Biomeetria

5. praktikum

Koosmõjud. Mudeli valik sammregressiooni abil.

Loeme sisse andmestikud:

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

Sisseloetud andmestikest kasutame andmestikku ravi2:

```
head(ravi2)

attach(ravi2)

table(ravi)
table(rs123)
```

Tunnuste selgitused: Inglismaal katsetati juhuslikult valitud patsientidel uudset ravimit. Osadel patsientidel oli üks genotüüp (rs123=A tähistab genotüüpi AA), teistel teine (rs123=G tähistab genotüüpe AG ja GG, mis käituvad sarnaselt – vaatlusalune mutatsioon käitub dominantselt). Märkus: mutatsiooni rs-number ei ole õige! Ravi efektiivsust mõõdeti voodipäevade arvu järgi – kui kaua tuli patsienti enne tervenemist (täpsemalt: kodusele ravile lubamist) haiglas hoida.

Proovime andmestiku põhjal vastata järgmistele küsimustele:

1. Kas ravi osutus Inglismaal kasulikuks?

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

   või

   ```
   m1=lm(voodipaevi~ravi)
   summary(m1)
   ```

Milline on otsus ravi kasulikkuse kohta?

Lisaküsimus tähelepanelikele: milline erinevus on toodud kahe lähenemise (t-test, lineaarne mudel) eeldustes?

Kuu aja jooksul vajab Inglismaal 1000 antud haiguse all kannatavat patsienti haiglaravi. Kui palju voodipäevi aitaks kokku hoida uuele ravile üleminek?

Vaata ka keerukamat mudelit tunnusele *voodipäevad*.

```
m2=lm(voodipaevi~factor(rs123)+factor(ravi)+factor(rs123)*factor(ravi))
summary(m2)
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        10.0600     0.1598  62.973   <2e-16 ***
factor(rs123)G                      0.2614     0.3064   0.853    0.394
factor(ravi)ravim                  -3.9857     0.2267 -17.583   <2e-16 ***
factor(rs123)G:factor(ravi)ravim    5.6470     0.4310  13.103   <2e-16 ***

Residual standard error: 1.957 on 408 degrees of freedom
Multiple R-squared:  0.5673,    Adjusted R-squared:  0.5641
F-statistic: 178.3 on 3 and 408 DF,  p-value: < 2.2e-16
```

Seleta kõigi hinnatud parameetrite tähendust!

Kontrollime ka anova-käsu abil koosmõju statistilist olulisust (kui koosmõjude kirjeldamiseks läheb vaja vaid ühte parameetrit, siis pole selline täiendav test tegelikult vajalik, sest hinnatud koosmõju-parameetri kohta raporteeritav p-väärtus langeb kokku anova-testi poolt tulemuseks tuleva p-väärtusega):

```
m3=lm(voodipaevi~factor(rs123)+factor(ravi))
anova(m2, m3)
```

Leia predict-käsu abil hinnatud keskväärtused kõigi ravi ja genotüübi kombinatsioonidele mõlemt mudelit kasutades.

Mudel 2 (koosmõjudega)

	Kontroll	Ravim
A		
G		

Mudel 3 (ainult peamõjud)

	Kontroll	Ravim
A		
G		

Leia andmestiku põhjal ravimit saanud genotüübiga A patsientide keskmine (kasutades mean-käsku). Kumma mudeli prognoosiga ühtib nähtud keskmine?

Kui täpselt sa ikkagi tead ravi saanud genotüübiga A inimeste keskväärtust?
Leia usaldusintervall antud patsientide grupi keskväärtusele nii klassikalisel viisil (valemiga $\bar{x} \pm t_{df=n-1;\,\alpha/2} \cdot s/\sqrt{n}$):

```
t.test(voodipaevi[ravi=="ravim" & rs123=="A"])
```

Saadud usaldusintervall ravi saanud genotüübiga A inimeste voodipäevade arvu keskväärtusele:

…… - ……..

kui ka hinnatud mudelit `m2` ja predict-käsku kasutades: …… - ……..

Võrdle saadud usaldusintervalle. Milles seisneb erinevus?

Muuseas, 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?

Viimasele küsimusele vastamisel võib olla kasu ka sellest, kui suudad ära mõistatada, mida (sisuliselt) hindavad järgmised käsud:

```
# install.packages("gmodels") - anna käsk ainult vajadusel (kui järgnevad
#                                 käsud ei tööta)

library(gmodels)

estimable(m2, c(1, 0, 0, 0) )


estimable(m2, c(1, 0, 1, 0) )


estimable(m2, c(1, 0, 0, 0)-c(1, 0, 1, 0) )


estimable(m2, c(1, 1, 0, 0)-c(1, 1, 1, 1) )


estimable(m2, 0.15*(c(1, 0, 0, 0)-c(1, 0, 1, 0))+0.85*(c(1, 1, 0, 0)-c(1, 1, 1, 1)))


estimable(m2, c(0, 0, -1, -0.85))


estimable(m2, c(0, 0, -1, -0.85), conf.int=0.95)


estimable(m2, 1000*c(0, 0, -1, -0.85), conf.int=0.95)
```

Koosmõju pideva tunnuse ja faktortunnuse vahel

Loeme sisse tudengite andmestiku:

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

ja üritame prognoosida tudengi kaalu tema soo ja pikkuse järgi:

```
m4=lm(kaal~pikkus+factor(sugu))
```

Märka, et tunnust pikkus me factor-käsuga ei ümbritse. Faktortunnuste puhul käsitleme iga konkreetse faktortunnuse väärtust kandvat inimeste gruppi eraldi, hindame selle konkreetse grupi eripära jne. Kui uuritav tunnus on pidev või paljude võimalike väärtustega diskreetne tunnus, siis iga pikkusega inimesi eraldi käsitleda on liiga vaevarikas – tuleks hinanta väga palju parameetreid (näiteks kuidas erinevad 156,6 cm pikkused inimesed võrdlusnivoost jne..). Kuna igasse gruppi sattuks vähe inimesi (andmestikus on vaid üks 156,6 cm pikkune inimene), siis oleksid leitud parameetrite hinnangud väga ebatäpsed. Lisaks ei saaks pikkust faktortunnusena käsitledes leida ka kaalu prognoosi näiteks 156,5cm pikkusele inimesele – sest täpselt sellise pikkusega inimest valimis ei olnud ja seega ei saa pikkust faktortunnusena käsitlev mudel hinnata, kui palju 156,5cm pikkused inimeste keskmine kaal erineb võrdlusnivoo keskmisest kaalust. Lahenduseks on pikkuse käsitlemine pideva tunnusena – iga täiendav ühik pikkust suurendab veidike ka kaalu prognoosi. Vajadusel tuleks mudelisse muidugi lisada ka pikkuse ruutliige jne.

Hinda ka mudel m5 mis sisaldaks pikkuse ja soo koosmõju. Kas pikkuse ees olev regressioonikordaja on meestel/naistel tõestatavalt erinev? Milline on regressioonsirge tõus meestel, milline naistel?
Kas pikkuse ja soo koosmõju sisaldav mudel muutub statistiliselt oluliselt paremaks, kui üritame mudelisse lisada ka pikkuse ruutu?

Iseloomustagem hinnatud mudelit ka graafiliselt:

```
plot(pikkus, kaal, col=c("red2", "skyblue")[sugu], pch=20)

xx=seq(140, 210, length=500)

# Meeste kaalude prognoosid erinevate pikkuste korral
y_mees=predict(m5, data.frame(pikkus=xx, sugu=2))

# Naiste kaalude prognoosid erinevate pikkuste korral
y_naine=predict(m5, data.frame(pikkus=xx, sugu=1))

# Kanname prognoosid graafikule:
lines(xx, y_naine, col="red3", lwd=2)
lines(xx, y_mees, col="blue", lwd=2)
```

Mudeli valikust – sammregressioon prognoosiva mudeli valikuks.

Parima prognoosiva mudeli otsimisel saab R meid teataval määral ka aidata. Esmalt tuleb koostada võimalikult rikas mudel, mis peaks sisaldama kõikmõeldavaid effekte ja mõjusid (midagi täiendavat R mudelile lisada ei proovi):

```
mudel_suur=lm(pikkus~kaal+factor(sugu)+kaal*factor(sugu)+
factor(sport)+factor(sport)*kaal+I(kaal^2)+
factor(olu)+factor(suitsetamine)+vanus+factor(suitsetamine)*vanus+
vanus*factor(sugu))
```

Seejärel võime R-l lasta AIC-väärtuse järgi valida välja parima prognoosiva mudeli:

```
valitud_mudel = step(mudel_suur)
summary(valitud_mudel)
```

Millised tunnused jäid mudelisse sisse ja milliseid tunnuseid ei peetud tudengi pikkuse prognoosimisel vajalikeks?

Leia ühe tudengi (iseenda või sõbra) pikkusele prognoos kasutades leitud mudelit valitud_mudel ja leia ka vastav prognoosiintervall. Kas prognoositava tudengi tegelik pikkus jäi etteantud vahemikku?

Muuseas, tunnus suitsetamine on kodeeritud järgnevalt:

1 - ei ole kunagi suitsetanud
2 - pean vahet või olen suitsetamisest loobunud
3 - suitsetan harvemini kui üks kord nädalas
4 - suitsetan mitmeid kordi nädalas, kuid mitte iga päev
5 - suitsetan kõige rohkem 9 sigaretti/sigareid/piibutubakat päevas
6 - suitsetan päevas 10-19 sigaretti/sigareid/piibutubakat
7 - suitsetan päevas vähemalt 20 sigaretti (või sama palju sigareid või piibutubakat)

Mudeli valikust 2

parim prognoosiv mudel ei pruugi olla tunnuse väärtuste tekkimist kirjeldav mudel!

Teeme järgmise eksperimendi. Võtame mingi suurusega valimi ja hindame antud valimi põhjal parima *Y*-tunnuse väärtuseid prognoosiva mudeli ja nn õige mudeli (*õige mudel* siin: tunnuse väärtuste kujunemist korrektselt kirjeldava mudeli).Vaatame, milline neist mudelitest tegelikult tulevasi vaatluseid täpsemalt prognoosib!

Loeme sisse abifunktsiooni mis meie eest eksperimenti teeb ja andmeid kogub:

```
source(url("http://www.ms.ut.ee/mart/biomeetria2015/eksperiment.R"))
```

Peale ülaltoodud käsu andmist R-le tekib R-i juurde uus funktsioon nimega v6tavalim.

Võtamegi selle uue funktsiooni abil valimi, uurime 100-t inimest (või uurimisobjekti):

```
andmestik=v6tavalim(100)
```

Saadud andmestikus on 6 tunnust. Laseme esmalt R-l valida välja parima prognoosiva mudeli (AIC-väärtuse järgi):

```
mudel_suur=lm(Y~X+X2+X3+X4+X5, data=andmestik)
mudel_valitud=step(mudel_suur)
summary(mudel_valitud)
```

Milline mudel osutus valituks? Pane valitud mudel ka kirja:

Y = …… + ε

Tegelikult on *Y*-tunnuse väärtuseid genereeritud järgmise valemi abil:

$$Y = c_0 + c_1 X + c_2 \cdot X_3 + c_3 \cdot X_5 + \varepsilon$$

Hinda ka ülaltoodud tegelik (nn põhjuslik või struktuurne) mudel nimega mudel_oige. Millised hinnangud saad kordajatele c0, c1, c2 ja c3?

c_0= …… c_1= …… c_2= …… c_3= ……

Võrdle nn õige mudeli ja R'i poolt valitud parima prognoosiva mudeli (vale mudeli) AIC-väärtuseid. Milline mudel peaks tulevasi vaatluseid paremini prognoosima otsustades AIC-väärtuse põhjal?

Kontrollime, kas vale mudel tõepoolest suudab täpsemaid prognoose leida.

Teeme 1000 uut vaatlust:

```
uued=v6tavalim(1000)
```

ja võrdleme uute vaatluste Y-tunnusele leitud prognooside täpsust. Teeme võrdluse kahel viisil. Esmalt leiame mõlema mudeli jaoks prognoosivigade ruutude keskmise (nn keskmine ruutviga, mean squared error ehk MSE):

```
# uute vaatluste y-tunnusele vale mudeli abil leitud prognoosid
prognoos1=predict(mudel_valitud, uued)

# uute vaatluste y-tunnusele õige mudeli abil leitud prognoosid
prognoos2=predict(mudel_oige, uued)

# Keskmine ruutviga vale mudeli (aga AIC-järgi parem mudel) jaoks:
mean((uued$Y-prognoos1)**2)

# Keskmine ruutviga õige mudeli jaoks:
mean((uued$Y-prognoos2)**2)
```

Vaatame leitud prognoosivigasid ka graafiliselt:

```
par(mfrow=c(2,1))
hist(uued$Y-prognoos1, xlim=c(-3,3), main="vale mudel (AIC järgi parem)")
hist(uued$Y-prognoos2, xlim=c(-3,3), main="õige mudel")
```

Kumma mudeli korral on prognoosivead nullile lähedamal (kumb mudel prognoosib täpsemalt)?