

Rakendustarkvara: R. Kevad 2018, 10. praktikum*

1 Tulemuste vormistamine

Sageli kõige tüütum andmeanalüüsi juures on tulemuste vormistamine ilusaks dokumendiks. R-i väljund on tavaliselt kas konsoolis või graafikaaknas. Isegi kui tulemused on konsooliaknas tabelina, on seda Wordi vms programmi ümber tõsta tülikas. Õnneks on ka selle probleemi lahendamiseks R-is mitmeid erinevaid pakette, mille tööpõhimõte on sarnane:

1. Kirjutada oma analüüsi tekst koos vahepealsete koodiridadega ühte tekstifaili.
2. See fail kompileerida sobivasse formaati (nt word, html, tex või pdf).

Kui kasutada RStudiot, siis on taolise analüüsiraporti koostamine väga lihtne. Oletame, et punktis 1. nimetatud fail on näiteks järgmise sisuga:

```
---
output: pdf_document
---
# Esimene arvutus
```{r,echo=FALSE}
mass <- read.table("http://kodu.ut.ee/~annes/Rkursus/mass.txt", sep = "\t", header = T)
```

Massachusettsi andmestikus on kokku `r nrow(mass)` inimest; sooline jaotus on järgnev:
```{r,echo=FALSE}
table(mass$SEX)
```

## Esimene alapunkt
Nüüd on lisaks tulemusele endale näha, milline koodijupp leiab keskmise vanuse:
```{r,echo=TRUE}
mean(mass$AGEP)
```
```

See fail peaks olema salvestatud laiendiga *.rmd* ja avatud RStudios, siis vastava pdf-faili saab kompileerida kui vajutada menüüribal lõngakeraga nuppu *Knit*. Kui avada sama nupu noolemärgist rippmenüü, siis saab ka muuta koostatava faili tüüpi: *Knit to PDF* asemel võib valida ka *Knit to HTML* või *Knit to Word*. (NB! Pdf-faili tegemine eeldab, et arvutis on olemas MikTeX vms).

Selleks, et raporti genereerimine ise läbi proovida kontrolli kõigepealt kas installeeritud on paketid **knitr** ja **rmarkdown**. Seejärel ava uus markdown fail valides menüüst **File -> New file -> R Markdown**. Pane analoogselt ülaltoodud näitega faili kirja oma tekst ja märgendite vahele R-i kood, mis teeb arvutused, salvesta fail laiendiga *.rmd* ja vajuta nuppu *Knit to ...*. Markdown märgendkeeles faili vormistamise abi saamiseks vaata näiteks http://rmarkdown.rstudio.com/pdf_document_format.html

Ülalolevas näites on teksti vahele pikitud R-i koodi nn juppide/plokkide ehk *chunk*idega. Jupp võib olla nn reasisene ehk *inline*, või täiesti eraldiseisev. Reasisene R-i koodijupp tuleb piirata graavise sümboliga (*backtick*), mida eesti paigutusega klaviatuuril leiab *backspace* klahvi kõrvalt; reasisese koodijupi alguses peab olema täht **r**.

Sageli on mõttekas kasutada täiesti eraldiseisvaid koodiplokke, mis tuleks piirata kolmekordse graavisega. Ploki alguses tuleb loogelistes sulgudes kõigepealt kirjutada täht **r**, seejärel võib kirjutada komaga eraldatult lisaargumente¹. Mõned olulisemad:

*Praktikumijuhendid põhinevad aine MTMS.01.092 Rakendustarkvara: R (2 EAP) materjalidel, mille autorid on Mait Raag ning Raivo Kolde

¹<http://yihui.name/knitr/options>

- `echo` – kas väljundis peaks ka R-i kood olema näha (TRUE/FALSE)
- `eval` – kas koodiplokis olev kood täita või mitte (TRUE/FALSE)
- `fig.width`, `fig.height` – kui koodiplokis tehakse joonis, siis mis mõõtmetega see peab olema.
- `results` – kuidas vormindada väljundit (konsooli väljatrukki); väärtus `'asis'` on paslik sel juhul, kui kasutame mingit spetsiifilist vorminduskäsku (nt `xtable(.)`). Väärtuse `'hide'` korral väljundite ei näideta.

Oluline on, et eraldiseisva koodiploki algus ja lõpumärgised oleks eraldi real ja ilma eelneva tühikuta. RStudios on koodiploki märgendeid lihtne lisada klahvikombinatsiooni `Ctrl + Alt + I` abil või vajutades vastavat nuppu (*Insert*) menüüreal.

Koodiplokkide vahelise teksti puhul mõned olulisemad märgendusvõtted²:

- `#` Pealkiri – esimese taseme pealkiri
- `##` `alapealkiri` – teise taseme pealkiri jne (kuni 6 taset)
- Uue lõigu alustamiseks jätta üks tühi rida vahele
- Nummerdamata loetelu elementide ette näiteks – või `*`; loetelu ees peab olema tühi rida
- Nummerdatud loetelu kõigi elementide ette 1.
- Käsitsi saab tabelit vormistada miinusemärgi ja püstkriipsude abil
- allmärkuseid saab nii: `mingitekst^[allmärkuse tekst]`

Nii on võimalik R-i väljundit mugavasti ühte faili saada, et ei peaks pidevalt kopeerima-kleepima. Kuidas aga R-i produtseeritavad tabelid ilusaks saada? Lihtsamad tabelid saab kujundada **knitr** käsu **kable** abil. Näiteks kui eelnevalt on moodustatud tabel `tabel0`:

```
library(dplyr)
tabel0 <- mass %>% group_by("Kodakondsus" = CIT) %>%
  summarise("Keskmine vanus" = mean(AGEP), "Vanuse standardviga" = sd(AGEP)/sqrt(n()),
            "Soo ülekaal" = ifelse(sum(SEX == "Male") > n()/2, "Mehed", "Naised"))
```

siis tabeli esitamiseks väljundis on vaja lisada koodiplokk kujul (`format` määrab väljundi tüübi näiteks `latex`, `html`):

```
```{r koodijupp tabeli tegemiseks 1, results = 'asis'}
library(knitr)
kable(tabel0, digits = c(1, 2), align = c("lccr"), format = "markdown")
```
```

| Kodakondsus | Keskmine vanus | Vanuse standardviga | Soo ülekaal |
|---|----------------|---------------------|-------------|
| Born abroad of American parent(s) | 38.60 | 2.8 | Naised |
| Born in Puerto Rico, Guam, the U.S. Virgin Islands, | 45.04 | 2.1 | Naised |
| Born in the U.S. | 38.95 | 0.3 | Naised |
| Not a citizen of the U.S. | 37.91 | 0.8 | Mehed |
| U.S. citizen by naturalization | 50.27 | 0.9 | Naised |

On olemas ka spetsiaalseid pakette R tabelite esitamiseks ja näiteks mudelite väljundi ilusaks esitamiseks. Üks sellistest pakettidest on **xtable**, mille kõige olulisem käsk `xtable(.)` produtseerib etteantud tabelist (või vähegi tabelit meenutavast objektist, nt `data.frame`'ist) sobiliku LaTeX või HTML koodi:

```
```{r koodijupp tabeli tegemiseks 2, results='asis'}
library(xtable)
tabel0 <- table(mass$CIT, mass$SEX)
tabel <- xtable(tabel0, caption = "Sagedustabel", align = c("r|cc"))
print(tabel, type = "latex", comment = FALSE) # type = "latex" tabel LaTeX koodis
```
```

²http://rmarkdown.rstudio.com/authoring_basics.html

| | Female | Male |
|---|--------|------|
| Born abroad of American parent(s) | 31 | 31 |
| Born in Puerto Rico, Guam, the U.S. Virgin Islands, | 46 | 33 |
| Born in the U.S. | 2785 | 2637 |
| Not a citizen of the U.S. | 211 | 237 |
| U.S. citizen by naturalization | 226 | 187 |

Tabel 2: Sagedustabel

Lineaarse mudeli koefitsentide esitus tabeli kujul:

```
```{r mudeli kordajad, results = 'asis'}
mudel <- lm(WAGP ~ WKHP, data = mass)
print(xtable(mudel))
```
```

| | Estimate | Std. Error | t value | Pr(> t) |
|--|-------------|------------|---------|----------|
| (Intercept) | -71544.4589 | 11165.5063 | -6.41 | 0.0000 |
| AGEP | 3143.3746 | 345.3799 | 9.10 | 0.0000 |
| I(AGEP^2) | -31.2506 | 3.9803 | -7.85 | 0.0000 |
| LANXYes, speaks another language | -8156.2759 | 3144.4080 | -2.59 | 0.0095 |
| CITBorn in Puerto Rico, Guam, the U.S. Virgin Islands, | -14376.4311 | 12470.3254 | -1.15 | 0.2490 |
| CITBorn in the U.S. | 451.5616 | 8806.3954 | 0.05 | 0.9591 |
| CITNot a citizen of the U.S. | 533.0460 | 9357.1599 | 0.06 | 0.9546 |
| CITU.S. citizen by naturalization | 1321.1767 | 9383.0085 | 0.14 | 0.8880 |
| SEXMale | -12695.2778 | 5180.4127 | -2.45 | 0.0143 |
| WKHP | 1112.2084 | 95.5204 | 11.64 | 0.0000 |
| SEXMale:WKHP | 652.1225 | 128.8311 | 5.06 | 0.0000 |

1.1 Veel võimalusi vormistamiseks

R koodi ja analüüsiteksti võib siduda ka LaTeX keelt kasutades. Sellega alustamiseks vali menüüst: **File -> New file -> R Sweave**. Tekkiv fail salvesta laiendiga *.rnw*. Sel juhul võib teksti vormistamise osas järgida LaTeX reegleid.

Kui tekstireal on vaja teha mingi arvutus, siis selleks on käsk `\Sexpr{}`, mille argumentiks võib anda mingi muutuja koodist või arvutustehte. Eraldiseisvate koodiplokkide algusmärgendiks on `<<>=` ning lõpumärgiks `@`, ka siin peavad koodijupi algus ja lõpumärgised olema eraldi real ja ilma eelnevate tühikuteta. Koodiploki argumentid on samad, mis välja toodud markdown variandi juures, argumentid tuleb panna märgendite vahele, näiteks `<<echo = TRUE, results = 'hold'>>`. Teksti ja koodi siduva faili genereerimiseks vajutada munüüreal nuppu *Compile PDF*.

2 Kordamiseks

Läbitud teemade kordamiseks tee läbi järgmised ülesanded. Kasutatavas andmestikus on andmed maailma riikide rikkuse kohta, see sisaldab infot riikide rahvaarvu, koguvara ning erinevate maavarade ja loodusressursside osa kapitalist. Kõik rahanumbrid on dollarites inimese kohta.

Ülesanded vormista RMarkdown keeles või LaTeX abil.

2.1 Ülesanded

1. Loe sisse andmestik WB.txt aadressilt <http://kodu.ut.ee/~annes/Rkursus/>. (NB! pane tähele kuidas on tähistatud tühjad lahtrid)
2. Andmeid vaadates on näha, et osade riikide kohta on süstemaatiliselt andmed puudu.
 - Kui palju neid riike on?
 - Tekita uus andmestik, kus need riigid on välja visatud. Kasuta selle ülesande täitmiseks tunnust `Population`.
3. Mitu erinevat regiooni on esindatud selles andmestikus? Kui palju erinevaid riike igast regioonist on?
4. Kuidas on jaotunud riigid regiooni ja jõukuse kaupa? Koosta tabel, kasutades sobivaid tunnuseid.
5. Tekita tunnus nn naftariikidest ja neist kellel seda pole (tunnuse `Oil` väärtus peab olema suurem kui 0).
6. Kuidas jagunevad naftariigid regioonide kaupa?
7. Millised on Lõuna-Aasia regiooni naftariigid?
8. Tunnus `Total.wealth` näitab riigi kõikide varade summat inimese kohta. Arvutage keskmine varade maht inimese kohta üle kõigi riikide. (Vihje: lihtsalt keskmise võtmine annab kallutatud tulemusi, kuna rahvaarv riikides on erinev.)
9. Arvutage nüüd keskmine varade maht elaniku kohta kõigis regioonides eraldi. Lisaks näidake sealjuures ära ka iga regiooni rahvaarv.
10. Joonistage graafik, mis võrdleb riikide produtseeritud kapitali ja loodusvarades peituvat kapitali. Kas joonistub välja mingi trend?
11. Proovige lisada pildile ka riikide sissetuleku grupid.
12. Lisage pildile ka mõnede huvitavamate riikide nimed.
13. Eralda andmestikust veerud `Region`, `Population`, `Natural.Capital`, `Intangible.Capital` ja `Produced.Capital`. Leia nagu ennegi iga regiooni keskmine inimese kohta kõigi sõna “Capital” sisaldava nimega suuruste jaoks. Kujuta tulemusi sobiva joonisega.