

Lineaarsed mudelid
2. praktikum
Näiteid ja töövõtteid, veidike lisatud ka koosmõju maitset.

Loeme sisse mõned antud praktikumis kasutatavad andmestikud:

```
print(load(url("http://www.ms.ut.ee/mart/linmud2018/praks2.RData")))
```

Võrdlusnivoo muutmine

Esimeses näiteandmestikus oleme mõõtnud paaris linnas tunnuse y väärtuseid. Esitame lihtsa küsimuse: kas uuritava tunnuse y keskvärtus võiks olla erinevates linnades erinev?

Küsimusele vastamiseks hindame lihtsa dispersioonanalüüsi mudeli:

```
mudel1=lm(y~factor(linn), data=andmed1)
summary(mudel1)
```

Küsimused:

1. Milline tuleb nullhüpoteesi $H_0: \mu_{Tartu} = \mu_{Tallinn} = \dots = \mu_{tallinn}$ kontrollimisel saadav p-väärtus?
p-väärtus:
2. Kuidas interpreteerida linna "tallinn"-mõju, kuidas linna "Tallinn"-mõju? Miks üks neist mõjudest tuli statistiliselt oluline, teine mitte?

Hindame mudeli uuesti, valides võrdlusgrupiks linna „Tallinn“:

```
mudel2=lm(y~relevel(factor(linn), ref="Tallinn"), data=andmed1)
summary(mudel2)
```

Lihtne teadlane ei saa nüüd enam aru, miks R on peast segi läinud. Väidab ju R, et uuritava tunnuse keskvärtused on „Tallinn“-as ja „Pärnu“-s statistiliselt oluliselt erinevad (usutav), aga uuritava tunnuse keskvärtus grupis „tallinn“ on sama mis „Tallinnas“ ja sama mis „Pärnus“. Selline jutt aga on ju arulage ja sobilik pigem paberitega hullule, arvab töökas ja lihtne teadlane. Selgita õnnetule teadlasele – mis siis ikkagi toimub!

Otsusta millise grupiga võiksim/peaksim liitma grupi „tallinn“, kas grupiga „Pärnu“ või grupiga „Tallinn“. Tee seda ja teosta korrektne analüüs.

Muuseas: mida teha siis, kui uuritava tunnuse keskvärtused gruppides „tallinn“ ja „Tallinn“ osutuksid statistiliselt oluliselt erinevaks? Miks see võib päriselus ka nii juhtuda?

Koosmõjust

Järgnevalt vaatame (samuti genereeritud) andmestikku `ravi1`. Esmapilk sisseloetud andmestikule:

```
head(ravi1)

attach(ravi1)

table(sugu)
table(ravi)
```

Nagu näed, on andmestikus kahest soost inimesi, kes on saanud eksperimentaalset ravi (tunnus *ravi* väärtus *ravim*) ja teised, kes on jäetud ravita (*kontroll*). Kahe nädala pärast ravi alustamist (või kontrollide puhul peale kahte nädalat platseebo söömist) on mõõdetud patsientide vererõhku (kõigil vaadeldud patsientidel oli arsti poole pöördudes liiga kõrge vererõhk). Mõõtmise tulemusel saadi tunnus *sbp*.

Hindame mudeli uuritava tunnuse – vererõhu – kirjeldamiseks.

```
m1=lm(sbp~factor(ravi)+factor(sugu)+factor(ravi)*factor(sugu))
summary(m1)
```

Küsimused:

- Kas *ravi1* on mõju?
- Kui on, siis milline?
- Mida tähendab statistiliselt oluline "naiste" mõju?

Soovime otsida vastust ka järgmistele küsimustele:

- Kui suur on erinevus ravitud meeste ja ravita jäänud meeste vahel? Kas see erinevus on statistiliselt oluline?
- Kui suur on erinevus ravitud meeste ja ravitud naiste vererõhu vahel? Kas see erinevus on statistiliselt oluline?

Viimastele küsimustele vastamiseks on mitmeid võimalusi. Üheks võimaluseks on valida osavalt sobivaid võrdlusnivoosid – teatavate valikute korral peaksid vastused ülaltoodud küsimustele nähtavale ilmuma ka `summary`-käsu poolt tekitatud tabelis. Aga vahel on mõistlikum kasutada võimsamaid tööriistu.

Vajadusel installeeri

```
# install.packages("gmodels")
```

ja võta kasutusele lisamoodul `gmodels`:

```
library(gmodels)
```

Selle lisamooduliga saame oma käsutusse vägagi vajaliku funktsiooni `estimable`.

Alustame lihtsamast.

Vaata oma mudeli väljundit

```
summary(m1)
```

ja nuputa, mida võiks teha järgmised käsud:

```
estimable(m1, c(1,0,0,0))  
estimable(m1, c(0,1,0,0), conf.int=0.95)  
estimable(m1, c(1,1,0,0))
```

Nuputa välja, milline `estimable`-käsk annaks tulemuseks:

- a) ravitud ja ravimata meeste võrdluse
- b) ravitud meeste ja ravitud naiste võrdluse

Ülesanne 1

Vaata kolmandat tänase praktikumi andmestikku:

```
head(ravi2)  
  
attach(ravi2)  
  
table(ravi)  
table(rs123)
```

Legend: Inglismaal viidi läbi korralikult planeeritud randomiseeritud uuring uude ravimi kasulikkuse tuvastamiseks. Kas keskmine voodipäevade arv (näitab, kui kaua peab patsient haigusega voodis lamama) väheneb tänu ravile? Vaata ka alljärgnevat analüüsi:

```
t.test(voodipaevi~ravi)
```

Arvatakse, et uuritava haiguse kulgu võib mõjutada ka üks geenimutatsioon. Antud mutatsiooni olemasolu patsientidel kirjeldab tunnus *rs123*: *rs123*=A tähistab genotüüpi AA, *rs123*=G tähistab genotüüpe AG ja GG, mis peaksid käituma sarnaselt. Sestap hinnati ka mudel, kuhu lisaks patsiendi poolt saadud ravile lisati patsiendi genotüüpi iseloomustav tunnus:

```
m3=lm(voodipaevi~factor(rs123)+factor(ravi)+  
                                             factor(rs123)*factor(ravi))  
summary(m3)
```

Vaata hinnatud mudelit m3 ja kirjelda katsetatava ravimi mõju!

Kuidas oleks lugu sama raviga Eestis? Eestis on teada (geenivaramu andmetel), et 85% eestlastest kannavad lookuses rs123 varianti G (st genotüüpe AG või GG – mis käituvad fenotüübiliselt sarnaselt). Millist voodipäevade arvu muutust (keskmiselt) ootaksime nägavat Eestis, kui rakendaksime kaalumisel olevat ravi siin (paraku pole praktikas võimalik patsiendi genotüüpi enne raviga alustamist määrata)? Kui täpselt me ülalmainitud keskmist teame (raporteeri ka 95% usaldusintervall)?

Ülesanne 2

Mõõdetakse kuusepuude kõrgust ja vanust, lisaks pannakse kirja ka see, kas puukene on kasvanud valguse käes või varjus. Kas puu kõrguse modelleerimisel lineaarse mudeli abil tuleks kasutada vanuse ja valguse koosmõju või mitte? Põhjenda oma otsust!

Ülesanne 3

Alltoodud tabel kirjeldab, kuidas muutub Y -tunnuse keskvärtus (tinglik keskvärtus) siis, kui muudame tunnuste A (saab/ei saa ravimit A) ja B (saab/ei saa ravimit B) väärtuseid.

Kas antud situatsiooni kirjeldamiseks on tarvis kasutada koosmõjudega või ilma koosmõjudeta mudelit? Pane kirja dispersioonanalüüsi mudel, mis vastaks tabelis kirjeldatud situatsioonile.

	ei saa ravimit B (B-)	saab ravimit B (B+)
ei saa ravimit A (A-)	2	4
saab ravimit A (A+)	6	12

Ülesanne 4

Loe sisse andmestik, mis kirjeldab geeniekspressiooni sõltuvust erinevatest geeni kirjeldavatest tunnustest (andmestiku kirjeldus:
<http://www.ms.ut.ee/BDA/datadesc.pdf>):

```
mydata=read.table("http://www.ms.ut.ee/BDA/m52orfs.txt",
                  header=TRUE)

head(mydata)

attach(mydata)
```

Vaata esmalt, kas geeniekspressioon (tunnus `cai`, mudelisse mõistlik lisada logaritmitud kujul) ja geeni alguskoodon (esimesed 3 tähte DNA-ahelas, millega geen algab) on omavahel seotud:

```
m1=lm(log(cai) ~ first_codon)
summary(m1)
```

Milline on sinu otsus? Otsuse tegemiseks võid võrrelda ka järgmist kahte mudelit

```
m1=lm(log(cai) ~ first_codon)
m2=lm(log(cai) ~ 1)
anova(m1, m2)
```

Geeni pikkus (tunnus `length_aa`) võib samuti mõjutada geeni ekspressiooni. Proovi järgi! Lisa mudelisse ka geeni pikkus, kontrolli ka koosmõju lisamise vajalikkust. Tee joonis kolme enamlevinud alguskoodoni jaoks, mis näitaks, kuidas on seotud geeni pikkus ja tema ekspressioon!