

Lineaarsed mudelid

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 uurimiselule 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, kui võrd 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:

```
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)
```

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 (mudelis *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

Hinnangute täpsusest.

Vahel tekib regressioonimudeli hindamisel probleeme arvutustäpsusega. Vaatame paari võimalikku tüüpsituatsiooni – ja otsime (või täpsemalt: sina otsid) võimalikku lahendust.

Vaata järgmist miniandmestikku:

x	y	
1	$1e25+1$	$(=1 \times 10^{25}+1)$
2	$1e25+2$	
3	$1e25+3$	

Soovime hinnata regressioonimudeli: $y = c_0 + c_1 x + e$.

Kas oskad aimata, milline peaks tulema regressioonisirge? Proovi ise õige vastus peast arvutades ära arvata!

Hinda regressioonimudel R'is käsuga *lm*.

```
x=1:3
y=c(1e25+1, 1e25+2, 1e25+3)
lm(y~x)
```

Kas said ootuspärase tulemuse?

Kuidas saaks regressiooniparameetreid R-i abiga korrektselt hinnata? Kuidas võiks uuritavat tunnust transformeerida (ja pärast hinnatud parameetreid korrigeerida) nii, et arvuti suudaks jõuda (sinu abiga) õige lahenduseni (kui y_i -i väärtused on kujul $y_i = \text{suur konstant} + \text{midagi}_i$)?

Milline on õige regressioonimudel ja millised hinnangud saab R järgmise andmestiku korral:

x	y
1	$1e25+1$
2	$2e25+1$
3	$3e25+1$

Mida võiks nüüd teha? Kui y_i -i väärtused on kujul $y_i = \text{midagi}_i + \text{väike konstant}$?

Enamasti tekivad probleemid pigem x -tunnustega. Näiteks soovime hinnata x_1 ees olevat kordajat. (Taas – leia peast, milline on tegelik mudel). Kas saame sama kordaja x_1 -le kui lahutaksime kõigist x_1 väärtustest $1e25$? Miks mitte? Kuidas saaksime nüüd leida korrektse tulemuse?

y	x_1	x_2
1011	$1e25+1$	1
1022	$1e25+2$	2
1043	$1e25+3$	4
1085	$1e25+5$	8