

Biomeetria bioloogidele
6. praktikum
Regressioonanalüüsi eeldustest

Näide 1

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

Andmed leiad:

load(url("http://www.ms.ut.ee/mart/biomeetria2013/reuma.RData"))

Muuda attach käsu abil andmestiku tunnused lihtsamini kasutatavaks.

Seetsinased anmed 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. Et vastata sellele küsimusele, tuleks uurida, kuidas muutub SR kiirenedes CRP väärtus

Andmetes:

SR1 – SR kiirus mm/h

CRP1 – CRP sisaldus veres mg/l.

Alusta lineaarse seose uurimisest, seda käskude abil:

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

Pane hinnatud mudel kirja. Kuidas kommenteerid mudeli sobivust (parandatud 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)  
AIC(m2) ; AIC(m1)
```

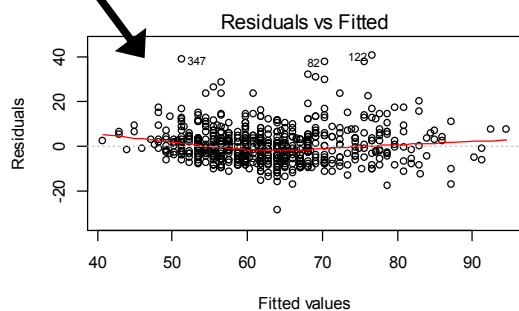
Mida ütleb parandatud determinatsioonikordaja? Milline peaks olema AIC-väärtuse põhjal tehtud otsus? Kas kuupliikme lisamisest oleks abi?

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

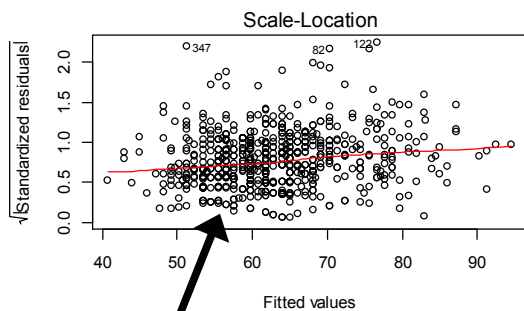
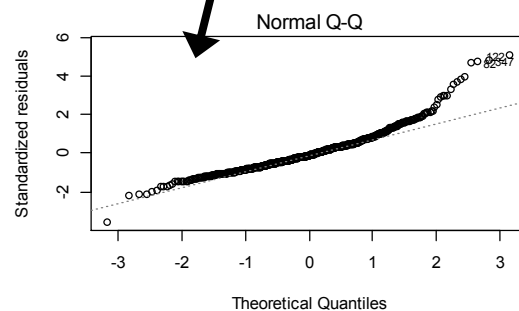
Vahepala: järgneva teoreetiline selgitus

Kui anda käsk `plot(mingi regressioonimudel)` siis R väljastab mudeli kohta neli diagnostilist graafikut, mille abil saab jälgida, kas mudeli eeldustega ikka on kõik korras.

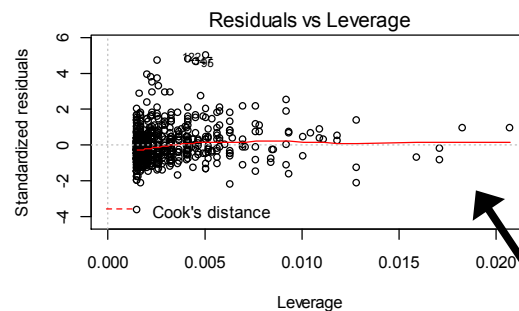
Kontrollib, kas valitud mudeli kuju on õige. Kui mudeli kuju on vale, siis peaksid mudeli jäägid mõnes piirkonnas olema süstemaatiliselt kas üleval või allpool nullijoont. Kui uuritav mudel sobib, siis peaks kõigis piirkondades olema nullist suuremaid ja nullist väiksemaid jääke ligikaudu samapalju.



Kontrollib, kas mudeli jäägid (prognosivead) on normaaljaotusega. Kui normaaljaotuse eeldusega on kõik korras, siis peaksid punktikesed toodud graafikul asuma enam-vähem ühel sirgel reas. Alltoodud graafiku korral on normaaljaotuse eeldus selgelt rikutud, seega on näiteks leitud prognoosiintervallid arvatavasti valed.



Kontrollib, kas vaatluste (täpsemalt: jääkide) hajuvus on alati sama. Kui konstantse hajuvuse eeldusega on kõik korras, siis ei tohiks joonisel olla näha mingit trendi, punktid peaksid kõikuma üsna juhuslikult ühe ja sama väärtuse ümber. Toodud joonisel aga on märgata kerget trendi: suuremate prognooside (fitted values) korral on jääkide – prognoosivigade – hajuvus suurem kui väiksemate prognooside korral.



Graafik on mõeldud erindite (näiteks sisestusvigade) tuvastamiseks. Joonise x-teljel on vaatluse mõjukus (Leverage) – mida suurem on vaatluse mõjukus, sena enam suudab antud vaatlus regressioonkõvera enda järgi paika panna, enda juurde „painutada“. Joonise y-teljel on toodud (standardiseeritud) jääk. Vaatlused mis on samal ajal nii mõjukad kui ka suure jäägiga väärivad väga hoolikat uurimist ja kontrollimist.

Jätkub töö konkreetse andmestikuga

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

```
par(mfrow=c(2,2))
plot(m1)

windows()
par(mfrow=c(2,2))
plot(m2, main="Mudel 2")
```

Vaatlus nr 40 on eripärane, tegemist on erindiga. Väärib edasist uurimist, miks see patsient on teistest erinev. Küsi õppejõult selle uurimise tagajärgde kohta!

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

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

Võrdle mudelite m1 ja m1a determinatsioonikordajaid, vaata kas peale 40. vaatluse eemaldamist vajame jätkuvalt oma mudelisse ruutliiget?

Vaata, kuidas mõjus erindi väljajätmine regressioonisirgele

```
plot(CRP1, SR1)
y=predict(m1, data.frame(CRP1=x))
lines(x,y, col=2, lty=2)
y=predict(m1a, data.frame(CRP1=x))
lines(x,y, col=2)

legend("bottomright", c("erindiga", "erindita"), lty=2:1, col=2:1)
points(CRP1[40], SR1[40], col=2, pch=20)
```

Ülesanne

Kujutame endale ette järgmist hüpoteetilist olukorda. Soovime võimalikult täpselt määrata patsiendi, täpsemalt tudengi, süstoolset vererõhku (SVR). Aga meditsiinisüsteemis on raha vähe ja vererõhu mõõtmine suhteliselt kallis (kulutab arsti väärtuslikku aega). Kulutuste mõttes oleks mõeldav mõõta patsiendi kas kaalu või pikkust, raha jätkub neist ühe mõõtmiseks. Kumba neist tunnust kasutades saaksime täpsemalt prognoosida tudengi süstoolset vererõhku? Proovi leida sellele küsimusele vastust kasutades varasematest praktikumidest tuttavat tudengite andmestikku!

Joonista saadud parim regressioonjoon ka välja, interpreteeri saadud tulemust. Kui hästi ikkagi on võimalik vererõhku prognoosida sinu poolt valitud tunnuse põhjal?

Teeme ühe abistava tunnuse, tunnuse *mees*, mis näitab, kas tegemist on mehega (1) või mitte (0).

```
mees=sugu-1  
table(mees)
```

Proovime prognoosida ka SVR-i vastloodud tunnuse abil:

```
regressioonmudel=lm(SVR~mees)  
summary(regressioonmudel)
```

Proovi interpreteerida saadud mudelit. Kui hästi prognoosib vastloodud mudel uuritava tunnuse väärtuseid, mida võiksid tähendada mudeli parameetrid?