

Lineaarsed mudelid. Praktikum

Mudeli valikust.

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)

f=function(x1,x2,x3,x4,x5,x6,x7,x8){
  return(10+1*x1+0.2*x2+0.08*x5+0.02*x6+0.01*x7)
}

y=f(x1,x2,x3,x4,x5,x6,x7,x8)+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 C_p kriteeriumi alusel.

Selleks võid näiteks kasutada anova-käsku.

NB! R-i sisseehitatud anova-käsk raporteerib vaikimisi Mallowsi C_p statistiku asemel tema veidralt transformeeritud väärtust, (raporteeritav number pole klassikaline Mallows'i C_p statistik), kuid mudelite paremusjärjestus jääb samaks. Soovi korral võid parandada anova-käsku selliselt, et ta raporteeriks korrektselt arvutatud (traditsioonilisel viisil leitud) Mallows'i C_p statistiku väärtust (mis on kooskõlas teiste R-i pakettide poolt väljastatavaga ja teiste tarkvarade poolt raporteeritavate väärtustega). Selleks anna käsk:

```
source(url("http://www-1.ms.ut.ee/mart/linmud2025/uusANOVA.R"))
```

Leiame seejärel kõigi meid huvitavate mudelite jaoks Mallows'i C_p statistiku väärtused:

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

Milline mudelistest annab hinnanguliselt kõige täpsemaid prognoose?

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

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

Täpseima модели Cp-väärtus:

```
min(tulemus$Cp)
```

Millised tunnused peaksid olema mudelis sees:

```
tulemus$which[tulemus$Cp==min(tulemus$Cp)]
```

Väärtus TRUE näitab, et vastav tunnus on (parimas) mudelis sees ja FALSE tähistab mudelist väljajäänud tunnust. Antud juhul peaksid siis parimas mudelis olema sees 1. 2. ja 5. tunnus.

- c) Võrdle Mallows'i Cp kriteeriumi alusel saadud parimat mudelit Akaike kriteeriumi (AIC) alusel saadud parima mudeliga:

```
AIC(m1,m2,m3,m4,m5,m6)
```

- d) Millise mudelini jõuaksid, kui viskaksid suurimast mudelist (m1) välja statistiliselt mitteolulisi tunnuseid (kasutades olulisuse nivood 0,05)?

- e) Milline on tegelikult kõige lihtsam õige mudel?

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

Uute vaatluste genereerimine (täpselt sama arvutuseeskirja kasutades):

```
n=10000
x1u=rnorm(n); x2u=rnorm(n); x3u=rnorm(n); x4u=rnorm(n)
x5u=rnorm(n); x6u=rnorm(n); x7u=rnorm(n); x8u=rnorm(n)

uusy = f(x1u, x2u, x3u, x4u, x5u, x6u, x7u, x8u) +rnorm(n)
```

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

Katse 2

Ühe andmestiku/olukorra puhul võis edu tagada ka lihtsalt erakordne õnn. Teeme teise katse veel.

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

```
source(url("http://www-1.ms.ut.ee/mart/biomeetria2017/eksperiment.R"))
```

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

Võtamegi selle uue funktsiooni abil valimi, uurime 1000-t inimest (või uurimisobjekti):

```
andmestik=v6tavalim(1000)
```

Saadud andmestikus on 6 tunnust. Meid huvitab eelkõige see, kuidas tunnus X võiks mõjutada tunnuse Y väärtust (teame lisaks, et kui mõju eksisteerib, siis X-tunnuse mõju Y-le on lineaarne).

Leia Mallows'i Cp kriteeriumi alusel kõige paremini prognoosiv mudel (mudel_cp):

```
mudel_cp=lm( Y~ ... )
```

Andmed on tegelikult genereeritud järgmist mudelit kasutades:

$$Y = c_0 + c_1 * X + c_2 * X^3 + c_3 * X^5 + \epsilon$$

Võrdle uute vaatluste korral Mallows'i Cp statistiku järgi väljavalitud mudeli prognooside täpsust ja õige mudeli prognooside täpsust!

```
uued=v6tavalim(2000)
```

```
# Leiame prognoosid mõlema mudeliga
prognoos_cp=predict(mudel_cp, uued)
prognoos_6igemudel=predict(mudel_6ige, uued)

# Leiame prognoosivigade ruutude keskmise
# (mida suurem, seda ebatäpsemad prognoosid)
mean((uued$Y-prognoos_cp)**2)
mean((uued$Y-prognoos_6igemudel)**2)
```

Kumb on täpsem? Kas palju või natuke täpsem?