velmi sofistikované a postavené na state-of-the-art znalostech) mohou být pro konkrétní data zcela mimo.

Pokud lineární modely poskytují takové výhody, možná si kladete otázku, k čemu nám vůbec nelineární modely jsou. Existují situace, kdy je použití lineárních modelů omezující. A to proto, že data v sobě nesou nelineární závislost, kterou nelze efektivně lineárními modely proložit. Pod termínem „efektivně" se skrývají (přinejmenším) tři aspekty: parsimonie, extrapolace a význam parametrů. Konkrétně: nelineární modely jsou často parsimonnější než ekvivalentní lineární modely, jsou-li odvozeny z mechanistického popisu studovaného jevu, chovají se často rozumněji za oblastí pokrytou pozorováními (měřenými daty použitými k identifikaci či fitu modelu), a konečně jejich parametry mají často relativně jasnou interpretaci, přímočarou a přitažlivou např. z biologického pohledu. Často přímo popisují nějakou teoretickou vlastnost studovaného procesu. To mohou být dostatečně silné důvody, proč se nelineárními modely zabývat. Na druhou stranu bychom se s jejich použitím nikdy neměli ukvapovat – jak uvidíme, mohou, zejména nepoučenému, nedostatečně teoreticky vybavenému nebo prostě jen nedostatečně přemýšlivému uživateli přinést mnoho problémů. Rozumný uživatel vždy dobře zváží nejen to, zda je možné žádaný model nějak exaktně převést na model lineární v parametrech, ale i možnost práce s lineárním modelem, který sice není zcela ekvivalentní modelu původnímu, ale může sloužit jako (pro požadovaný účel) dobrá aproximace. Aproximativní lineární modely mohou být také důležité pro odvození dobrých startovacích hodnot nezbytně potřebných pro nelineární odhadovací proceduru.

Nelineární model je typicky odvozen mechanisticky (např. je řešením diferenciální rovnice sestavené na základě teoretických argumentů o povaze populačního růstu). Také se vyskytující, ale méně vhodnou motivací je prostě detailní fenomenologický popis nějakého komplikovaného trendu. Jako příklad si podrobně probereme jeden takový model (používaný běžně v oblasti ekologie) v následující kapitole.

Co do počtu modelů je třída modelů nelineárních v parametrech podstatně větší než třída modelů lineárních v parametrech. Lineární modely jsou (velmi) speciálním případem nelineárních modelů. Jak jsme uvedli výše, lineární modely poskytují mnoho matematicky a statisticky dobře zdůvodněných užitečných funkcí (např. pro diagnostiku modelu, hodnocení kvality fitu apod.), jež jsou často implementovány ve výpočetním softwaru ve formě velmi pohodlné pro uživatele. Jak jsme se již zmínili výše, mnohé z analogů pro nelineární modely jsou komplikované, vybavené méně podloženou (např. jen asymptotickou) teorií. Oproti tomu lineární modely jsou založeny na známých explicitních (maticově zapsaných) vzorcích pro odhady (nevyžadujících iterativní řešení, a tedy ani počáteční odhady) a na velkém množství relativně jednoduché, ale velmi užitečné teorie. S modely nelineárními v parametrech je to jiné – pro uživatele mnohem méně příjemné. Už GLM (jako relativně „příjemná" podtřída nelineárních modelů) k odhadu potřebuje iterativní algoritmy (typicky založené na IRLS, tedy *iterative reweighted least squares*). Ty sice také potřebují startovací hodnoty, ale jejich vliv nebývá ve většině případů nikterak zásadní. Úspěšný „nástřel" startovacích hodnot je tak možné získat relativně snadno a většinou se lze spolehnout na to, co např. **`glm`** funkce pro startovací hodnoty vyprodukuje automaticky. Proto také mnozí uživatelé ani netuší, že nějaké startovací hodnoty jsou pro GLM modelování zapotřebí. Navenek to uživateli umožňuje stejně pohodlný život jako v případě lineárního modelu (fitovaného metodou nejmenších čtverců se známým explicitním řešením bez nutnosti iterovat a zadávat startovací hodnoty). V případě obecnějších nelineárních modelů než těch pocházejících

z GLM podtřídy je situace zcela jiná. Nejenže jsou iterační algoritmy pro moderní techniky odhadu parametrů nutností, kvalitní volba startovacích hodnot parametrů hraje často zcela zásadní roli a jejich plně automatická konstrukce mnohdy nemusí být k dispozici. To klade na uživatele nemalé nároky – už zadání dobrých startovacích hodnot často vyžaduje jistý vhled do struktury modelu, role jednotlivých parametrů i určitou matematickou hbitost (a aktivní znalost takových pojmů, jako je limita, derivace, spojitost apod.). Dekódování problémů s konvergencí (pochopení významu jednotlivých chybových hlášek použitého softwaru) a následný „troubleshooting" (poučené zkoušení různých variant fitu) může vyžadovat i některé další znalosti (např. o detailech použitého iterativního algoritmu). Také proto je potřeba vždy *dobře* zvážit, jestli použiji lineární, nebo nelineární regresi.

Často je velmi vhodné předem konzultovat použití složitějších nelineárních modelů se statistikem (a vyhnout se tak frustraci z neúspěšných pokusů o fit příliš ambiciózního modelu na nevhodná a/nebo málo informativní data), nebo dokonce blamáži s výsledky založenými na nekorektních odhadech parametrů s jinou interpretací, než jakou uživatel naivně předpokládá.

1.1 Jak číst tuto knihu

Tato kniha se zabývá dvěma skupinami statistických metod: parametrickou a neparametrickou nelineární regresí. Je napsána tak, že je čtenář může studovat odděleně. Na začátku každé části je představena trocha teorie, která je nezbytná k pochopení dané metody. Dále pak následují kapitoly, jež obsahují praktická řešení příkladů.

V první kapitole je vysvětlen rozdíl mezi lineárním a nelineárním modelem, přičemž je kladen důraz na praktické rozhodování, kdy který modelový přístup použít.

Ve druhé kapitole zopakujeme něco málo ze středoškolské matematiky. Cílem je ukázat čtenáři význam parametrů různých nelineárních funkcí. Na závěr kapitoly jsou ilustrovány některé základní vlastnosti vybraných funkcí. Pak se hlouběji podíváme na to, jak lineární, resp. nelineární funkční závislosti vznikají a jaké mají vlastnosti důležité z pohledu použití při analýze dat.

Třetí kapitola obsahuje teorii nutnou k pochopení toho, jak parametrická nelineární regrese pracuje a jaké jsou s ní spojeny problémy. Je zde stručně představen způsob odhadu neznámých parametrů v nelineárním regresním modelu. Dále se zde probírají různé rozšiřující specifikace základního nelineárního regresního modelu, například různé korelační struktury, které mohou být přítomny v reziduích. Nakonec jsou v této kapitole představeny některé vybrané **R** implementace založené na použití uživatelsky velmi přátelských tzv. *self starting functions*.

Kapitola čtvrtá je velice krátká, přesto důležitá. Má pomoci začátečníkovi zorientovat se v metodách a jejich **R** implementacích používaných v tomto díle pomocí jednoduchého klíče, nikoli nepodobného určovacím klíčům používaných biology.

Kapitoly 5 až 7 poskytují podrobná řešení celé řady konkrétních příkladů pomocí parametrické nelineární regrese. Jsou rozděleny dle typu modelu vzhledem k typu vysvětlujících proměnných a typu závisle proměnné. Jednotlivé příklady jsou prezentovány od popisu věcného problému přes zdůvodnění jednotlivých částí modelu až k formulaci závěru o zkoumaném jevu. Různé příklady se snaží ukázat vždy něco nového (jiný typ modelu, jinou korelační nebo kovarianční strukturu apod.). Postupy, které by se daly/měly pro různé příklady opakovat (takových je docela hodně), jsme ve výkladu vynechali. Vy je naopak při svých praktických analýzách s výhodou použijete.

Kapitola osmá je první kapitolou, která představuje neparametrický přístup k regresi. Neparametrické a semiparametrické regresi se detailně věnujeme v kapitolách 8–12. V teoretické části osmé kapitoly se dozvíte, jak neparametrická regrese pracuje, jaké typy jsou dostupné, a nakonec je jeden typ použit v konkrétní analýze.

Kapitoly 8 a 11 jsou především teoretické. Jsou zaměřeny na semiparametrickou regresi, na její stručné teoretické základy a implementaci v prostředí **R**.

Kapitoly 10 a 12 obsahují příklady z praktické analýzy dat pomocí semiparametrické regrese. Mají podobnou strukturu jako kapitoly 5 až 7. Také obsahují několik stejných příkladů jako kap. 5 a 6, pouze jsou zde analyzovány jinou metodou.

Datové soubory použité v této knize si můžete stáhnout z adresy https://www.press.muni.cz/analyza-3-dil. Na rozdíl od předchozích dílů jsou všechna data uložena v jednom excelovském souboru, každý datový soubor je na zvláštním listu označeném číslem kapitoly. Uživatel si je pak importuje do **R** přes schránku (clipboard), tedy nejprve označí do bloku, stiskne kombinaci kláves CRTL+C a pak napíše v prostředí **R** nasledující příkaz:

```
> dat<-read.delim("clipboard")
```

1.2 Konvence

Při psaní jsme se drželi stejných obecných konvencí pro zápis modelů, proměnných a **R** funkcí jako v předchozích dílech. To znamená, že v textu používáme 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. Na druhou stranu jsme příkazy a jejich argumenty zapsali s mezerami, abychom zlepšili přehlednost zápisu. Mezery však nemají pro zápis v příkazovém řádku vůbec žádný význam. Názvy proměnných, hodnoty parametrů a matematické formulace jsou psány kurzivou. Spojité kovariáty jsou psány malými písmeny, zatímco kategorické vysvětlující proměnné (factor v **R** terminologii) kapitálkami. Názvy úrovní faktorů uvádíme ve strojopisných uvozovkách. Jména packages (balíčků) jsou podtržena.

V kapitolách s příklady jsou data frame vždycky „připojena“ příkazem **`attach`**. To proto, abychom je pokaždé nemuseli specifikovat explicitně a zdlouhavě. Mnohé funkce však mají argument **`data`**, kterým lze vybraný data frame specifikovat bez jeho předchozího připojování. To je důležité, pokud střídavě budete pracovat s několika datovými objekty apod.

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). Mnohé grafy byly vytvořeny 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 v tomto dílu použita verze **R** 3.4.1 (R Core Team 2017). K výpočtům byly použity tyto balíčky: ape, lattice, nlme, mgcv, msm. Některé jsou součástí každé **R** verze, jiné je nutné si doinstalovat zvlášť.

2 FUNKCE

2.1 Funkční vztah

Jelikož nelineární modely jsou vesměs o funkčních vztazích mezi závislou (y) a vysvětlující proměnnou (x) či několika vysvětlujícími proměnnými, je dobré si připomenout a následně zapamatovat některé základní vlastnosti často používaných funkčních vztahů na příkladech zvládnutelných se znalostmi středoškolské matematiky. Pro konkrétní model, s nímž se chystáme pracovat, je třeba si uvědomit matematický význam jednotlivých parametrů. Ve zkratce funkci značíme písmenem f a v závorce se typicky uvádějí jména vysvětlujících proměnných a parametrů, např. $f(x, \theta)$, kde x je vektor vysvětlující proměnné (jejíž hodnoty jsou známé), θ je neznámý parametr, jenž bude odhadován z dat. Obecně je parametrů více než jeden (řekněme k) a pak θ bereme jako parametr vektorový, tedy $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$. Možných funkčních závislostí je (i pro jediný parametr) nekonečně mnoho, takže je samozřejmě nemůžeme probrat všechny. Stručně si nyní představíme jen několik relativně často používaných funkcí. Jmenovitě to jsou mocninné, racionální, exponenciální, logaritmické a trigonometrické třídy funkcí.

Mocninná funkce typu $y = x^b$ (pro $x > 0$) s parametrem b se používá pro jednoduchou aproximaci geometrického růstu, kde b určuje rychlost růstu. Pokud je parametr b větší než nula, jde o (monotónní) růst. Pokud je parametr b záporný, jde o pokles. A konečně, pro $b = 0$ máme konstantní funkci (nabývá konstantní hodnoty b bez ohledu na to, jakou hodnotu má vysvětlující proměnná x). Mocninná funkce může být doplněna o další parametry, např. $y = a \cdot x^b$ či $y = a \cdot (x - c)^b + d$ (pro $x > c$), kde c posunuje křivku podél osy x, a „škáluje", tedy násobí hodnotu $(x - c)^b$ a tím zrychluje nebo zpomaluje růst funkce, a d posunuje celou křivku podél osy y (obr. 2-1A).

Racionální funkce typu $y = a + \dfrac{b}{x + c}$ popisuje (pro $x > c$) asymptotický růst nebo pokles k nějaké hodnotě. Parametr a posunuje křivku podél osy y a definuje vodorovnou asymptotu (hodnotu, ke které se funkce blíží pro *hodně velké* hodnoty x). Znaménko parametru b pak určuje, zda se jedná o funkci rostoucí, či klesající (obr. 2-1B).

Exponenciální funkce typu $y = A^x$ (pro $A > 0$ a každé reálné x) má argument x v exponentu (na rozdíl od mocninné funkce, kde v exponentu máme konstantu). Používá se pro popis jevů, které se s rostoucím x velmi rychle zvětšují nebo zmenšují. Pro každé $A > 0$

můžeme tuto funkci ekvivalentně přepsat jako $y = A^x = e^{x \cdot \log(A)}$. Přirozeným zobecněním exponenciální funkce je funkce se třemi parametry: $y = a \cdot e^{b \cdot x} + d$, kde parametr *b* určuje základní tvar, který pak tento parametr „škáluje". Oba tedy ovlivňují derivaci (tj. „okamžitou rychlost") v kterémkoli bodě *x*. Parametr *d* posunuje celou křivku podél osy *y*. Tuto křivku samozřejmě můžeme dále zobecnit na křivku se čtyřmi parametry $y = a \cdot e^{b \cdot (x-c)} + d$, kde parametr *c* posunuje křivku podél osy *x* (obr. 2-1C). To se může jevit jako vždy výhodná a žádoucí vlastnost. Koneckonců, obecně, čím větší je počet parametrů funkce, tím je funkce „plastičtější" (ale pozor: nárůst „flexibility" pro další přidaný parametr se může s jejich počtem hodně snižovat – efekty různých parametrů spolu často interagují). Je dobré si však hned uvědomit, že větší počty parametrů přinášejí značné komplikace. Uvedenou exponenciální funkci totiž nebudeme moci bez dodatečných externích informací identifikovat („fitovat") jen z naměřených dat. Ptáte se proč? Tuto funkci lze rozepsat jako:

$$a \cdot e^{b \cdot (x-c)} + d = \left(\frac{a}{e^{b \cdot c}} \right) . e^{b \cdot x} + d = \widetilde{a} . e^{b \cdot x} + d \text{ pro } \widetilde{a} = \frac{a}{e^{b \cdot c}}.$$

Zvolíme-li pak např. $a = k \cdot a_0$ a $c = c_0 + \frac{\log(k)}{b}$ pro libovolné kladné *k*, dostaneme *naprosto* stejný průběh modelu (a tedy i *naprosto* stejný fit), jako kdybychom volili $a = a_0$ a $c = c_0$ Jinak řečeno, pro parametry *a* a *c* nejsme schopni dvojici jejich hodnot $(k \cdot a_0,\ c_0 + \frac{\log(k)}{b})$ a (a_0, c_0) z dat samotných naprosto nijak odlišit. To znamená, že bez dodatečných informací o hodnotě některého z parametrů budeme schopni identifikovat jen tři parametry, nikoliv všechny čtyři, a to z jakkoli velkých dat. Čtyřparametrový model tedy není identifikovatelný (*non-identifiable*). Avšak pokud bychom z externích zdrojů přesně znali hodnotu parametru *c* (třeba z nějaké teorie), lze zbylé tři parametry z dat odhadnout. Zmíněný problém není jen nějakou teoretickou hříčkou. Jde o zásadní potíž, se kterou se můžeme běžně setkat v praxi. Zatímco tříparametrový model může jít z dat velmi hladce odhadnout, model čtyřparametrový odhadnout prostě nelze (ať je dat jakékoli množství a ať jsou jakékoli kvality). Dobrá numerická odhadovací procedura by to měla poznat – a (většinou) i pozná a poví, i když pro laika ne vždy na první pohled pochopitelným způsobem. Například vydá hlášku o singularitě hesiánu (*Hessian singularity*), o překročení povoleného počtu půlících kroků (*halving steps*) nebo i hlášky jiné, ještě kryptičtější. V komplikovanějších modelech se můžeme navíc setkat i s tím, že singularitu (tedy to, že model z dat není identifikovatelný) prostě nepozná a jen bude mít „nějaké problémy s konvergencí". Je vždy na uživateli, aby o svém modelu a jeho důsledcích pro fitovaná data přemýšlel. Numerická procedura použitá k odhadu může chytrému napovědět, ale v žádném případě nemůže převzít odpovědnost za vše, co uživatel při formulaci pokazí, třeba přehnanými ambicemi na fit „velkého"/komplikovaného modelu. Řada v praxi se často vyskytujících potíží je však ještě subtilnějších/záludnějších. Model totiž může být teoreticky identifikovatelný, ale konkrétní data, na kterých se fit modelu provádí, identifikaci nemusí dovolovat. Intuitivně jasné to asi je v případě dat malého rozsahu, ale k obdobným potížím může dojít i u dat velkých a „nevhodně uspořádaných". Například data jen s malými středními hodnotami vysvětlující proměnné *x* pro model s několika parametry, z nichž jeden má význam asymptoty pro nekonečně velké *x*. S narůstající komplexitou modelu je těžší a těžší podobné problémy detekovat. Další komplikace přináší situace, ve které

se několik parametrů navzájem částečně kompenzuje (jejich odhady jsou asymptoticky relativně silně korelované) a k jejich dobré separaci je zapotřebí hodně a „dobře uspořádaných" dat. Analyzovat, co slova „dobře uspořádaných" přesně znamenají, nemusí být vůbec jednoduché. Přesně to však může obnášet návrh experimentálního designu v podmínkách konkrétního nelineárního modelu. To většinou vyžaduje spolupráci se specialistou – statistikem.

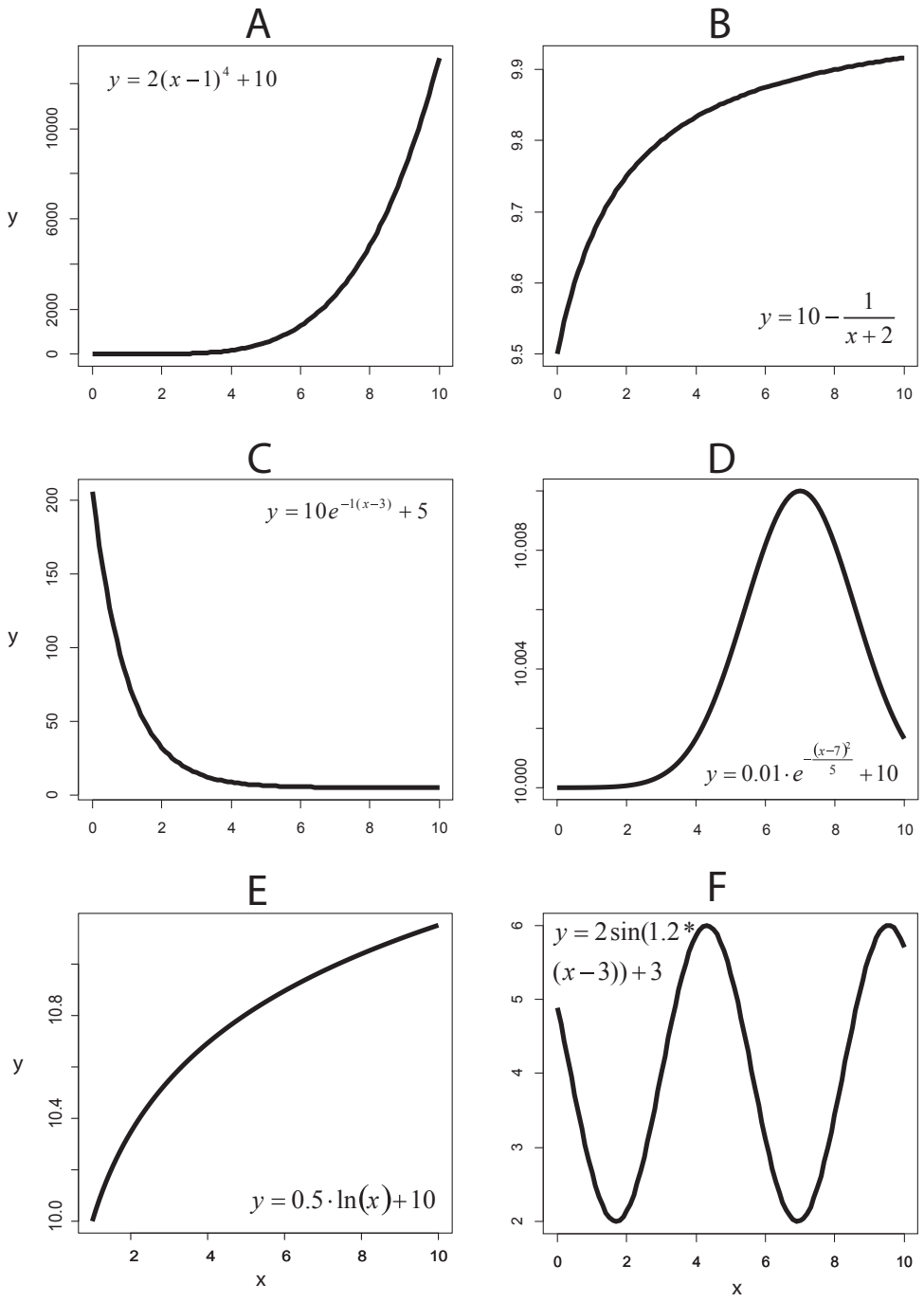
Dalším, často používaným čtyřparametrovým modelem je exponenciála s kvadratickým členem: $y = a \cdot e^{-\frac{(x-c)^2}{b}} + d$ (pro $b > 0$, $a > 0$). Povšimněme si, že takový tvar má i hustota normálního či Gaussova rozdělení (Gaussovského kopce), kde a udává výšku vrcholu, b souvisí se šířkou kopce, c definuje pozici kopce na ose x a d posunuje celý kopec podél osy y (obr. 2-1D). V praxi však může být velkým problémem všechny čtyři parametry (a, b, c, d) odhadnout z dat. Přestože model je formálně identifikovatelný, pro konkrétní velikost, kvalitu a uspořádání dat může být těžké (ba i nemožné) takový model fitovat. Jedním z důvodů, proč je s nelineárními modely „těžká práce", je i to, že identifikovatelnost je nutnou, ale nikoli postačující podmínkou k úspěšnému fitu na konkrétní reálná data. To je jev, se kterým se například u lineárních modelů normálně nesetkáváme (snad s výjimkou extrémní kolinearity). Při fitu čtyřparametrového modelu nastane problém, pokud třeba v datech nebudou zastoupeny dostatečně velké (absolutní) hodnoty vysvětlující proměnné x.

Logaritmickou funkci používáme nejčastěji ve tvaru $y = a \cdot \log_c(x) + d$, pro $x > 0$ a $c > 0$, kde c označuje základ logaritmu (obr. 2-1E). Typicky ovšem pracujeme s přirozeným logaritmem, tedy $c = e = 2.718282\ldots$, který budeme označovat jako log, případně s dekadickým logaritmem, tedy pro $c = 10$. Parametr a „škáluje" a parametr d posouvá křivku ve vertikálním směru.

Konečně trigonometrické funkce, jako jsou sinus nebo kosinus, popisují jevy, jež se opakují v pravidelných cyklech. Obecná sinová funkce $y = a \cdot \sin(b \cdot (x - c)) + d$ má čtyři parametry: a určuje výšku amplitudy, b definuje délku periody (či frekvenci), c posouvá křivku podél osy x, zatímco d podél osy y (obr. 2-1F). Z podobných důvodů, jako jsou ty zmíněné u čtyřparametrové exponenciální funkce, pro zcela obecnou fázi c není tento model identifikovatelný (protože funkce sinus je periodická s periodou 2π), fázi tedy potřebujeme pro identifikovatelnost intervalově omezit.

Obecně platí, že daný tvar křivky může být aproximován celou řadou funkcí z různých tříd a s různým počtem parametrů. Například, průběh racionální funkce z obr. 2-1B může být aproximována těmito funkcemi: exponenciální $y = a - b \cdot e^{-c \cdot x}$, mocninnou $y = a \cdot x^c$, nebo jinou racionální funkcí: $y = c + \frac{a \cdot x}{1 + b \cdot x}$ (funkce známá jako řešení Michaelis-Mentenové rovnice). Podobně kopcovito-hyperbolický průběh (obr. 3-1E) můžeme aproximovat exponenciální funkcí $y = a \cdot e^{-b \cdot x}$ (Rickerova křivka) nebo racionální funkcí: $y = \frac{a \cdot x}{(1 + c \cdot x)^b}$.

Výběr vhodné funkce se pak odvíjí od dalších vlastností, které požadujeme.



Obr. 2-1. Příklady některých funkcí. **A**. Mocninná. **B**. Racionální. **C**. Exponenciální. **D**. Exponenciální s kvadratickým členem. **E**. Logaritmická. **F**. Trigonometrická.

Kromě tvaru funkce, jejího matematického zápisu a významu jednotlivých parametrů, je často žádoucí poznat takové vlastnosti funkce, které jsou důležité pro daný praktický účel. Často nás například zajímají lokální extrémy. Lokální maximum nebo minimum funkce zjistíme tak, že spočítáme nulové body (kořeny) první derivace funkce. Jak víme ze střední školy, rozhodnutí o tom, zda takový kořen je maximem či minimem, souvisí s druhou derivací (pokud v bodě kořene existuje). Tiše přitom předpokládáme, že nějaký extrémní bod (maximum či minimum) pro vyšetřovanou funkci existuje uvnitř zkoumaného (otevřeného) intervalu, že existují i derivace v každém bodě atd. Tyto a další podmínky nejsou ani zdaleka vždy splněny – jejich ověření je plně na zodpovědnosti „uživatele" (typicky s nimi ale nebude problém u jednoduchých, velmi hladkých modelů podobných těm, které budeme používat k ilustracím v této knize). Nalezení první derivace (natožpak vyšších derivací, jež jsou důležité při hledání inflexních bodů, posuzování toho, zda jde o maximum či minimum apod.) funkce není pro nematematicky vzdělaného člověka vždy tak úplně snadné. Naštěstí lze v programu **R** leccos zvládnout jak numericky (vyhodnocením aproximace k požadované derivaci v požadovaných bodech), tak symbolicky (analytickým vyjádřením derivace obdobným tomu, které bychom dostali při odvozování na papíře). K výpočtu symbolických derivací funkcí jedné či více proměnných v **R** slouží procedura **D**. Ta umí derivovat leccos, ale samozřejmě ani zdaleka ne vše (zájemce o derivace či integrály komplikovanějších funkcí jednoznačně odkazujeme na specializované programy typu Mathematica®). Při výpočtu derivace nejprve definujeme požadovanou funkci pomocí příkazu **expression**, pak ji příkazem **D** zderivujeme podle některé z vysvětlujících proměnných, např. *x* (definovaná funkce může mít více vysvětlujících proměnných – v takovém případě pak **D** vůči jednomu z nich odpovídá parciální derivaci). Prvním argumentem **D** je rovnice funkce, kterou chceme derivovat, a druhým pak proměnná, podle níž se derivuje (ta je uvedena v uvozovkách – jako znaková proměnná). Postup si ukážeme na výpočtu derivace pro následující, relativně komplikovaně vypadající funkci:

$$y = 2 \cdot e^{-\frac{(x-5)^2}{3}} + 8 .$$

```
> y <- expression(2*exp(-(x-5)^2/3)+8)
> D(y, "x")
-(2 * (exp(-(x - 5)^2/3) * (2 * (x - 5)/3)))
```

Odpovědí programu je rovnice první derivace zadané funkce vůči jediné vysvětlující proměnné, *x*. S pomocí příkazu pro vykreslování analyticky zadaných křivek, **curve**, vyneseme nyní původní funkci i její derivaci do stejného grafu, abychom viděli, jak spolu souvisí.

```
> curve(2*exp(-(x-5)^2/3)+1,xlim=c(0,10), ylim=c(-1,3), lwd=2)
> curve(-(2 * (exp(-(x - 5)^2/3) * (2 * (x - 5)/3))), add=T, lty=2)
> abline(0, 0, lty=3)
```

Jak víme, derivace funkce odpovídá „okamžité rychlosti" (obr. 2-2). V bodě maxima (či minima) protíná osu *x*, v inflexních bodech dosahuje maxima či minima, v bodě maximální rychlosti nárůstu dosahuje vrcholu, v bodě maximální rychlosti poklesu dosahuje minima. Kladné hodnoty derivace znamenají, že původní funkce je rostoucí, záporné pak klesající. Mění-li první derivace znaménko, máme co do činění s ne-monotónní funkcí. Už z první derivace tak