Lineaarsed mudelid

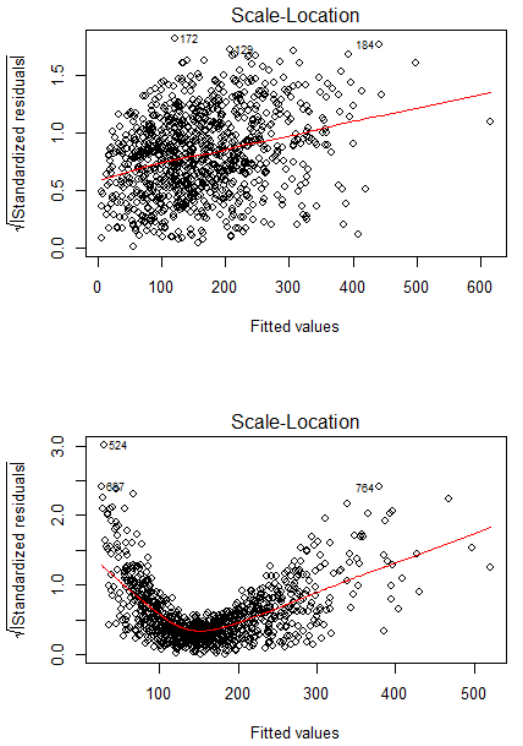
Eeldustest I

Mudeli eelduseid R-is kontrollides on esimeseks ja üheks olulisemaks abivahendiks plot-käsk. Nimelt produtseerib käsk `plot(mudel)` meile vaikimisi 4 diagnostilist graafikut, mis on mõeldud mudeli kuju kontrollimiseks (*Residuals vs Fitted*), normaaljaotuse eelduse kontrollimiseks (*Normal Q-Q*), konstantse hajuvuse eelduse kontrollimiseks (*Scale-Location*) ja rämedate sisestusvigade või andmevigade leidmiseks mõeldud graafik (*Residuals vs Leverage*).

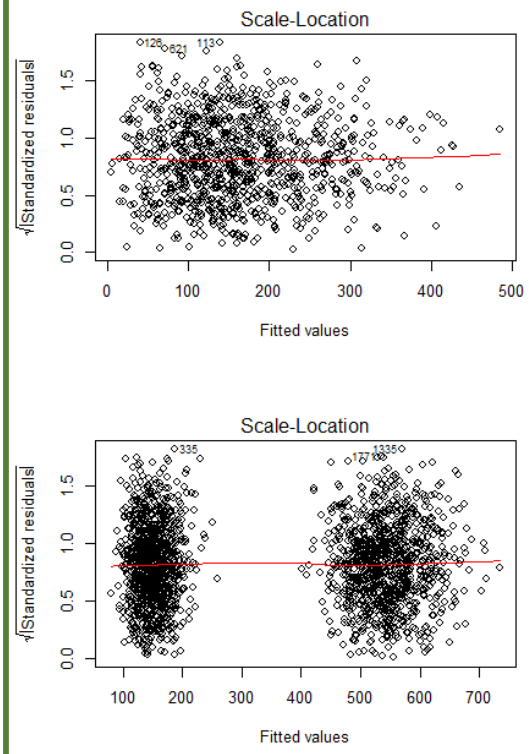
Alustame viimase kahe diagnostilise graafiku uurimisega/mõistmisega.

Esmalt konstantse hajuvuse eeldusest: $D(Y_i)=\sigma^2$ mistahes i väärtuse korral.

Konstantse hajuvuse eeldus on rikutud



Konstantse hajuvuse pole rikutud



Miks kasutatakse just sellist graafikut, mida ta tuvastab (ja mida ei märka)?

Katsetame seda kahel viisil simuleeritud andmestikuga.

```
# Variant 1:
x1=rep(0:1, c(90, 10))
y1=2+3*x1+rnorm(100, sd=rep(c(1,10), c(90, 10)))
mudel=lm(y1~x1)
summary(mudel)
par(mfrow=c(2,2))
plot(mudel)

# Variant 2
x2=rep(c(0,0,0,0,0,0,0,0,0,1), 10)
y2=2+3*x2+rnorm(100, sd=rep(c(1,10), c(90, 10)))
mudel2=lm(y2~x2)
summary(mudel2)
par(mfrow=c(2,2))
plot(mudel2)
```

Mille poolest need kaks simuleeritud andmestikku (y1, x1 vs y2, x2) teineteisest erinevad? Kas mõnes neist andmestikest on konstantse hajuvuse eeldus rikutud? Kas konstantse hajuvuse eelduse kontrollimiseks mõeldud diagnostiline graafik seda märkab? Kuidas on normaaljaotuse eeldusega?

Uurime neid kahte juhtu veidi põhjalikumalt, simuleerited palju andmestikke mõlemal viisil:

```
# Variant 1
hinnang=rep(NA, 1000);
se=rep(NA, 1000)

for (i in 1:1000){
      x=rep(0:1, c(90, 10))
      y=2+3*x+rnorm(100, sd=rep(c(1,10), c(90, 10)))
      hinnang[i]=coef(summary(lm(y~x)))[2,1]
      se[i]=      coef(summary(lm(y~x)))[2,2]
}
mean(hinnang)
mean(se)
sd(hinnang)

# Variant 2
for (i in 1:1000){
      x=rep(c(0,0,0,0,0,0,0,0,0,1), 10)
      y=2+3*x+rnorm(100, sd=rep(c(1,10), c(90, 10)))
      hinnang[i]=coef(summary(lm(y~x)))[2,1]
      se[i]=      coef(summary(lm(y~x)))[2,2]
}
mean(hinnang)
mean(se)
sd(hinnang)
```

Mida oskad simulatsioonide põhjal öelda? Kas hinnangud on nihketa? Mida märkad?

Probleemsete jääkide leidmiseks on sageli kõige mugavam kasutada (standardiseeritud) jäägid vs mõjukus (*Residuals vs Leverage*) graafikut. Mida suurem on vaatluse mõjukus, seda suuremat probleemi kujutab võimalik viga (näiteks sisestusviga või mõõtmisviga) nendes vaatlustes. Kui aga standardiseeritud jääk on väga suur või väga väike (-2 väiksemaid või +2 suuremaid jääke peaks esinema tõenäosusega 5%, -3 väiksemaid või +3 suuremaid jääkide esinemistõenäosus peaks normaaloludes olema aga juba kaduvväike (0,0027). Seega kui reaalselt näeme mõnda väga suurt või väga väikest standardiseeritud jääki võivad need olla tekkinud näiteks vaatlusandmete arvutisse sisestamisel tehtud veast. Seega tasuks nendele jääkidele vastavate objektide andmeid põhjalikult üle kontrollida. Sellele joonisele kantakse ka Cook'i kaugustele 0,5 ja 1 vastavad jooned – kui jääk jääb teisele poole Cooki kaugust 0,5 iseloomustavat joont, siis on vaatluse Cook'i kaugus suurem 0,5-st. Cooki kauguste graafikut saab soovi korral ka eraldi paluda: `plot(mudel,4)`.

Mida näitavad Cook'i kaugused? Seda selgitab järgmine simulatsioon (proovi see näide korra läbi):

Genereerime vaatlused

```
set.seed(1)
y=rnorm(90)
grupp=rep(1:3, each=30)
```

Tekitame ühe sisestusvea

```
y[1]=6.1
and=data.frame(y,grupp)
```

Hindame mudeli

```
m1=lm(y~factor(grupp)-1, data=and)
```

Mida Cooki kaugus näitab? Hindame teise mudeli ilma selle vaatluseta:

```
m1a=lm(y~factor(grupp)-1, data=and[-1,])
```

#Hinnatud mudelite võrdlus:

```
summary(m1)
coef(m1a)
```

Näeme, et muutus ainult 1. parameetri väärtus. Muutus umbes ühe standardvea võrra.
Kokku muutuvad mudeli parameetrid seega keskmiselt 1/3 standardvea võrra.
Seega peaks 1. vaatluse cooki kaugus olema umbes 1/3:

```
cooks.distance(m1)[1]
```

Antud arvutluskäik on täpne vaid siis, kui mudeli parameetrite hinanngud on
teineteisest sõltumatud, nagu näeme näiteks siit (hinnangute kovariatsioonid on 0-id)

```
vcov(m1)
```

Kui hinnangud oleksid sõltuvad, siis 2 tugevalt sõltuva parameetri hinnangu
muutus suurendab cooki kaugust vähem kui teistest hinnangutest sõltumatu parameetri
hinnangu muutus.

Cooki kauguseid saad kasutada näiteks selliste käskude abil:

Cooki kaugused arvuliselt:
```
cooks.distance(m1)
```

Cooki kaugused graafikul:
```
plot(m1,4)
```

Cooki kauguste kasutamine erindite leidmiseks mõeldud graafikul:
```
x=runif(100, 0, 10)
y=2+3*x+rnorm(100)
x[25]=15
naidismudel=lm(y~x)
plot(naidismudel, 5)
```

Õpitu kasutamine reaalse andmestiku analüüsimisel.

Kasutatud on dr. Marika Tammaru 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-1.ms.ut.ee/mart/linmud2025/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)
```

Kas mudel on hea? Kommenteeri mudeli sobivust (determinatsioonikordaja, olulisustõnäosus) ja arutle selle üle, kas tegemist on sama näitajaga – kas piisab, kui mõõdame patsientidel vaid ühte neist näitajatest?

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)
```

Kas ruutliige on statistiliselt oluline? Mis juhtub determinatsioonikordajaga?

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

Uuri, mis lisainformatsiooni annab mudelite eelduste kontrollimine. Kasuta käske:

```
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");
```

Graafikute järgi otsustades on vaatlus nr 40 on eripärane.

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

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

Võrdle mudelite m2 ja m2a determinatsioonikordajaid.

Võrdle ka 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?

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)
```

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

Mida teha erindiga, vaatlusega nr 40?