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 uurimisalusele sõltumatule tunnusele – snipile – on sageli mudelis veel teisi tunnuseid, nagu inimese *vanus*, *sugu, kehamassiindeks* (*KMI*) jms. Kuna hinnatavaid mudeleid on palju (sageli kontrollitakse kümnete miljonite SNP-de mõju, seega siis tuleb hinnata ka kümneid miljoneid regressioonmudeleid), siis üritatakse töö mahtu võimalikult väikeseks teha. Vahel kasutatakse järgmist strateegiat: *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 korraga sees nii KMI kui geneetiline marker) tulemust.

Näidisandmete genereerimine:

```
set.seed(1)
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 eemaldatud KMI mõju:

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

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

Mis juhtuks tulemustega, kui geneetilise markeri mõjud on väga väikesed? Muuda geneetilise markeri mõju nii kehamassiindeksile kui y-tunnusele 100 korda väiksemaks. Sedavõrd väikeste mõjude tuvastamiseks on muidugi tarvis kasutada ka suuremat valimit, seega suurenda valimi suurust 1000 korda, ehk tee eelnenud programmis järgmised parandused:

```
n=200                                -> n=200*1000
KMI = 10+5*gen_marker + rnorm(n)     -> KMI=10+5/100*gen_marker + rnorm(n)
y = 2*gen_marker - 2*KMI + rnorm(n) -> y=2/100*gen_marker - 2*KMI + rnorm(n)
```

Kas lihtsustatud mudel m1 annab nüüd mõistliku hinnangu geneetilise markeri mõjule?

Ülesanne 1

Taasta esialgne olukord – vaatame geneetilist markerit, millel tegelikult ka on (reaalne) mõju nii kehamassiindeksile kui ka *y*-tunnusele. Kuidas saaksid hinnata tunnuse *gen_marker* mõju korrektselt ilma tunnust *KMI* kasutamata (tee nii, et mudelis `m3` poleks sees tunnust *KMI*, aga sealt mudelist saaksime siiski kätte geneetilise markeri (kehamassiindeksile kohandatud) mõju *y*-tunnusele:

```
m2 = lm(… ~ … )
m3 = lm(y_jaak ~ … )
summary(m3)
```

Kas oskad arvata, miks selline lähenemine praktikas mutatsioonide mõju uurivatele inimestele ei meeldi?

Ehk märkad ka, et mudelist `m3` saame küll õige hinnangu geneetilise markeri mõjule, kuid teised näitajad (p-väärtused, standardvead jne) on õige natuke erinevad. Mudelite `m3` ja `mudel` jääkide ruutude summad on täpselt samad:

```
sum(residuals(m3)**2)
sum(residuals(mudel)**2)
```

kuid standardvead jms tulevad (veidi) erinevad:

```
summary(mudel)$sigma
summary(m3)$sigma
```

Miks? Kuidas saaksid näiteks leida jääkide standardveale korrektset hinnangut mudelit `m3` kasutades?

Ühe katse planeerimisest

Eksperimentaator Ann soovis uurida kahe tunnuse (x_1 ja x_2) mõju *y*-tunnusele. Ta tekitas endale x_1 ja x_2 väärtuste tabeli ja kavatses siis iga tekitatud x_1 ja x_2 väärtuse korral läbi viia eksperimendi ja mõõta *y*-tunnuse väärtust. Tunnuste x_1 ja x_2 väärtused tekitas ta nii:

```
x1=rep(1:4, each=30)
x2=rep(1:3, each=40)
```

Enne eksperimendi algust rääkis ta kavandatavast eksperimendist sõbrast statistikule Bertale. Sõbrast statistik tegi aga järgmise arvutuse:

```
y=rep(1:2, 60)
library(car)      # sisaldab funktsiooni vif
vif(lm(y~x1+x2))
```

ja väitis siis, et kavandatud katseplaan pole just kõige parem. Ta arvas, et sama arvu katsetega saaks x_1 ja x_2 mõju hinnata vähemalt 6 korda väiksema dispersiooniga – kui kasutada mõistlikumat katseplaani.

Ülesanne 2
Statistik Berta kasutas oma arvutustes väga imelikku *y*-tunnust. Tegelikud katsest saadavad *y*-tunnuse väärtused on täiesti teistsugused (ootuspärased *y*-tunnuse väärtused võiksid hoopis olla 20 ringis). Kas Berta poolt arvutatud VIF-i väärtused võiksid ikkagi olla ligikaudu õiged või mitte? Põhjenda oma arvamust!

Kontrollime!
Proovime järgi- kas mõistlikum katseplaan tõepoolest suurendaks hinnangu täpsust 6 korda (st. vähendaks hinnangu dispersiooni 6 korda).

Genereerime eksperimendi tulemused esialgset katseplaani kasutades:

```
set.seed(1)
x1=rep(1:4, each=30)
x2=rep(1:3, each=40)
y=17+2*x1-x2+rnorm(length(x1))
m1=lm(y~x1+x2)
summary(m1)
```

Genereeri eksperimendi tulemused paremat katseplaani kasutades:

```
set.seed(1)
x1=rep(1:4, each=30)
x2=…
y=17+2*x1-x2+rnorm(length(x1))
m2=lm(y~x1+x2)
summary(m2)
```

Võrdle x_1 mõju hinnangu dispersioone mudeli `m1` ja mudeli `m2` korral. Mitmekordset erinevust näed?

Kuidas saaksid näiteks x_2 mõju hinnangut veelgi täpsemaks muuta (ilma eksperimentide arvu muutmata)?

Katsetus 3

Genereerime taas endale vaatlusandmed (eeldame seekord, et tegemist on tõesti mõõtmistulemustega, mitte meie poolt kontrollitud eksperimendiga, st x1, x2 ja x3 väärtuseid ei määra meie vaid tegemist on juhuslikult valitud inimese peal mõõdetud tunnuste väärtustega):

```
set.seed(1)
n=200
x1=rnorm(n); x2=rnorm(n); x3=rnorm(n)+4*x2
y=10+1*x1+1*x2+1*x3+rnorm(n)
mudel2=lm(y~x1+x2+x3)
summary(mudel2)
vif(mudel2)
```

Osade tunnuste korral on dispersiooni puhitusteguri väärtused suured. Kas näiteks x3 eemaldamine mudelist aitab teiste tunnuste hinnanguid täpsemaks teha?