

Lineaarsed mudelid. Praktikum

Mudeli valikust I.

Alustame lihtsa genereeritud näitega. Tekitame mõned tunnused, mis on seotud y-tunnusega, teised pole. Hindame ka mõned mudelid y-tunnuse prognoosimiseks.

```
n=1000

set.seed(1)
x1=rnorm(n); x2=rnorm(n); x3=rnorm(n); x4=rnorm(n)
x5=rnorm(n); x6=rnorm(n); x7=rnorm(n); x8=rnorm(n)

y=10+1*x1+0.2*x2+0.08*x5+0.02*x6+0.01*x7+rnorm(n)

m1=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
m2=lm(y~x1+x2+x5+x6+x7)
m3=lm(y~x1+x2+x5+x6)
m4=lm(y~x1+x2+x5)
m5=lm(y~x1+x2)
m6=lm(y~x1)
m7=lm(y~1)
```

Valime parima mudeli välja mitmel erineval viisil.

- Kindlasti õige mudel (oletame, et teame – kõige rikkam mudel m1 on igal juhul õige).
- Mallows'i Cp kriteeriumi alusel
Selleks parandame esmalt R'i anova-käsu Mallows'i Cp statistiku arvutusvalemi

```
source(url("http://www.ms.ut.ee/mart/linmud2017/uusANOVA.R"))
```

ja laseme seejärel arvutada kõigi meid huvitavate mudelite jaoks Mallows'i Cp statistiku väärtuse:

```
anova(m1,m2,m3,m4,m5,m6, test="Cp")
```

Milline mudelistest annab hinnanguliselt kõige täpsemaid prognoose?

Mallows'i Cp kriteeriumi järgi parima mudeli otsimiseks on võimalik kasutada ka lisamoodulit `leaps`:

```
tulemus=leaps(x=cbind(x1, x2, x3, x4, x5, x6, x7, x8), y=y,
              method="Cp")
tulemus
```

Täpseima mudeli Cp-väärtus:
`min(tulemus$Cp)`

Millised tunnused peaksid olema mudelis sees:
`tulemus$which[tulemus$Cp==min(tulemus$Cp)]`

- Millise mudelini jõuad, kui viskad mudelist välja statistiliselt mitteolulisi tunnuseid (kasutades olulisuse nivood 0,05)?
- Milline on tegelikult kõige lihtsam õige mudel?

Vaatame nüüd, milline mudelistest suudab uusi vaatluseid tegelikult kõige täpsemalt prognoosida. Genereeri uued vaatlused ja täida alljärgnev tabel:

Uute vaatluste genereerimine:

```
n=10000
uued=data.frame(x1=rnorm(n), x2=rnorm(n), x3=rnorm(n), x4=rnorm(n),
                x5=rnorm(n), x6=rnorm(n), x7=rnorm(n), x8=rnorm(n))

uusy = with(uued, 10+1*x1+0.2*x2+0.08*x5+0.02*x6+0.01*x7+rnorm(n) )
```

		Keskmine ruutviga (uued vaatlused)
Kindlasti õige mudel:	m1.....	
Tegelikult õige mudel:		
Mallows'i Cp kriteeriumi arvates parim mudel:		
Statistilise olulisuse abil valitud mudel:		

Näide 2

Vaatame nüüd olukorda, kus tegelik mudel on lihtsake kuid vaatluseid on kasutada palju:

```
set.seed(1)
n=10000
x1=rnorm(n); x2=rnorm(n); x3=rnorm(n); x4=rnorm(n); x5=rnorm(n)
x6=rnorm(n); x7=rnorm(n); x8=rnorm(n); x9=rnorm(n); x10=rnorm(n)

y=10+1*x1+0.2*x2+rnorm(n)
```

Otsime parimat mudelit nii Mallows'i Cp kriteeriumi abil kui ka kasutades sammregressiooni ja AIC ning BIC-kriteeriume:

```
m1=lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10)
# Mallowsi Cp järgi otsimine kirjuta ise!

# sammregressioon + AIC
m2=step(m1)

# sammregressioon + BIC
m3=step(m1, k=log(n))
```

Vaata millised tunnused valis välja Mallows'i Cp, AIC, millised BIC. Milline meetoditest jõudis tõelisele mudelile (tõele) lähemale? Kommenteeri tulemust!