

Osa I. Osamudeli hindamisest

Geneetikas tuleb ette olukordi, kus sama y -tunnust (fenotüüpi) kasutades tuleb hinnata miljoneid erinevaid regressioonmudeleid, kusjuures hinnatavad mudelid erinevad vaid ühe sõltumatu tunnuse poolest (kontrollitakse erinevate SNP-de ehk ühenukleotiidi pikkuste mutatsioonide mõju). Lisaks uurimiselusele sõltumatule tunnusele – snipile – on sageli mudelis veel teisi tunnuseid, nagu inimese *vanus*, *sugu*, *kehamassiindeks (KMI)* jms. Kuna hinnatavaid mudeleid on palju, siis üritatakse töö mahtu võimalikult väikeseks teha. Sageli kasutatakse järgmist strateegiat, kus y -tunnusest eemaldatakse nende alati samaks jäävate tunnuste (*vanus*, *sugu*, *KMI* jms) mõju – leitakse y -tunnuse prognoosimisel nende tunnuste abil tekkivad jäägid ja kasutatakse siis saadud jääke kontrollimaks, kas ühel või teisel SNP-l on mõju uuritavale tunnusele. Kontrollime, kuivõrd hästi võiks selline strateegia (eemaldame eelnevalt *KMI* mõju y -tunnusele, ja seejärel hindame geneetilise markeri mõju korrigeeritud y -tunnusele) matkida korrektse suure mudeli (kus on korraka sees nii *vanus* kui geneetiline marker) tulemust.

Näidisandmete genereerimine:

```
n=200
gen_marker = rbinom(n, 2, 0.3)
KMI = 10+5*gen_marker + rnorm(n)
y = 2*gen_marker - 2*KMI + rnorm(n)

# Hinnatud õige mudel:
mudel=lm(y~gen_marker+KMI)
summary(mudel)

# Sageli kasutatav lihtsustatud analüüs,
# kus y-tunnusest on eemaldatakse vanuse mõju:

y_jaak=residuals(lm(y~KMI))
m1 = lm(y_jaak~gen_marker)
summary(m1)
```

Kas lihtsustatud analüüs annab geneetilise markeri mõjule sarnase hinnangu võrreldes õige, suure mudeliga?

Hinda mudel või mudelid mis ei kasuta tunnust *gen_marker* sõltumatu tunnuseks (nn x -tunnuseks). Nende mudelite abil saadud tulemusi kasuta teise mudeli hindamiseks, kus mudelis kasutatakse vaid ühte sõltumatut tunnust. Vali see tunnus teise mudelisse selliselt, et tema ees oleva kordaja hinnang tuleks täpselt samasugune, nagu tunnuse *gen_marker* kordaja suures kahte sõltumatut tunnust sisaldavas regressioonanalüüsi mudelis (mudel *mudel*).

Proovi leida ka võimalust saada lihtsamast regressioonmudelist õige hinnang jääkide dispersioonile (kuidas jõuda samasuguse jääkide hajuvuse hinnanguni nagu saadi algsest suurest mudelist)....

Osa II

... et lihtsad kuid tähtsad asjad ei ununeks

Uuringu taustast

Soovitakse uurida, kas paljude preilidega sahmerdamisel on ka negatiivseid külgi. Bioloogidel on pikka aega püsti olnud hüpotees, et paljude järglaste andmine kurnab ära organismi ja viib varase surmani. Kuna loomulikes tingimustes saavad palju järglaseid eelkõige terved ja tugevad isendid, ei pruugi järglaste kahjulik mõju vaatlusandmetes näha olla. On vaja (randomiseeritud) katset.

Samastame end nüüd hetkeks suurte teadlastega ja liigume umbes aastasse 1980. Antud hetkeks on juba näidatud (putukate peal), et ohtra järglaskonna produtseerimine vähendab emaste isendite eluiga. Soovime näidata, et ka mitmekordne isaks saamine mõjub laastavalt isaste loomakeste elueale. Sestap teeme järgmise katse. Hoiame isaseid äädikakärbseid pisikestes purgikestes. Osadesse purkidesse me emaseid kärbeid ei lase ja vaatame siis, kui kaua meie isane kärbekeene seal purgis vastu peab. Teistesse purkidesse laseme iga paari päeva tagant sisse neitsist äädikakärbe. Kolmandates purkides elavate isaste juurde laseme aga iga paari päeva tagant 8 neitsit.

Paraku varitseb oht, et purkides, kuhu lasti palju äädikakärbeid, võib loomakeste eluiga väheneda ka muudel põhjustel: konkurents söögi pärast, üleasustatusest tingitud stress vms. Sestap laseme veel osadesse purkidesse iga paari päeva tagant selliseid emaseid äädikakärbeid, kes pole seksist huvitatud (rasedaid). Seega on katses kokku viis gruppi: 25 äädikakärbest kes elasid oma elu ilma emast isendit nägemata, 25 isendit kes pidid iga paari päeva tagant 1 neitsiga tegelema, 25 isendit kes pidid iga päeva tagant 8 neitsiga vaeva nägema, 25 isendit kes pidid iga paari päeva tagant koos elama 8 raseda äädikakärbe ja 24 isendit, kes pidid iga paari päeva tagant oma purki jagama 1 raseda emasega.

Ülesande üldine kirjeldus: Leia, kas järglaste produtseerimine mõjub halvasti isaste elueale!

Andmestiku kirjeldus

Andmestik pärineb ajakirjas Nature 1981. a. avaldatud artiklist "Sexual Activity and the Lifespan of Male Fruitflies" (Nature 294, 580-581), täpsemalt on ta saadud artiklis avaldatud joonise 2 hoolikal inspekteerimisel. Andmestiku on originaalartiklist välja peilinud Hanley ja Shapiro, vaata ka nende artiklit Hanley, J. A., and Shapiro, S. H. (1994), "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention," *Journal of Statistics Education*, Volume 2, Number 1.

Andmestikus on järgmised tunnused:

- ID* – äädikakärbe isikukood;
- pruute* – mitu emast äädikakärbest antud isendiga kohtamas käis (väärtused: 0, 1, 8)
- ptyyp* – pruudi tüüp: 0 – rase; 1- neitsi; 9- pruute polnudki
- eluiga* – meie isase äädikakärbe eluiga (päevades)
- rind* – rindmiku pikkus (millimeetrites).

Juhised

Loe sisse andmestik karbes.dat:

```
andmed=read.table(url("http://www.ms.ut.ee/mart/linmud2018/karbes.dat"),
  header=TRUE)
andmed[1:2,]
attach(andmed)
```

Vaata andmeid:

```
# Joonis 1 -----
boxplot(eluiga~paste(pruute,ptyyp),main="Eluiga",
  names=c("üks", "1 rase", "1 neitsi", "8 rasedat", "8 neitsit"),
  col="grey90",cex.axis=.9, xlab="Kaaslased")

# Joonis 2 -----

library(beeswarm)
varv=ptyyp+1
varv[ptyyp==9]=3

par(mar=c(4.4,4.2,2,3.5))
beeswarm(eluiga~pruute, pch=21, pwb=c("red2","powderblue","gray50")[varv] )
legend(3.4, 95, c("üks", "rase", "neitsi"),
  pch=21, pt.bg=c("gray50", "red2","powderblue"),
  xpd=NA, bg="white")
```

Küsimus:

Kas paljude järglaste soetamine lühendab eluiga? Tee sobiv test/võrdlus! Millise p-väärtuse saad?