

Lineaarsed mudelid

Eeldustest I

Näide jääkidest pärisandmestikus.

Kasutatud on dr. Marika Tammari andmeid (reuma) ja selgitusi andmetele.

Näide sellest, kuidas üks erind paljut muuta võib ja ka sellest, mida sellise erindiga peale hakata.

Andmed leiad:

```
load(url("http://www.ms.ut.ee/mart/linmud2018/reuma.RData"))
```

Muuda attach käsu abil andmestiku tunnused lihtsamini kasutatavaks.

Seetsinased andmed on kogutud 50-lt reumatoidartriidi (RA) haigelt kahe järjestikuse visiidi käigus. Uuringu eesmärgiks oli hinnata RA-haigete elukvaliteediküsimustiku usaldusväärsust, kuid kogutud andmed võimaldavad uurida nii mõndagi muud põnevat.

Näiteks: kas haiguse ägeduse kirjeldamisel saab piirduda vaid settereaktsiooni (SR) kiiruse ära märkimisega või tuleb SR kiirus ja C-reaktiivse valgu (CRP) väärtus mõlemad ära nimetada. Eesti arstide seas üsna levinud arvamuse kohaselt mõõdavad nad ühte ja sedasama (põletiku esinemist) ja sestap piisab vaid ühe neist määramisest. Vaatame, kui hästi on üks nendest põletikunäitajatest teise kaudu prognoositav.

Andmetes:

SR1 – SR kiirus mm/h

CRP1 – CRP (C-reaktiivne valk) sisaldus veres mg/l.

Alusta lineaarse seose uurimisest, seda käskude abil:

```
m1=lm(SR1~CRP1)
```

```
summary(m1)
```

Kuidas kommenteerid mudeli sobivust (determinatsioonikordaja, olulisustõenäosus)?

Joonista hajuvusgraafik ja kanna sellele mudeliga kirjeldatud regressioonisirge

```
plot(CRP1, SR1)
```

```
x=seq(0,70)
```

```
y=predict(m1, data.frame(CRP1=x))
```

```
lines(x,y)
```

Uuri, mis muutub, kui mudelisse lisada ruutliige

```
m2=lm(SR1~CRP1+I(CRP1^2))
```

```
summary(m2)
```

Mida ütleb determinatsioonikordaja?

Kanna mudeli m2 poolt kirjeldatud regressioonijoon hajuvusgraafikule.
Mis torkab silma?

Uuri, mis lisainformatsiooni annab mudelite eelduste hindamine. Kasuta kāske:

```
par(mfrow=c(2,2))
plot(m1,1, main="Mudel 1"); plot(m1,5, main="Mudel 1");
plot(m2,1, main="Mudel 2"); plot(m2,5, main="Mudel 2");
```

Vaatlus nr 40 on eripärane. Mis võiks olla selle eripära põhjuseks?

Viska mudelist m1 välja vaatlus nr 40 ja vaata, mis saab

```
m2a=lm(SR1~CRP1+I(CRP1**2), data=reuma[-40,])
summary(m2a)
```

Vaata, kuidas mõjus erindi väljajätmine regressioonisirgele

```
plot(CRP1, SR1)
y=predict(m2, data.frame(CRP1=x))
lines(x,y, col=2, lty=2)
y=predict(m2a, data.frame(CRP1=x))
lines(x,y, col=2)
```

Vaata, kas ruutliige on jätkuvalt vajalik?

Proovi ka lihtsamat, ilma ruutliikmeta, mudelit:

```
m1a=lm(SR1~CRP1, data=reuma[-40,])
summary(m1a)
```

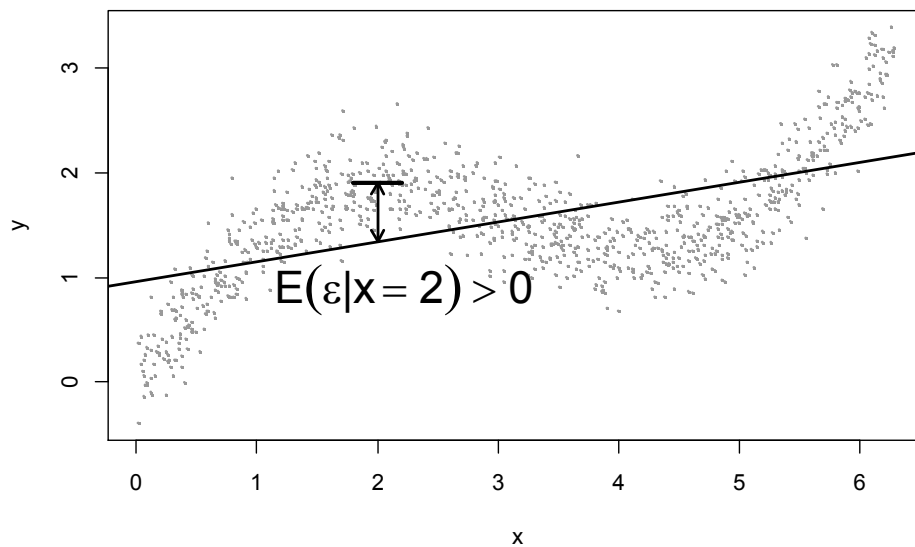
Võrdle mudelite m2 ja m2a determinatsioonikordajaid.

Kuidas liikuda edasi? Mida järgmisena teha? Millise mudeli kasuks otsustad lõpuks sina?

Lisatähelepanek: võrdle hinnatud mudelite m2 ja m2a (40. vaatlus eemaldatud) parameetreid. Vaatlusele nr. 40 vastav Cook'i kaugus oli ligikaudu 1. Mida ütles Cook'i kaugus parameetrite hinnangute muutuse kohta? Kas mudelite m2 ja m2a hinnangute erinevus on enam-vähem ootuspärane?

$$E \varepsilon = 0 \quad \text{ehk} \quad EY = X\beta$$

Lineaarse mudeli oluliseks eelduseks on nõue, et jääkide keskvärtus on null. Antud eeldus on rikutud näiteks siis, kui valim pole esinduslik, või kui kasutatud mudel on vale. Vaata näiteks lisatud võimalikku probleemi illustreerivat joonist:



Esimese ettekujutuse mudeli headusest annab sageli hajuvusgraafik. Hajuvusgraafiku pealt märkame näiteks suhteliselt kergesti, kas tegelik seos on lineaarne või mitte. Kui aga kasutame keerukamat mudelit, siis ei pruugi olla kuigi kerge märgata hajuvusgraafikul oma mudeli puudujääke. Hoopis paremaks alternatiiviks võib osutuda graafik, kus vaatleme mudeli jääke vs x -tunnuse väärtuseid. Vaata ja otsusta ise!

Tekitame näidisandmed (milline on tegelik mudel?):

```
set.seed(2)
x1=runif(200, 1, 10)
y=2+3*x1+0.8*x1*x1+9*sin(x1)+rnorm(200, sd=7)
```

Hindame (vigase) mudeli, vaatleme hajuvusgraafikut:

```
m1=lm(y~x1+I(x1**2))
# Graafik 1
plot(x1,y)
xx=seq(-1,11, length=1000)
lines(xx, predict(m1, data.frame(x1=xx)))
```

Kas märkad, et kasutame vale mudelit?

Võrdluseks joonista graafik x-tunnus vs jäägid:

```
# Graafik 2
plot(x1, residuals(m1))
abline(h=0)
```

Kumma graafiku pealt märkad kergemini mudeli ebatäiuslikkust?

Sageli eelistatakse hoopis prognoos vs jääk graafikut:

```
# Graafik 3
plot(predict(m1), residuals(m1))
abline(h=0)
```

Kui meie hädine mudel prognoosib y-tunnuse väärtuseks 70, siis milline võiks prognoos olla tegelikult?

Võrdle ka:

```
plot(m1, 1)
```

Paranda mudel vastuvõetavaks – ilma mudeli tegelikku kuju piilumata – ja vaata siis jääkide graafikuid! Kas tunduvad nüüd vastuvõetavad? Kanna hajuvusgraafikule nii oma mudeli prognoosid kui ka tegelik seos – kas jõudsid tõe lähedale?

Ülesanne

Loe sisse andmestik ja uuri hinnatud mudelit graafikute abil. Kas mudel sobib? Kas y-tunnuse prognoosimisel on arvesse võetud kogu informatsioon, mis sisaldub tunnustes x_1 ja x_2 ? Kas graafikuid vaadates on võimalik öelda, kas mudeli prognoos kohal $x_1=1120$, $x_2=90$ on pigem ala- või üleprognoos? Kuidas saaks esitatud prognoosi parandada?

```
andmed=read.csv(url(
  "http://www.ms.ut.ee/mart/linmud2018/sacracoronaunita.csv"),
  header=TRUE); attach(andmed)
m1=lm(y~x1+x2, data=andmed)
summary(m1)

plot(m1, 1)
```

Vahel saab mudeli puudujääkidest aru juba prognoos vs jääk graafikult, sageli on aga mõttekas ka suuremate mudelite korral iga sõltumatut tunnust eraldi vaadelda. Kas märkas mõnda puudujääki oma mudelis? Kas tunnuseid x_1 ja x_2 kasutades saaks anda veel paremat prognoosi y-tunnusele kui senikasutatud mudel seda teeb? Kuidas tuleks mudeli m1 prognoosi parandada?

```
windows(width=8, height=6)
plot(x1, residuals(m1), pch=20, cex=0.5)
abline(v=1120, col=2, lwd=2)
```

Tee sarnane graafik ka x_2 vs jäägid jaoks. Milline on sinu otsus / soovitus?