Biomeetria

4. praktikum

Dispersioonanalüüs ja näide kovariatsioonanalüüsist

Nominaalse (faktortunnuse) abil üritame prognoosida pideva tunnuse väärtuseid.

Loe sisse tudengite andmestik:

```
print(load(url("http://www.ms.ut.ee/mart/biomeetria2015/andmed.RData")))
```

Esimene katsetus. Proovime prognoosida tudengite pikkust kasutades informatsiooni tudengite sporditegemise tavade kohta.

```
table(sport)
```

Tunnus *sport* on andmestikus kodeeritud järgmiselt:

1 – ei tee sporti
2 – 1..2 korda nädalas
3 – 3..4 korda nädalas
4 – 5 või enam korda nädalas
5 – sisestusviga (seda väärtust ei tohiks andmebaasis esineda, kuid esineb siiski…)

```
m1=lm(pikkus~factor(sport))
summary(m1)
```

Tulemuseks saame järgmise väljundi:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    168.9363     0.7343 230.051  < 2e-16 ***
factor(sport)2   1.6167     0.8454   1.912   0.0563 .
factor(sport)3   4.3281     1.0516   4.116 4.36e-05 ***
factor(sport)4   9.0012     1.6214   5.552 4.12e-08 ***
factor(sport)5   4.8137     4.1541   1.159   0.2470

Residual standard error: 8.177 on 654 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.05959,   Adjusted R-squared:  0.05384
F-statistic: 10.36 on 4 and 654 DF,  p-value: 3.851e-08
```

Testib, kas tänu tunnuse sport kasutamisele on võimalik saada täpsemaid pikkuse prognoose.
H0: y-tunnuse keskmine on sama hea (või paremgi) prognoos kui meie mudeli (kasutades tunnust sport) abil leitav prognoos.
Antud juhul on p-väärtus väike ja saame tõestada, et meie mudel prognoosib täpsemalt kui tudengite kohta lisainformatsiooni mittekasutav mudel.

Paneme kirja ka meie poolt hinnatud mudeli matemaatilise valemi kuju:

$$E(pikkus|sport) = 168{,}9 + 1{,}6\, I_{sport=2} + 4{,}3\, I_{sport=3} + 9{,}0\, I_{sport=4} + 4{,}8\, I_{sport=5}$$

Indikaatorfunktsioon:
$I_{sport=2} = 1$, kui sport=2
$I_{sport=2} = 0$, kui sport≠2

$I_{sport=3} = 1$, kui sport=3
$I_{sport=3} = 0$, kui sport≠3

Proovi ise saadud väljundi põhjal vastata järgmistele küsimustele:

Milliseks hindab mudel mittesportivate (sport=1) tudengite keskmise pikkuse:

……

Milliseks hindab mudel 1..2 korda nädalas sporti tegevate (sport=2) tudengite keskmise pikkuse:

……..

Kontrolli enda poolt leitud vastuseid kahel viisil:

1. Kasutades hinnatud mudelit m1, prognoosi pikkust predict-käsu abil:
   `predict(m1,  data.frame(sport=c(1,2)))`

2. mean- käsu abil leia lihtsalt iga grupi keskmine:
   `mean(pikkus[sport==1], na.rm=TRUE)`
   `mean(pikkus[sport==2], na.rm=TRUE)`

Kas kõigil kolmel viisil arvutades jõuad sama tulemuseni?

Paku välja interpretatsioon hinnatud mudeli parameetritele: mida siis näitab antud mudeli vabaliige (Intercept), mida näitab real `factor(sport)2` toodud hinnang (1.6167)? Millise testi kohta käib `factor(sport)2` real toodud p-väärtus (0.0563), kuidas me seda p-väärtust saame interpreteerida? Kuidas interpreteerida Intercept-real toodud p-väärtust (< 2e-16)?

Hindame ka teise mudeli pikkuse prognoosimiseks:

```
m2=lm(pikkus~relevel(factor(sport), ref="5"))
summary(m2)
```

Saadud tulemus:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     173.7500     4.0887  42.496   <2e-16 ***
relevel(factor(sport), ref = "5")1  -4.8137     4.1541  -1.159    0.247
relevel(factor(sport), ref = "5")2  -3.1970     4.1101  -0.778    0.437
relevel(factor(sport), ref = "5")3  -0.4856     4.1574  -0.117    0.907
relevel(factor(sport), ref = "5")4   4.1875     4.3367   0.966    0.335

Residual standard error: 8.177 on 654 degrees of freedom
Multiple R-squared:  0.05959,   Adjusted R-squared:  0.05384
F-statistic: 10.36 on 4 and 654 DF,  p-value: 3.851e-08
```

Kas pikkuse prognoosiks tundub olevat parem mudel m2 või varem hinnatud mudel m1? Kas mudeli m2 abil saab tudengi pikkust prognoosida paremini (täpsemalt) kui lihtsalt alati kõigi tudengite keskmist pikkust prognoosiks pakkudes?

Võrdle ka mõlema mudeli (m1 ja m2) abil leitud prognoose ja prognoosiintervalle:

```
predict(m1, data.frame(sport=1:5))
predict(m2, data.frame(sport=1:5))

predict(m1, data.frame(sport=1:5), interval="prediction")
predict(m2, data.frame(sport=1:5), interval="prediction")
```

Mille poolest leitud prognoosid erinevad või sarnanevad?
Interpreteeri mudeli m2 parameetreid,

Ülesanne
Reparametriseeri mudel ümber selliselt, et kõik võrdlused oleksid tehtud 1..2 korda nädalas sportivate (*sport*=2) tudengite grupi suhtes (mudeli vabaliige peaks näitama 1..2 korda sportivate tudengite keskmist pikkust). Kas oskad seda teha?

Ülesanne
Kontrolli kas tudengi tervisliku seisundi (tunnus *tervis*) järgi on võimalik tudengi pikkust prognoosida (kas tervisliku seisundi põhjal antud pikkuse prognoos on tõestatavalt parem – täpsem – sellisest prognoosimisviisist, mis kõigile tudengitele pakub sama prognoositud pikkuse – tudengite keskmise pikkuse)?

Vaatame ka järgmist mudelit:

```
mudel1=lm(pikkus~factor(olu))
summary(mudel1)
```

Kas tudengi õlletarbimise järgi saab pikkust prognoosida?

Proovime ka võrrelda kahte mudelit. Hindame lisaks mudeli, kus pikkuse prognoosimiseks pole kasutada ühtegi tunnust – prognoosiks pakutakse selle mudeli järgi lihtsalt tudengite keskmist pikkust:

```
mudel0=lm(pikkus~1)
```

Võrdleme mudeleid mudel0 ja mudel1. Testime, kas lihtsam mudel prognoosib sama hästi kui keerukam (H_0: lihtsam mudel – *mudel0* – prognoosib sama täpselt kui keerukam mudel ehk *mudel1*):

```
> anova(mudel0, mudel1)
Analysis of Variance Table

Model 1: pikkus ~ 1
Model 2: pikkus ~ factor(olu)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    659 46512
2    655 41098  4    5413.1 21.568 < 2.2e-16 ***
```

Prognoosivigade (jääkide) ruutude summad mõlema mudeli jaoks

Prognoosivigade ruutude summa muutus (kui palju väiksemaks muutusid prognoosivead tänu keerukama mudeli kasutamisele)

Testi „H0: Lihtsam mudel prognoosib tulevasi vaatluseid sama hästi kui keerukam mudel“ p-väärtus.

Näeme, et tudengi poolt tarbitava õlle koguse teadmine aitab pikkust täpsemalt prognoosida…

Oleme teinud hea alguse ja leidnud lootustandva mudeli, mis aitab tudengi pikkust prognoosida. Proovime leitud mudelit täiendada ja prognoosime tudengi pikkust nii tema õlletarbimise kui ka tudengi soo (1- naine; 2- mees) järgi.

```
> mudel2=lm(pikkus~factor(olu)+factor(sugu))
> summary(mudel2)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     168.02792    0.37514 447.904   <2e-16 ***
factor(olu)<1    -0.19703    0.52584  -0.375    0.708
factor(olu)1-5    0.58049    0.78367   0.741    0.459
factor(olu)5-12  -1.76034    1.25775  -1.400    0.162
factor(olu)>13    0.08445    2.36404   0.036    0.972
factor(sugu)2    14.11890    0.64873  21.764   <2e-16 ***

Residual standard error: 6.037 on 654 degrees of freedom
Multiple R-squared:  0.4875,    Adjusted R-squared:  0.4836
F-statistic: 124.4 on 5 and 654 DF,  p-value: < 2.2e-16
```

Saame mudeli, mida võib soovi korral kirja panna ka järgmisel kujul:

$$Pikkus = 168{,}03 - 0{,}20\, I_{olu<1} + 0{,}58\, I_{1<olu<5} - 1{,}76\, I_{5<olu<12} + 0{,}08\, I_{olu>12} + 14{,}12\, I_{sugu=mees} + \varepsilon$$

Proovi ise arvutades leida milline tuleb selle mudeli prognoos:

a) Õlut mittetarbiva (olu="ei joo") naistudengi pikkusele:

……..

b) Õlut mittetarbiva (olu="ei joo") meestudengi pikkusele:

……..

c) Nädalas 5-12 pudelit õlut tarbiva naistudengi pikkusele:

……..

d) Nädalas 5-12 pudelit õlut tarbiva meestudengi pikkusele:

……..

Kontrolli enda poolt arvutatud prognoosid üle arvuti predict-käsu abil:

```
# Koostame andmestiku nende inimeste andmetega, kelle pikkust
# soovime prognoosida:
progandmed=data.frame(
            olu=c("ei joo","ei joo","5-12","5-12"),
            sugu=c(1,2,1,2)
         ) #data.frame-käsk lõppeb

# Pane tähele, et andmestikus on olemas kõik need tunnused,
# mida kasutati sõltumatute tunnustena (nn x-tunnustena) mudeli
# hindamisel:
progandmed

#Leiame nende nelja juhu jaoks oma mudeli prognoosid
predict(mudel2, progandmed)
```

Kuidas interpreteerida real `factor(sugu)2` hinnatud parameetrit (`14.11890`)?
Kuidas interpreteerida real `factor(olu)5-12` hinnatud parameetrit (`-1.76034`)?

Veendume ka selles, et saadud uus ja keerukam mudel (kasutame ju prognoosimiseks nii tudengi õlletarbimist kui tema sugu) prognoosib tõestatavalt paremini kui eelnev vaid tudengi õlletarbimist kasutanud mudel:

```
anova(mudel1, mudel2)
```

Võid võrrelda mudelite mudel1 ja mudel2 prognoosivõimet ka AIC-näitajat kasutades.

Ülesanne
Hinda mudel3, mis prognoosib pikkust vaid tunnust sugu kasutades. Mudel3 on samuti lihtsam kui mudel2 – testi, kas nende kahe mudeli (mudel3 vs mudel2) võrdluses saame tõestada, et keerukam mudel prognoosib pikkust paremini kui lihtsam mudel (ainult tudengi sugu kasutav mudel)? Interpreteeri saadud tulemust!

Seni on meie poolt leitud pikkust-prognoosivatest mudelitest parimaks osutunud mudel3 (miks?). Üritame siiski leida veel paremat mudelit – üritame kasutada pikkuse prognoosimisel ka tudengi kaalu.

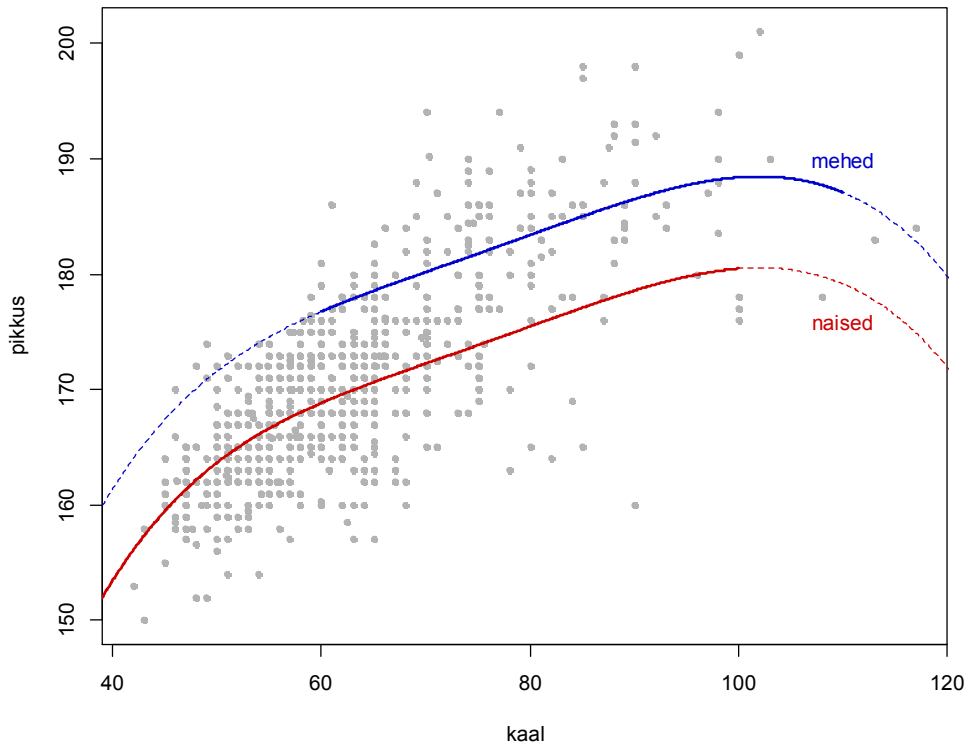
Hinda järgmise mudeli parameetrid (millised tulevad c_0, c_1, c_2 väärtused?):

$$pikkus = c_0 + c_1\, kaal + c_2\, I_{sugu=2} + \varepsilon$$

Mida näitab saadud mudelis vabaliige (c_0)?
Mida näitab saadud mudelis kaalu ees olev kordaja (c_1)?
Mida näitab saadud mudelis sugu=2 mõju (c_2)?

Seos pikkuse ja kaalu vahel ei pruugi olla lineaarne. Proovi lisada mudelisse ka kaalu kõrgemaid astmeid (tuleta meelde kuidas me seda eelmises praktikumis tegime). Kas nende lisamisel läheb mudeli prognoosivõime paremaks?

Kasuta pikkuse prognoosimiseks sugu ja kaalu. Proovi oma mudeli iseloomustamiseks teha ka joonis, näiteks midagi sellist:



Proovi ka järgmiseid käske. Kas mõistad, mida tulemuseks saad (meenuta ka viimast loengut)?

Kui Sinu arvutis pole eelnevalt multcomp-lisamoodulit paigaldatud, siis tee seda:

```
install.packages("multcomp")
```

Loe lisamoodul multcomp R'i sisse

```
library(multcomp)
```

Ja proovi järgmiseid käske:

```
oluF=factor(olu)
mudel=lm(pikkus~oluF)
abi=glht(mudel, linfct = mcp(oluF = "Tukey"))
abi

summary(abi)

confint(abi)

par(mar=c(4,7,3,2))
plot(abi)
```