Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika II – cvičení pro antropology

Pokyny k domácímu úkolu
jarní semestr 2023

Zdeňka Geršlová, Vojtěch Šindlář

211215@math.muni.cz

26. dubna 2023

Instrukce

Pro řádné splnění domácího úkolu je nutné odevzdat následující soubory:

1. Soubor MAS20c-UCO-prijmeni-jmeno-2023.R se zdrojovým R-kódem kompletního postupu ke všem řešeným příkladům (tzv. "R-skript")

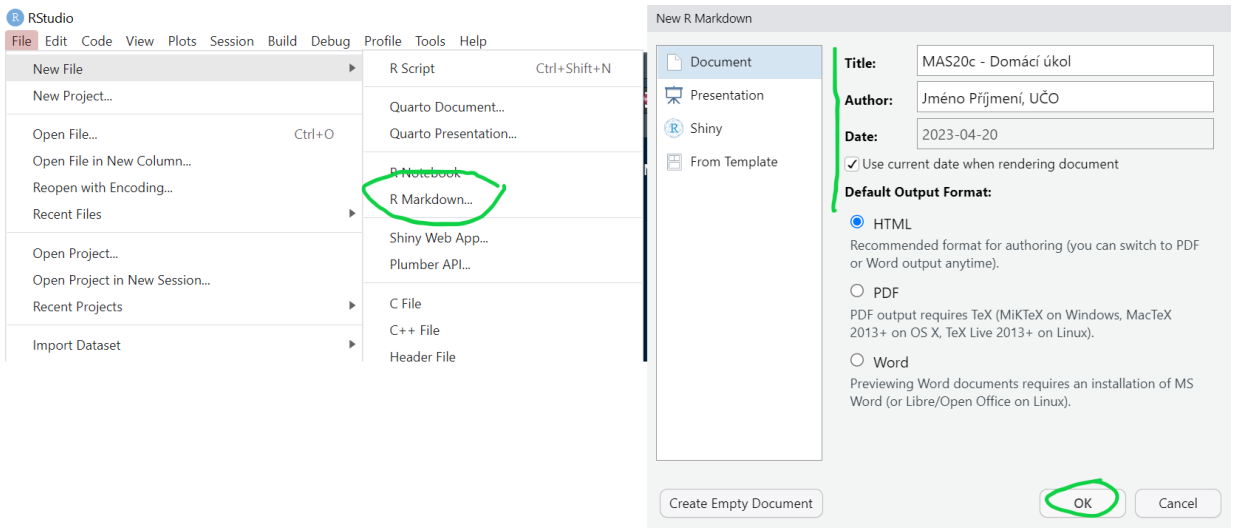
2. pdf soubor MAS20c-UCO-prijmeni-jmeno-2023.pdf, který obsahuje zpracované řešení příkladů (tj. komentáře k postupu, odpovědi na otázky ze zadání, obrázky, interpretace apod.)

Tento pdf soubor s řešením můžete vygenerovat některým z níže uvedených postupů, popř. použijte libovolnou jinou formu, ale vždy dbejte na to, aby výstup byl přehledný a zachovejte pro odevzdání formát .pdf. Jiné formáty (.doc, .docx apod.) nebudou akceptovány.

R Markdown

R Markdown je framework, který umožňuje vytváření reportů obsahujících obrázky, textové komentáře nebo ukázky R-kódu ve formátu pdf či html přímo z RStudia.

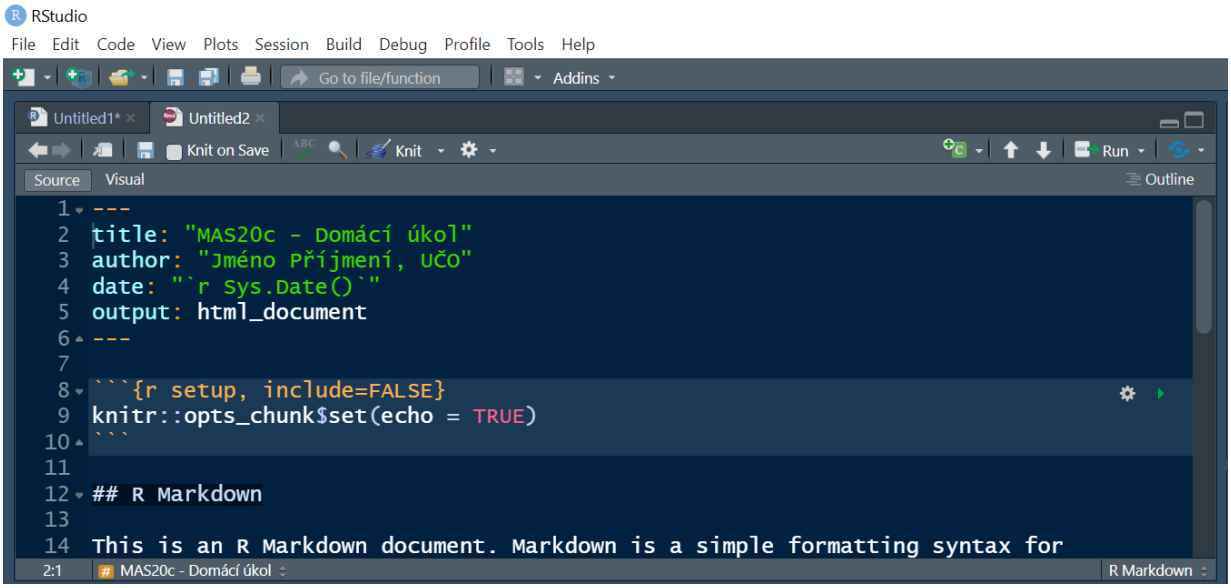
1. Pro vytvoření nového dokumentu typu markdown zvolte v RStudiu možnost File - New file a z nabídky vyberte možnost R Markdown.



Obrázek 1: Vytvoření dokumentu R markdown v RStudiu

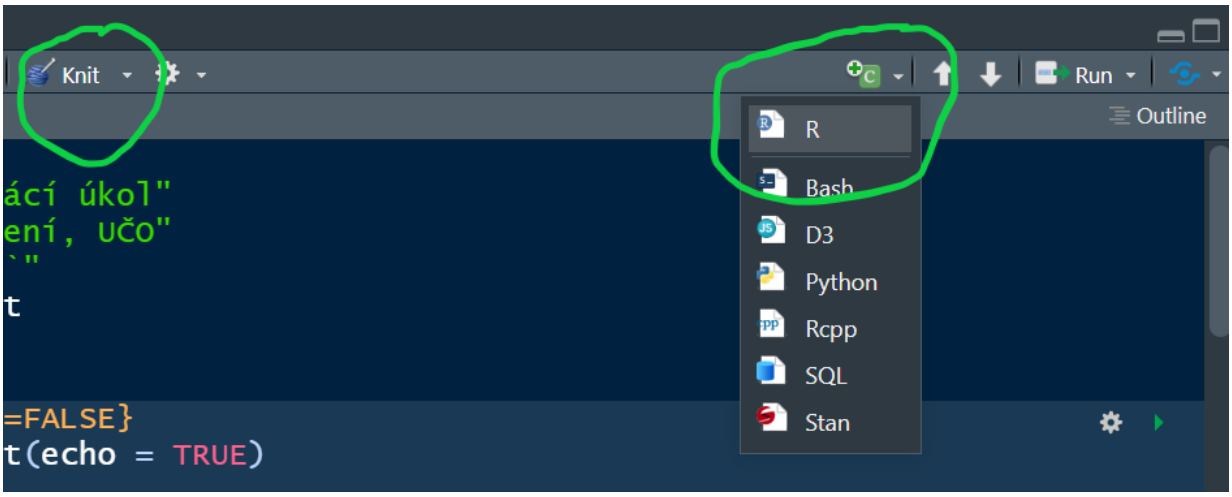
2. Následně se otevře okno pro zadání základních informací o dokumentu, které prosím vyplňte podle přiloženého ilustračního obrázku: na místo "Jméno Příjmení, UČO" vyplňte své osobní údaje a zaškrtněte možnost "Use current date when rendering document", volbu potvrďte kliknutím na OK.

3. Tímto byl vytvoření základní dokument, ve kterém je defaultně obsažena krátká ukázka (v AJ), jak vkládat obrázky, kód či nadpisy. Pro správné vygenerování výstupu ponechte prvních 10 řádků beze změny!



Obrázek 2: Hlavička vytvořeného markdown dokumentu

4. Pro vložení bloku R-kódu využijte zelené tlačítko +C a vyberte první možnost R (viz obr.3). Výstup potom vygenerujete pomocí tlačítka Knit, kde máte možnost zvolit formát html (pro rychlý náhled, zda dokument na výstupu obsahuje vše, co potřebujete) či pdf (výsledný výstupní formát, který budete odevzdávat).



Obrázek 3: Přidávání bloků kódu a generování výstupu

5. Výstupní pdf soubor prosím pojmenujte dle pokynů v sekci Instrukce (bod 2).

Pokud si zvolíte formu vypracování úkolu pomocí R Markdown, máte pro odevzdání R-kódu s postupem dvě možnosti:

1. Kód samostatně vyseparujete do zvláštního R-skriptu (tak, aby byl kompletně spustitelný, tj. včetně načítání dat apod.) a ten odevzdáte (tj. odevzdáváte soubor .pdf a soubor .R).

2. Odevzdáte přímo generující Markdown dokument, tj. soubor s příponou .Rmd, který obsahuje zdrojový kód k vygenerovanému pdf souboru. I v tomto případě prosím dodržte pojmenování dle sekce Instrukce, pouze příponu .R vyměníte za .Rmd.

Více o syntaxi v R Markdown najdete např. na https://rmarkdown.rstudio.com

Pozn: Kromě samotného Markdownu lze pro vygenerování souboru typu .pdf přímo z R využít např. Sweave, který je propojením R a LaTeXu.

Hodnocení

Hodnocení DÚ zahrnuje jednak správnost dosažených výsledků a interpretací, jednak formální stránku odevzdaných souborů. Hodnotí se následující:

1. přítomnost obou výše zmíněných souborů a jejich názvy (při uploadu v ISu 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 odevzdávaného pdf souboru a přehlednost R-kódu s ohledem na instrukce v prezentaci Standards of programming in R: R style guide.

Zápočet je udělen na základě hodnocení výše uvedených kritérií cvičícím předmětu.
Pokud by byly v odevzdaném DÚ nalezeny vážné nedostatky, bude úkol vrácen k přepracování

Domácí úkol odevzdejte nejlépe ještě před koncem semestru. Pro udělení zápočtu dostatečně včas (vzhledem ke zkoušce), pak prosím úkol odevzdejte nejpozději 7 dní před termínem zkoušky, na kterou se hlásíte.

Zadání

Příklad 1 *Pracujte s datovým souborem* du-kanga.txt*, který obsahuje údaje z měření lebek klokanů (jde o upravenou verzi souboru* kanga *z knihovny* faraway*). Soubor obsahuje údaje o pohlaví (proměnná* sex*), druhu (proměnná* species*) a 12 rozměrů naměřených na lebkách. Vaším úkolem je provést analýzu hlavních komponent (PCA) pro spojité proměnné.*

1. *Načtěte datový soubor a prohlédněte si strukturu dat. Pokud jsou v souboru nějaké chybějící hodnoty, odstraňte příslušná pozorování.*
2. *Vypočítejte korelační matici pro spojité proměnné. Napište příklad proměnných, které mezi sebou mají:*
   (a) *silnou pozitivní korelaci*
   (b) *negativní korelaci*
   (c) *slabou korelaci*
3. *Vypočítejte průměry pro všechny spojité proměnné a vykreslete přehled krabicových diagramů (tj. všechny diagramy v jednom obrázku). Rozhodněte na základě těchto údajů o tom, zda bude nutné v PCA použít škálování proměnných.*
4. *Proveďte PCA pro spojité proměnné. Vypište podíl variability a kumulativní podíl variability jednotlivých komponent. Jaký podíl variability vysvětluje druhá hlavní komponenta? Jaký podíl variability je vysvětlen prvními třemi hlavními komponentami společně?*
5. *Rozhodněte o počtu hlavních komponent, se kterými budete nadále pracovat. Kolik komponent vyberete podle*
   (a) *Kaiserova kritéria,*
   (b) *sutinového grafu,*
   (c) *požadavku na vysvětlení alespoň 80 % variability?*
6. *Vypočítejte korelaci původních proměnných s komponentami, které jste vybrali v předchozím bodě na základě Kaiserova kritéria. Pokuste se jednotlivé komponenty interpretovat, pokud je to možné (tj. slovně popište, v jakém vztahu jsou jednotlivé komponenty s původními proměnnými).*
7. *Vykreslete pozorování a proměnné v rovině prvních dvou hlavních komponent (tzv. biplot). V grafu označte pozorování druhem příslušného jedince. Je podle Vás rovina prvních dvou hlavních komponent dostačující pro rozlišení klokanů podle druhů?*
8. *Vypočítejte reziduální korelační matici.*

Příklad 2 *V tomto příkladu budeme opět pracovat s datovým souborem* du-kanga.txt*, ale nyní se zaměříme pouze na klokany druhu* fuliginosus*. Vaším úkolem je sestrojit model, který na základě proměnných* basilar.length*,* zygomatic.width*,* orbital.width*,* crest.width*,* mandible.length *a* mandible.width *určí pravděpodobnost, že neznámý jedinec je samec.*

1. *Načtěte datový soubor a odstraňte případná pozorování s chybějícími hodnotami. Potom z celého datového souboru vyfiltrujte pouze jedince druhu* fuliginosus*.*

2. *Dále pracujte pouze s proměnnými* basilar.length*,* zygomatic.width*,* orbital.width*,* crest.width*,* mandible.length *a* mandible.width*. Vypočítejte výběrové průměry a výběrové směrodatné odchylky těchto proměnných zvlášť pro samce a pro samice.*

3. *Prozkoumejte vztah mezi pohlavím a jednotlivými vysvětlujícími proměnnými pomocí t-testů. Nezapomeňte ověřit předpoklady (do řešení uveďte, jakou metodu jste na ověření použili a také patřičně okomentujte výsledky).*

4. *Sestrojte model logistické regrese závislosti pohlaví na 6 vybraných proměnných a proveďte celkový test významnosti modelu (tj. porovnání s modelem konstanty).*

5. *Najděte nejlepší model pomocí stepwise procedury.*

6. *Porovnejte model vybraný stepwise procedurou s plným modelem (sestaveným v kroku 4) pomocí:*

   (a) *AIC kritéria,*

   (b) *Nagelkerkova koeficientu,*

   (c) *relativní četnosti správně určených jedinců,*

   (d) *ROC křivek a AUC hodnoty.*

   Vyberte model, který je podle Vás nejlepší (své rozhodnutí odůvodněte).

7. *Pro model, který jste v předchozím kroku vybrali, vypočítejte poměry šancí a intervaly spolehlivosti. Popište slovně, jak se vzhledem k jednotlivým proměnným mění pravděpodobnost, že dané pozorování bude zařazeno mezi samce.*