Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Lineární statistické modely II

Pokyny a zadání k domácímu úkolu
jarní semestr 2019

Stanislav Katina
Vojtěch Šindlář
Markéta Janošová

26. května 2019

Vzorový domácí úkol. V souboru stka-vzor-du-template.pdf máte k dispozici vzorovou šablonu pro domácí úkol, vygenerovanou z následujících *.tex souborů:

1. stka-vzor-du-template.tex,
2. stka-vzor-text-template.tex a
3. stka-vzor-title-page-template.tex.

V odevzdávaném pdf souboru s domácím úkolem zachovejte styl použitý v šablonách.

Domácí úkol odevzdejte v jedné ze dvou níže uvedených forem. V názvech souborů nepoužívejte diakritiku a dodržujte velká a malá písmena podle návodu.

1. Forma Sweave

Tento způsob kombinuje k vytvoření řešení RSkript a flexibilní systém Sweave. Odevzdává se jeden pdf soubor nazvaný UCO-prijmeni-jmeno-LSM2-2019.pdf (obsahuje řešení příkladů, tabulky, obrázky, komentáře a náhled R-kódu), jeden zdrojový soubor naprogramovaných funkcí UCO-prijmeni-jmeno-funkce-LSM2-2019.R a jeden Sweave soubor UCO-prijmeni-jmeno-LSM2-2019.Rnw, z něhož byl vygenerován výsledný pdf soubor a který využívá zdrojový soubor naprogramovaných funkcí. V R Sweave se při používání LaTeX šablon postupuje identicky jako v LaTeXu.

K vygenerování R-kódu v požadované formě použijte v LaTeX-ovské hlavičce Rnw dokumentu balíček listings. Následujícím kódem umístěným taktéž v LaTeX-ovské hlavičce Rnw dokumentu upravíte původní nastavení vzhledu R-kódu a R-výstupů do požadované formy.

```
\definecolor{dgray}{gray}{0.35} % barva textu komentaru
\definecolor{lgray}{gray}{0.95} % barva pozadi R-kodu
\definecolor{llgray}{gray}{0.98} % barva pozadi R-vystupu

\lstdefinestyle{Rstyle}{ % nastaveni vzhledu R-kodu
language=R, % nastaveni jazyka R
basicstyle=\ttfamily\small, % typ a velikost pisma R-kodu
backgroundcolor=\color{lgray}, % barva pozadi R-kodu
commentstyle=\ttfamily\small\itshape\color{dgray}, % barva komentare k funkcim
showstringspaces=false, % zakaz zvyraznovani mezer
numbers=left, % cislovani vlevo
numberstyle=\ttfamily\small, % typ pisma a velikost cislovani
stepnumber=1, % cislovani po kroku jedna
firstnumber=last, % kumulativni cislovani radku v po sobe nasledujicich Chunk prostedich
breaklines=T} % automaticke zalamovani kodu na konci radku

\lstdefinestyle{Routstyle}{ % nastaveni vzhledu R-vystupu
language=R, % nastaveni jazyka R
basicstyle=\ttfamily\small, % typ a velikost pisma R-vystupu
backgroundcolor=\color{llgray}, % barva pozadi R-vystupu
showstringspaces=true, % zakaz zvyraznovani mezer
numbers=right, % cislovani vpravo
numberstyle=\ttfamily\small, % typ pisma a velikost cislovani
firstnumber=last, % kumulativni cislovani radku v po sobe nasledujicich Chunk prostedich
breaklines=T} % automaticke zalamovani kodu na konci radku
```

Dále je potřeba nastavit, aby byl balíček `listings` i s výše uvedenými nastaveními použit při překládání Rnw souboru do pdf souboru. Toto nastavení již vkládáme do těla dokumentu za příkaz `\begin{document}`.

```
<< setup >>= # Setup Chunk
render_listings()
@
```

Po přeložení Rnw souboru se náhled R-kódu automaticky zobrazí ve výsledném pdf souboru, pokud v hlavičce Chunk prostředí, obsahujícím R-kód, nastavíme argument `echo=T`.

```
<< echo=T >>=
x <- 1:10
plot(x)
@
```

Další informace o systému Sweave najdete např. zde: Chunk options and package options.

Při psaní R-kódu postupujte podle instrukcí v prezentaci Standards of programming in R: R style guide.

2. Forma LaTeX

Tento způsob kombinuje k vytvoření řešení RSkript a LaTeX. Odevzdává se jeden pdf soubor nazvaný UCO-prijmeni-jmeno-LSM2-2019.pdf (obsahuje řešení příkladů, tabulky, obrázky, R-kód napsaný v LaTeXu), jeden zdrojový soubor naprogramovaných funkcí UCO-prijmeni-jmeno-funkce-LSM2-2019.R a jeden soubor R-kódu konkrétních řešení DÚ UCO-prijmeni-jmeno-LSM2-2019.R, který používá tento zdrojový kód. Na psaní R-kódu použijte LaTeX-ovský balíček `listings` k vytvoření prostředí v hlavičce dokumentu pomocí následujícího kódu:

```
\definecolor{dgray}{gray}{0.35} % barva textu komentaru
\definecolor{lgray}{gray}{0.95} % barva pozadi R-kodu

\lstset{ % nastaveni vzhledu R-kodu
language=R, % nastaveni jazyka R
basicstyle=\ttfamily\small, % typ a velikost pisma R-kodu
backgroundcolor=\color{lgray}, % barva pozadi R-kodu
commentstyle=\ttfamily\small\itshape\color{dgray}, % barva komentare k funkcim
showstringspaces=false, % zakaz zvyraznovani mezer
numbers=left, % cislovani vlevo
numberstyle=\ttfamily\small, % typ pisma a velikost cislovani
stepnumber=1, % cislovani po kroku jedna
firstnumber=last, % kumulativni cislovani radku v po sobe nasledujicich castech s R-kodem
breaklines=T} % automaticke zalamovani kodu na konci radku
```

V textu potom R-kód vkládáme do prostředí `\begin{lstlisting}` a `\end{lstlisting}`. Při psaní R-kódu postupujte podle instrukcí v prezentaci Standards of programming in R: R style guide.

Pokud nemáte nainstalovaný LaTeX, můžete pro vygenerování souboru UCO-prijmeni-jmeno-LSM2-2019.pdf s textem domácího úkolu použít **Overleaf**.

- Na Overleaf si vyberte template UWE dissertation report. Tím dojde k vytvoření projektu, který pojmenujte např. DU-LSM2. Automaticky se vytvoří adresář files s několika soubory, z nichž některé je nutné přejmenovat a následně nahradit jejich obsah obsahem vzorových souborů a jiné vymazat takto:
  - přejmenujte main.tex na UCO-prijmeni-jmeno-LSM2-2019.tex,
  - přejmenujte Chapter1.tex na UCO-prijmeni-jmeno-LSM2-text.tex,
  - přejmenujte titlepage.tex na UCO-prijmeni-jmeno-LSM2-title-page.tex,
  - obsah souboru UCO-prijmeni-jmeno-LSM2-2019.tex nahraďte obsahem souboru stka-vzor-du-template.tex,
  - obsah souboru UCO-prijmeni-jmeno-LSM2-text.tex nahraďte obsahem souboru stka-vzor-text-template.tex,

  - obsah souboru UCO-prijmeni-jmeno-LSM2-title-page.tex nahraďte obsahem souboru stka-vzor-title-page-template.tex,
  - vymažte soubory Abstract.tex, biblio.bib a references.bib.

- V souboru UCO-prijmeni-jmeno-LSM2-title-page.tex modifikujte následující

```
\textbf{Nazev predmetu}
\textbf{Jmeno Prijmeni}
\textbf{UCO}
Obor XY
```

- V souboru UCO-prijmeni-jmeno-LSM2-2019.tex vyplňte následující

```
\fancyhead[L]{Nazev predmetu} %% hlavicka vlevo
\fancyhead[R]{Jmeno Prijmeni} %% hlavicka vpravo
```

- V souboru UCO-prijmeni-jmeno-LSM2-2019.tex modifikujte následující

```
\input{stka-vzor-title-page-template} %% nacteni souboru s titulni strankou
\input{stka-vzor-text-template} %% nacteni souboru s hlavnim textem ukolu
```

  zaměňte za své názvy souborů

```
\input{UCO-prijmeni-jmeno-LSM2-title-page} %% nacteni souboru s titulni strankou
\input{UCO-prijmeni-jmeno-LSM2-text} %% nacteni souboru s hlavnim textem ukolu
```

- Pro psaní ve slovenštině v hlavičce souboru UCO-prijmeni-jmeno-LSM2-2019.tex namísto

```
\usepackage[czech]{babel} %% zabezpeci ceske nastaveni
```

  použijte

```
\usepackage[slovak]{babel} %% zabezpeci slovenske nastaveni
```

- V souboru UCO-prijmeni-jmeno-LSM2-text.tex je zapotřebí postupovat takto:
  - text svého projektu pište buď v módu `Source` nebo `Rich Text`,
  - vkládání obrázků – vedle ikony `files` je šipka a z vyrolovaného menu vyberete `Computer` a uploadujete své obrázky jako `*.pdf`.
  - použití obrázků – příklad pro  logo v textu

```
\includegraphics[angle=0,width=0.025\textwidth]{Rlogo.jpg}
```

    Argument `width` určuje, jaká proporce šířky textu na stránce odpovídá šířce obrázku.
  - použití obrázků – příklad pro samostatný obrázek

```
%% prostredi obrazku
\begin{figure}[ht]
\centering
\includegraphics[angle=0,width=0.45\textwidth]{nazev-obrazku}
\caption{Popisek ...}
\end{figure}
```

  - použití tabulek – příklad

```
%% prostredi tabulky
%% zarovnani vpravo (r), pocet pismen "r" predstavuje pocet sloupcu
%% h - here, na tomto miste, t - top, v horni casti stranky
%% velikost pisma \footnotesize (10pt), \scriptsize (8pt)
\begin{table}[ht]
\caption{Popisek ...}
\footnotesize
\centering
\begin{tabular}{r||rrr|rrr}
%% telo tabulky
\end{tabular}
\end{table}
```

Export tabulek z R umožňuje knihovna `xtable` a její funkce `xtable`. Nastavení počtu desetinných míst je možné pomocí argumentu `digits`, kde první číslo vektoru je nula, neboť popis řádků je text.

– Vkládání R-kódu umožňuje prostředí `listings`

```
%% prostredi pro R-kod
\begin{lstlistings}
%% R kod
\end{lstlistings}
```

Ukázku vloženého kódu najdete v souboru stka-vzor-text-template.tex.

- Po dokončení domácího úkolu exportujete celý adresář DU-LSM2 (obsahující zdrojové soubory, obrázky) kliknutím na šipku pod ikonou DOWNLOAD AS ZIP, kde vyberete možnost Input and Output Files.
- Bližší informace o LaTeX-u najdete např. zde: The Not So Short Introduction to LaTeX.

DÚ je nejprve po formální stránce hodnocen cvičícím. Toto hodnocení zahrnuje:

1. *přítomnost tří výše zmíněných souborů a jejich názvy (při uploadu se nezaškrtne "přidat UČO, příjmení a jméno" a uploadujte jednotlivé soubory, nikoli *.zip, *.rar či jiné archivy),*
2. *kompletnost zpracování (každý příklad musí být vypracovaný, žádný nesmí chybět),*
3. *dostatečný opis Vašich úvah, zvoleného postupu a interpretace výsledků, ať už tabulkových nebo grafických,*
4. *přehlednost R-kódu a dodržování instrukcí v prezentaci Standards of programming in R: R style guide.*

DÚ je potřeba odevzdat do odevzdávárny přibližně 7 dní před termínem zkoušky, na který se přihlásíte. (Přesný termín odevzdání bude oznámen společně se zkušebními termíny.)

Zadání úkolů

Příklad 1 (Jednofaktorová ANOVA) *Mějme jednofaktorový ANOVA model* $\mathcal{F}_{H_1}$*:* $Y_{ji} = \mu_j + \varepsilon_{ji}$*, kde* $j = 1, \ldots, J$ *a* $i = 1, \ldots, n_j$ *(nevyvážené třídění), který zapíšeme maticově jako*

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

kde

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_J} & \mathbf{0}_{n_J} & \cdots & \mathbf{1}_{n_J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{J1} \\ \varepsilon_{J2} \\ \vdots \\ \varepsilon_{Jn_J} \end{pmatrix}.$$

Testujeme nulovou hypotézu H_0*:* $\mu_1 = \mu_2 = \cdots = \mu_J = \mu$ *oproti alternativě* H_1*: existuje alespoň jedno* $i \neq j$ *takové, že* $\mu_i \neq \mu_j$*. Dále označme jako* $\hat{\mathbf{y}}_0$ *(resp.* $\hat{\mathbf{y}}_1$*) odhad* $\mathbf{Y}$ *za platnosti* H_0 *(resp.* H_1*). Ukažte, že platí (doplňte vynechané části výpočtu):*

(a) $\mathrm{SS_e} = \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 = \cdots = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{j\cdot})^2,$

(b) $\mathrm{SS_A} = \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0\|^2 = \cdots = \sum_{j=1}^{J} n_j \, (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2,$

(c) $\mathrm{SS_T} = \mathrm{SS_A} + \mathrm{SS_e} = \cdots = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{\cdot\cdot})^2.$

Příklad 2 (Dvoufaktorová ANOVA) *V datovém souboru* anova-head.txt[1] *máme k dispozici antropometrické údaje mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). Vašim úkolem je zjistit, zda existuje vliv pohlaví (proměnná* sex*) na šířku hlavy v milimetrech (proměnná* head.W*) modifikovaný vlivem sexuální orientace (proměnná* sexor: op *– výlučně na opačné pohlaví,* sa *– ostatní, tj. jiné než výlučně na opačné pohlaví (bisexuální, homosexuální)).*

(a) Prohlédněte si data a odstraňte řádky obsahující chybějící hodnoty. Vykreslete krabicové diagramy popisující šířku hlavy v závislosti na pohlaví a sexuální orientaci. Vytvořte také krabicové diagramy v závislosti pouze na pohlaví. Zvolte stejnou škálu na ose y pro oba obrázky a nezapomeňte správně popsat jednotlivé „krabice“. Přidejte do nich zářezy, nastavte, aby jejich šířka odpovídala proporčně rozsahům a dokreslete do nich aritmetické průměry jako červené body.

(b) Modelujte závislost střední hodnoty šířky hlavy na pohlaví a sexuální orientaci. Vyzkoušejte různé varianty složitosti modelu: (1) model se vzájemnou interakcí obou faktorů, (2) model bez interakce a (3) model bez vlivu proměnné sexor*. Vyberte ten nejvhodnější z nich a své rozhodnutí zdůvodněte a podpořte příslušným výstupem z* R*.*

[1] Podrobnější popis datového souboru naleznete na `http://www.math.muni.cz/soubory/STKA/WEB-pdf/16-anova-head.pdf`.

(c) *Pro vámi vybraný model model z (b) vypište do tabulky odhady středních hodnot pro jednotlivé skupiny (tj. v případě modelu (1) skupiny symbolicky zapíšeme jako* f.op, f.sa, m.op *a* m.sa*, ale např. v případě modelu (3) uvažujme pouze skupiny* f *a* m*).*

(d) *Odhadnuté střední hodnoty šířky hlavy z (c) vykreslete do grafu zvlášť pro ženy (červená úsečka) a zvlášť pro muže (modrá úsečka). Na ose x budou úrovně faktoru* sexor *a na ose y šířka hlavy v mm. Uveďte také, jak získáte jednotlivé odhady středních hodnot pomocí příslušných koeficientů* β*.*

Příklad 3 (ANCOVA) *Opět pracujte s antropometrickými údaji studentů vysokých škol, tentokrát však se souborem* lrm-foot.txt *obsahující proměnné pohlaví (*sex*), délku chodidla v milimetrech (*foot.L*) a tělesnou výšku v milimetrech (*body.H*). Zajímá nás efekt pohlaví na tělesnou výšku adjustovaný na délku chodidla.*

(a) *Prohlédněte si data – pokud obsahují nějaké chybějící hodnoty, odstraňte příslušné řádky. Vykreslete do jednoho obrázku krabicové diagramy popisující výšku studentů v závislosti na pohlaví (stejnou formou jako v předchozím příkladu).*

(b) *Modelujte závislost střední hodnoty tělesné výšky na pohlaví a délce chodidla. Vyzkoušejte různé varianty složitosti modelu: model se vzájemnou interakcí pohlaví a délky chodidla (1 – všeobecný, různé sklony přímek), model bez interakce (2 – ANCOVA, stejné sklony přímek) a model bez vlivu proměnné pohlaví (3 – jedna přímka). Vyberte ten nejvhodnější a své rozhodnutí zdůvodněte a podpořte příslušným výstupem z R. Vypište také odhad vektoru* $\boldsymbol{\beta}$ *vašeho modelu.*

(c) *Vykreslete všechna pozorování jako bodový graf, kde na ose x bude délka chodidla a na ose y tělesná výška. Barevně rozlište muže i ženy a do obrázku umístěte i jednoduchou legendu. Tento postup třikrát zopakujte, přičemž do každého grafu vykreslete také přímky znázorňující odhad střední hodnoty postupně pro modely (1), (2) a (3). Všechny obrázky škálujte stejným způsobem, aby se mezi sebou daly porovnat.*

(d) *Dále vytvořte graf, který bude obsahovat všechna pozorování spolu s regresní přímkou odpovídající modelu (3). Vypočtěte z dat minimální a maximální délku chodidla a vytvořte v R posloupnost* 100 *bodů od získaného minima po maximum. Ve všech bodech této posloupnosti zkonstruujte oboustranné simultánní intervaly spolehlivosti s Bonferroniho korekcí hladiny významnosti a jejich hranice vyznačte do grafu. Přidejte také Scheffého pás spolehlivosti (v obou případech zvolte* $\alpha = 0{,}05$*) a okomentujte rozdíly.*

(e) *Pomocí* 95% *intervalu spolehlivosti predikujte výšku jedince, jenž má délku chodidla* 240 *mm. Vypište dolní a horní hranici tohoto IS a do obrázku z (d) příslušný interval znázorněte.*