

Lineaarsed mudelid

Eeldused III

Näide 1. Pärissandmed. Kui halb on olla vana?

On hästi teada, et vanematel emadel on suurem tõenäosus saada Downi sündroomiga last (last kellel on kolm 21. kromosoomi). Aga kas ka muude geenimutatsioonide tekketõenäosus on vanematel emadel suurem? Loeme sisse andmestiku, kus on sees ema vanused lapse sündimise ajal ja lapsest leitud uute ühetäheliste geenimutatsioonide arvud (mitu tähte on lapse genoomis teistsugused võrreldes selle osaga tema vanemate genoomist, mis lapsele pärandati). Kas eksisteerib seos mutatsioonide arvu ja ema vanuse vahel?

```
print(load(url("http://www.ms.ut.ee/mart/linmud2018/mutatsioonid.RData")))
attach(mutandid)

mutandid[1:3,]
plot(emavanus, mutatsioone)

m1=lm(mutatsioone~emavanus)
summary(m1)
abline(m1)
```

Kas seos ema vanuse ja mutatsioonide arvu vahel eksisteerib? Milline on hinnanguline seos?

Vaata ka mudeli eelduseid!

```
plot(m1)
```

Kommenteeri!

Vaatame, mida meile soovitab Box-Cox-i transformatsioon!

```
library(MASS)
boxcox(m1)
```

Transformeeri vajadusel uuritavat tunnust, proovi andmeid uuesti analüüsida. Millise järelduseni jõuad?

Näide 2. Lõpuks midagi ilusat ja selget. Genereeritud andmed.

Genereerime ühe andmestiku, kus vead on pigem multiplikatiivsed (juhuslikud kõrvalekalded keskmisest – teiste, mõõtmata jäänud tunnuste mõjud - on pigem protsentuaalsed, mitte absoluutsed). Vaatame milliseid tulemusi me sellise andmestiku analüüsimisel võiksime näha.

```
# Andmete genereerimine

x=runif(200, 5, 20)
f=function(x){
  6*sin(x)+8*x-0.25*x**2
}

set.seed(1)
y=f(x)*exp(rnorm(length(x), sd=0.3))
```

Esimene mudel:

```
m1=lm(y~poly(x, 5))
plot(m1)
```

Mida oskad nendelt graafikutelt välja lugeda?

Vaata, millist transformatsiooni soovitab Box-Cox'i meetod?

```
boxcox(m1)
```

Tee joonis esialgse mudeli jaoks ja lisa samale joonisele prognoos/prognoosiintervall mille saad, kui võtad kuulda Boxi ja Coxi soovitusi!

```
plot(x,y)
xx=seq(4,21, length=1000)
lines(xx, f(xx), col=2, lwd=2)
m1=lm(y~poly(x, 5))
summary(m1)
y1=predict(m1, data.frame(x=xx), interval="prediction")
lines(xx, y1[,1], col="blue")
lines(xx, y1[,2], col="blue", lty=2)
lines(xx, y1[,3], col="blue", lty=2)
```

Kui vaatluste hajuvus pole konstantne, siis meeldib inimestel sageli kasutada üldistatud lineaarseid mudeleid. Võrdle kahe järgneva üldistatud lineaarse mudeli prognoose enda poolt saadud tulemustega. Kommenteeri!

```
m3=glm(y~poly(x, 5), family=gaussian(link="log"))
y3=predict(m3, data.frame(x=xx), type="response")
lines(xx, y3, col="orange", lwd=2)

m3a=glm(y~poly(x, 5), family=quasi(link="log", variance="mu^2"))
y3a=predict(m3a, data.frame(x=xx), type="response")
lines(xx, y3a, col="green3", lwd=2)
```