Lineaarsed mudelid

3. praktikum

Dispersioonanalüüs; lisatud ka koosmõju maitset

Loeme sisse mõned antud praktikumis kasutatavad andmestikud:

```
print(load(url("http://www.ms.ut.ee/mart/linmud2021/praks3.RData")))
```

Võrdlusgrupp (ka referentstase või referentsnivoo)

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

Küsimusele vastamiseks hindame lihtsa dispersioonanalüüsi mudeli:

```
m1=lm(y~linn, data=andmed1)
summary(m1)
```

```
[…]

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.0221     3.1095   4.509 4.46e-05 ***
linnPärnu    -0.2797     3.2982  -0.085   0.9328
linnTallinn   2.3422     3.1627   0.741   0.4627
linnTartu     6.0660     3.2365   1.874   0.0673 .

Residual standard error: 3.11 on 46 degrees of freedom
Multiple R-squared:  0.3287,    Adjusted R-squared:  0.2849
F-statistic: 7.507 on 3 and 46 DF,  p-value: 0.0003434
```

Mõned tähelepanekud kasutatud R-i koodi kohta. Oleme seletava tunnusena kasutanud tunnust linn. Mudelit võime kujul y~linn kirja panna vaid selle tõttu, et linn on sisseloetud andmestikus juba faktortunnuseks teisendatud. Üldjuhul peaksime kasutama mudeli kirjapanekul kuju y~factor(linn) . Kui tunnus linn oleks esitatud kodeeritud kujul (näiteks 1-Tartu; 2- Tallinn; jne) siis hindaks mudel y~linn väga sobimatu lineaarse regressioonanalüüsi mudeli, käsk y~factor(linn) jõuaks aga sarnase analüüsini (kus igas linnas võiks olla uurtiava tunnuse keskväärtus täiesti omanäoline).

Küsimused:

1. Kas tunnusel linn on mõju tunnusele *y*?

   Vastus: jah, tunnusel linn on statistiliselt oluline mõju tunnusele *y*. Kust näeme seda? Mudeli kui terviku olulisustõenäosus (0,0003434) on väga väike – järelikult on mudelist kui tervikust kasu tunnuse y väärtuste prognoosimisel. Kuna praegu on mudelis seletavateks tunnusteks vaid tunnus „linn“ siis peab tunnuse linn mõju olema statistiliselt oluline. Antud juhul on mudeli kui terviku p-väärtus ka tõlgendatav kui hüpoteeside paari

   H_0: E(y|*linn*=Tartu) = E(y|*linn*=Tallinn) = E(y|*linn*=Pärnu) = ...
   H_1: leiduvad linnad *i* ja *j* nii et E(y|*linn*=*i*) ≠ E(y|*linn*=*j*)

2. Kuidas interpreteerid linna "Tallinn"-mõju (milliste keskväärtuste erinevust see mõju kirjeldab)?

Vastus: see näitab referentsgrupi ja nende, kellel linnaks on määratud „Tallinn“ keskväärtuste erinevust. Vaikimisi valib R referentsgrupiks faktortunnuse esimese taseme. Mis on esimeseks tasemeks näeme näiteks sagedustabelit koostades:

```
table(andmed1$linn)
```

Referentsgrupis on uuritava tunnuse keskväärtuseks antud mudeli järgi

```
14.0221
```

Ning „Tallinn“-as on uuritava tunnuse keskväärtuseks

```
14.0221+2.3422
```

Seega näitab hinnatud efekt uuritava tunnuse keskväärtuste erinevust referentsgrupi ja „Tallinn“-a vahel (hinnatud erinevust).

Referentsgrupiks valimine tähendab sisuliselt seda, et antud tasemele vastav indikaatortunnus visatakse mudelist välja (ehk vastavale tasemele vastav faktoranalüüsi efekt defineeritakse nulliks).

Referentstaset saab muidugi soovi korral muuta. Näiteks võime sama mudeli hinnata uuesti kasutades seekord võrdlusgrupiks linna „Tallinn“:

```
attach(andmed1)
m2=lm(y~relevel(linn, ref="Tallinn"))
summary(m2)
```

Lihtne teadlane ei saa nüüd enam aru, miks R on peast segi läinud. Väidab ju R, et uuritava tunnuse keskväärtused on „Tallinn“-as ja „Pärnu“-s statistiliselt oluliselt erinevad (usutav), aga uuritava tunnuse keskväärtus grupis „tallinn“ on sama mis „Tallinnas“ ja sama mis „Pärnus“. Selline jutt aga on ju arulage ja sobilik pigem paberitega hullule, arvab töökas tavateadlane. Selgita õnnetule teadlasele – mis siis ikkagi toimub? Kuidas saab „tallinna“ keskväärtus olla sama mis „Tallinna“ keskväärtus ja sama mis „Pärnu“ keskväärtus, aga samas on „Tallinna“ ja „Pärnu“ keskväärtused erinevad? Mis toimub?

Võrdlustaset saab muuta mitmel moel. Selgita mille poolest erinevad mudelid m2, m2a, m2b, m2c, m2d:

```
m2=lm(y~relevel(linn, ref="Tallinn"))
summary(m2)


linn2=factor(linn,
      levels=c("Tallinn", "Tartu", "Pärnu", "tallinn"))
m2a=lm(y~linn2, data=andmed1)
summary(m2a)


linn3=factor(linn,
      levels=c("Tallinn", "Tartu", "Pärnu", "tallinn"),
      labels=c("Reval", "Dorpat", "Pernau", "Reval2"))
m2b=lm(y~linn3, data=andmed1)
summary(m2b)


linn4=factor(linn, levels=c("Tallinn", "Tartu", "Pärnu"))
m2c=lm(y~linn4, data=andmed1)
summary(m2c)

linn5=factor(linn,
      levels=c("Tallinn", "Tartu", "Pärnu", "tallinn"),
      labels=c("Tallinn", "Tartu", "Pärnu", "Tallinn"))
m2d=lm(y~linn5, data=andmed1)
summary(m2d)
```

Selgitamisel võib abi olla ka sellest, kui vaatad defineeritud faktortunnuse sagedustabelit:

```
table(linn4, exclude=NULL)
```

Küsimus 1
Mida tuleks teha siis, kui uuritava tunnuse keskväärtused gruppides „tallinn“ ja „Tallinn“ osutuksid statistiliselt oluliselt erinevaks? Miks see võib päriselus ka niimoodi juhtuda (näiteks kui andmed on kogutud veebiküsitluse teel kus inimene ise peab oma elukoha sisestama võib täiesti oodata, et elukohaga „Tallinn“ ja „tallinn“ inimestel on uuritava tunnuse keskväärtus statistiliselt oluliselt erinev – oskad arvata, miks)? Kas sellisel juhul eelistaksid mudelit m2d või m2? Miks?

Küsimus 2
Kuidas tuleks hinnata mudelit, kui tahaksime saada kõiki linnu võrdluses Pärnuga?

Küsimus 3
Miks mudeli m1 korral kõik hinnatud parameetrid osutusid statistiliselt mitteoluliseks? Miks see võiks olla ootuspärane tulemus?

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.

```
mudel1=lm(sbp~factor(ravi)+factor(sugu)+
                        factor(ravi)*factor(sugu))
summary(mudel1)
```

```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         180.2800     0.9738 185.134  < 2e-16 ***
factor(ravi)ravim                     0.2300     1.3771   0.167    0.867
factor(sugu)naine                     5.8600     1.3771   4.255 2.61e-05 ***
factor(ravi)ravim:factor(sugu)naine -11.7400     1.9476  -6.028 3.80e-09 ***
```

Küsimused:

Kas ravil on mõju?

Vastus: Ravil on selle mudeli järgi mõju olemas ja see mõju on statistiliselt oluline. Selle nägemiseks võrdleme näiteks ravi saanud ja ravi mittesaanud naisi.

Uuritava tunnuse keskväärtus ravi mittesaanud naiste jaoks:

```
180.28 + 0    + 5.86 + 0
```

Uuritava tunnuse keskväärtus ravi saanud naiste jaoks:

```
180.28 + 0.23 + 5.86 -11.74
```

Seega ravi saanud naistel langes vererõhk (võrreldes mitteravitud naistega) 11,51 ühiku võrra. Antud erinevus on statistiliselt oluline – seda võime soovi korral ka estimable-käsu abil kontrollida:

```
# Hinnang suurusele E(sbp | ravi, naine) - E(sbp | ravita, naine)
library(gmodels)
estimable(mudel1, c(0, 1, 0, 1))

           Estimate Std. Error   t value  DF     Pr(>|t|)
  (0 1 0 1)   -11.51   1.377131 -8.357956 396 1.110223e-15
```

Vaadeldud erinevus tuli statistiliselt oluline. Selle mõistmiseks, et ravil on mõju ei pea aga tingimata estimable-käsku kasutama, piisab summary-käsu väljundist. Kui koosmõju tunnuste ravi ja millegi muu vahel on statistiliselt oluline, siis paratamatult peab ravil olema uuritavale tunnusele mõju (vähemalt mingis alagrupis). Alati ei pruugi muidugi õnnestuda tõestada millises alagrupis ravi just toimib, aga alati same kindlad olla, et kuskil või kellegi jaoks on ravil mõju olemas kui koosmõju tuleb statistiliselt oluline.

Mida näitab statistiliselt oluline "naiste" mõju?

Naiste mõju näitab seda, kuidas teineteisest erineb ravita jäänud meeste ja ravita jäänud naiste vererõhkude keskmine.

Ravita naiste keskmine vererõhk mudeli järgi:

```
180.28 + 0 + 5.86 + 0
```

Ravita meeste keskmine vererõhk mudeli järgi:

```
180.28 + 0 + 0    + 0
```

Kuidas iseloomustada koosmõju (`-11.74`) ?

Variant 1. Koosmõju näitab, kuidas muutub ravi mõju liikudes meeste (referentgrupp) juurest naiste juurde. Ravi mõju meestel (ravita meeste ja ravitud meeste erinevus): 0,23. Ravi mõju naistel (ravita naiste ja ravitud naiste erinevus): 0,23-11,74 = - 11,51. Koosmõju näitab kuidas muutub ühe sõltumatu tunnuse (ravi) mõju siis, kui muudame teise sõltumatu tunnuse (sugu) väärtust.

Variant 2. Koosmõju näitab, kuidas muutub meeste-naiste erinevus (ehk tunnuse sugu mõju) kui liigume ravita jäänud inimeste juurest ravi saanud inimeste juurde. Meeste ja naiste vererõhkude erinevus ravi mittesaanud inimeste korral: 5,86 (naistel kõrgem). Ravi saanud meeste ja naiste vererõhkude erinevus: 5,86-11,74=-5,88 (naistel madalam). Ehk koosmõju näitab kuidas muutub ühe sõltumatu tunnuse (sugu) mõju siis, kui muudame teise sõltumatu tunnuse (ravi) väärtust.

Reaalses elus ei võta mitte kõik patsiendid ravimit. Kui 70% naistest ja 60% meestest, kellele ravi kirjutatakse võtaksid ravimit, siis milline oleks meespatsientide ja naispatsientide keskmiste vererõhkude erinevus?

Antud küsimusele võime vastust otsida estimable-käsu abil. Leiame esmalt mõned abitulemused ja seejärel otsitava suuruse (kui saavutad estimable-käsu kasutamisel suurema vilumuse võid otse vajaliku arvutuse teha):

```
# Ravita meeste keskmine
estimable(m2, c(1, 0, 0, 0) )

# Ravitud meeste keskmine
estimable(m2, c(1, 1, 0, 0) )

# Meeste keskmine kui 60% meestest võtab ravi, E(X) = E E(X|Y)
estimable(m2, c(1, 0.6, 0, 0) )

# Ravita naiste keskmine
estimable(m2, c(1, 0, 1, 0) )

# Ravitud naiste keskmine
estimable(m2, c(1, 1, 1, 1) )

# Naiste keskmine kui 70% neist jälgib ravi
estimable(m2, c(1, 0.7, 1, 0.7) )

# Hinnang meeste ja naiste keskmiste erinevusele
#   E(sbp|mees)- E(sbp| naine):
estimable(m2, c(0, -0.1, -1, -0.7) , conf.int=0.95)
```

Näeme, et keskmised vererõhud meespatsientidel ja naispatsientidel ei pruugi teineteisest erineda kui 60% meestest ja 70% naistest saaksid ravi.

Ülesanne

Vaata kolmandat tänase praktikumi andmestikku:

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

Lugu andmete juurde: Inglismaal viidi läbi korralikult planeeritud randomiseeritud uuring uudse 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)
```

Põhimõtteliselt võiks uuringu sellega lõpetada. Oleme saanud tulemuse selle kohta, kas ravimit tasub kasutada või mitte.

Küsimus 4: Milline tuli otsus, selgita mille põhjal otsuseni jõudsid!

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 jätkasid uuringu tegijad ja hindasid puhtast huvist 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.

Küsimus 5: Kirjelda katsetatava ravimi mõju! Kas mõju eksisteerib, milline see mõju on?

Küsimus 6:
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 nägevat 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)? Kas antud ravi määramine Eesti inimestele oleks mõistlik kui raviga alustamise ajal ei tea me ravitava eestlase genotüüpi?