Biomeetria

3. praktikum

Loe sisse andmestik lapsed2:

```
andmed = read.csv2(
      "http://www.ms.ut.ee/mart/biomeetria2015/lapsed2.csv",
      header=TRUE)
```

Vaata andmestiku esimesi ridu:

```
andmed[1:3,]
```

Antud andmestikus on järgmised tunnused:

vanus – lapse vanus mõõtmise tegemise hetkel (aastates)
kaal – lapse kaal (kg)
pikkus – lapse pikkus (cm).
sugu – lapse sugu

NB! Millise käsu peaksid andma, et saaksid hiljem selle andmestiku tunnuseid (lihtsalt) kasutada?

Meid huvitab, kas (ja kuidas) lapse vanemaks saades muutub lapse pikkus – tahame näha nn kasvukõveraid.

Esialgne mudel – meenutuseks…

```
m1=lm(pikkus~vanus)
summary(m1)
```

Milline näeb välja hinnatud mudel? Pane see kirja!

Pikkus = ……. + ……×Vanus + ……..

Meenutuseks eelmisest praktikumist: prognoosime oma mudelit kasutades 1,8 aasta vanuse lapse pikkust. Leiame ka 95%-usaldusintervalli 1,8 aastaste laste keskmisele pikkusele ja 95%-prognoosiintervalli 1,8 aastase lapse pikkusele:

```
> predict(m1, data.frame(vanus=1.8))
1 89.25716
> predict(m1, data.frame(vanus=1.8), interval="confidence")
       fit      lwr      upr
1 89.25716 89.17328 89.34105
> predict(m1, data.frame(vanus=1.8), interval="prediction")
       fit     lwr      upr
1 89.25716 82.3231 96.19123
```

Joonistame vanuse-pikkuse hajuvusgraafiku ning lisame graafikule meie poolt leitud regressioonisirge:

```
plot(vanus, pikkus, xlab="Vanus (aastates)",
                                ylab="Pikkus (cm)")
x=seq(0,2.2,0.01)
y=predict(m1, data.frame(vanus=x))
lines(x, y, col="red", lwd=2)
```

Millist probleemi märkad joonist vaadates?

Üritame sama probleemi märgata ka mudeli jääkide graafikult:

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

Kas märkad probleemi olemasolu ka sellel graafikul?

Varem saime 1,8 aasta vanuse lapse pikkuse prognoosiks 89cm. Vaata mudeli jääkide (tegelik väärtus – mudeli prognoos) graafikut ja ütle (ligikaudu) kuidas tuleks esialgset prognoosi korrigeerida saamaks mõistlikku prognoosi (kas esialgne mudel keskmiselt üle- või alahindab lapse pikkust, kui palju)?

Üritame nähtud viga parandada ja hinnata parema mudeli. Alustame otsides andmeid kõige paremini kirjeldava ruutpolünoomi:

```
m2=lm(pikkus~vanus+I(vanus^2))
```

```
> summary(m2)
[…]

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.93209    0.03487  1546.6   <2e-16 ***
vanus       29.21899    0.09447   309.3   <2e-16 ***
I(vanus^2)  -6.49518    0.05102  -127.3   <2e-16 ***
[…]
```

Näeme, et kõige paremini andmetega sobivaks vanuse ruutliiget sisaldavaks mudeliks on järgmine mudel:

$$Pikkus = 53{,}93 + 29{,}22*vanus - 6{,}495*vanus^2 + \varepsilon$$

Saadud mudel on selgelt parem AIC-väärtuse (Akaike Informatsioonikriteerium) põhjal otsustades (väiksem AIC väärtus näitab R-is paremini prognoosivat mudelit):

```
> AIC(m1)
[1] 268073.8
> AIC(m2)
[1] 254040.3
```

Võime ka küsida mõlema mudeli AIC-väärtuseid ühe käsu abil (soovitatav):

```
AIC(m1, m2)
```

Paremus on antud juhul ka lausa tõestatav (rikkam mudel m2 on tõestatavalt parem vaesemast mudelist m1):

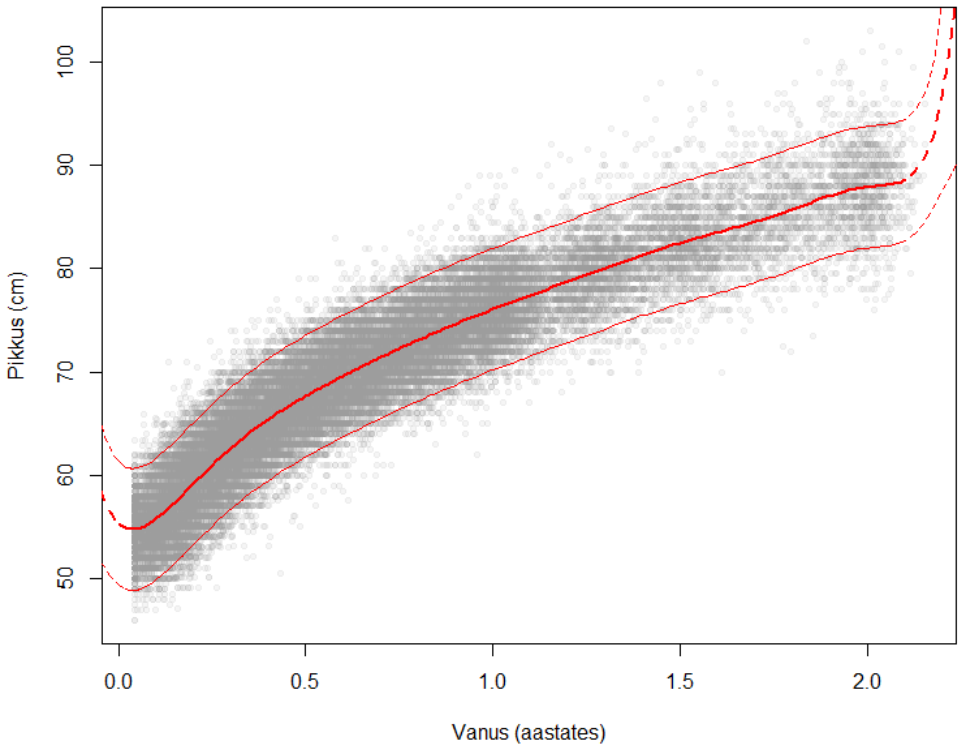
```
> anova(m1,m2)
Analysis of Variance Table

Model 1: pikkus ~ vanus
Model 2: pikkus ~ vanus + I(vanus^2)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1  49967 625285
2  49966 472163  1    153122 16204 < 2.2e-16 ***
```

Kui rikkam mudel on tõestatavalt parem vaesemast siis tuleb p-väärtus väike (nagu siin)

Paraku pole ka saadud ruutmudel piisavalt hea. Lisa mudelisse veel vanuse kuup ja soovi (vajaduse) korral ka vanuse kõrgemat järku astmeid. Leia hea mudel, mis kirjeldaks vanuse ja pikkuse vahelist seost tõepäraselt. Kanna saadud mudeli poolt kirjeldatud regressioonijoon (koos 95%-prognoosiintervalliga) ka graafikule:

```
mudel_hea = lm(pikkus~vanus+I(vanus^2)+ .....
x=seq(0,2.1,0.01)
y=predict(mudel_hea, data.frame(vanus=x))
plot(vanus, pikkus, xlab="Vanus (aastates)", ylab="Pikkus (cm)",
                                    pch=20, col=rgb(0.6,0.6,0.6,0.1))
lines(x,y, col= "red", lwd=2)
```



Saadud polünoomi abil konstrueeritud regressioonjoont võime võrrelda teise universaalse meetodi – b-splainide abil saadud regressioonjoontega. Splainide kasutamiseks peame aga kasutusse võtma lisamooduli splines:

```
library(splines)
```

Hindame b-splaine kasutades ka (võrdluseks) kaks mudelit – lihtsa murdjoone (1-järku b-splain) ja tavapärase kuupsplaini (3-järku b-splain):

```
# Tavaline murdjoon:
mudel_bspline1=lm(pikkus~bs(vanus, degree=1, df=16))
# "Peidetud" murdjoon (3-järku splain):
mudel_bspline2=lm(pikkus~bs(vanus, degree=3, df=15))
```

Kanname ka nende mudelite regressioonjooned hajuvusgraafikule (lisaks sinu poolt leitud regressioonjoonele):

```
plot(vanus, pikkus, xlab="Vanus (aastates)",
     ylab="Pikkus (cm)", xlim=c(0,2.5), ylim=c(45,120))

x=seq(0,2.6,0.01)

# Sinu poolt leitud mudeli regressioonjoon:
y=predict(mudel_hea, data.frame(vanus=x))
lines(x,y, col= "red", lwd=2)

# Murdjoon (1-järku b-splain):
y=predict(mudel_bspline1, data.frame(vanus=x))
lines(x,y, col= "pink", lwd=2)

# Kuupsplain (3-järku b-splain):
y=predict(mudel_bspline2, data.frame(vanus=x))
lines(x,y, col= "blue", lwd=2)
```

Mille poolest saadud regressioonjooned erinevad, mille poolest sarnanevad?

Muuseas, ka spliine kasutavate mudelite korral saab võrrelda keerukamat ja lihtsamat mudelit omavahel nii AIC-väärtuste abil või soovi korral testida kas keerukam mudel (suuremate vabadusastmete arvuga) on tõestatavalt parem lihtsamast (väiksema vabadusastmete arvuga ehk vähemate hinnatavate parameetritega mudelist):

```
mudel_bspline2=lm(pikkus~bs(vanus, degree=3, df=15))
mudel_bspline3=lm(pikkus~bs(vanus, degree=3, df=16))

AIC(mudel_bspline2, mudel_bspline3)

anova(mudel_bspline2, mudel_bspline3)
```

Näeme, et antud juhul otsustavad nii AIC-kriteerium kui ka p-väärtusel baseeruv mudeli valik lihtsama mudeli kasuks (mudeli mudel_bspline2 AIC-väärtus on väikem;

keerukam mudel mudel_bspline3 pole tõestatavalt parem lihtsamast). Alati ei pruugi need kaks mudeli valiku viisi siiski viia samasuguste tulemusteni…

Ülesanne 1.

Prognoositakse tavaliselt lapse kasvu ikkagi tütarlastele ja poisslastele eraldi. Hinda sinagi kasvukõver vaid poisslaste jaoks (nende jaoks, kellel tunnuse sugu väärtuseks on M) !

Erindid

Vahel häirivad andmeanalüüsi erakordsete väärtustega vaatlused – nn erindid (outliers). Erindite tekkepõhjused on erinevad, väga sageli osutuvad erindid sisestusvigadeks (mõni väärtus on valesti sisestatud, näiteks on ühe inimese pikkus meetrite asemel sentimeetrites kirja pandud vms). Vahel on aga tegemist täiesti õige ja korrektselt sisestatud vaatlusega, mis mingil põhjusel siiski käitub erinevalt teistest vaatlustest. Kuidas erinditega käituda ja millist mõju erindid võivad omada, seda vaatame järgmist näidet kasutades (*Kasutatud on dr. Marika Tammaru andmeid ja tema selgitusi andmetele*).

Andmed leiad:
load(url("http://www.ms.ut.ee/mart/biomeetria2015/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. Et vastata sellele küsimusele, tuleks uurida, kuidas muutub SR kiirenedes CRP väärtus ja kuivõrd täpselt on ühe näitaja abil võimalik teist prognoosida (kui prognoos on piisavalt täpne, siis pole mõlemaid näitajaid mõtet kallite laborianalüüside abil eraldi mõõta…):

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, olulisustõnäosus)?

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?

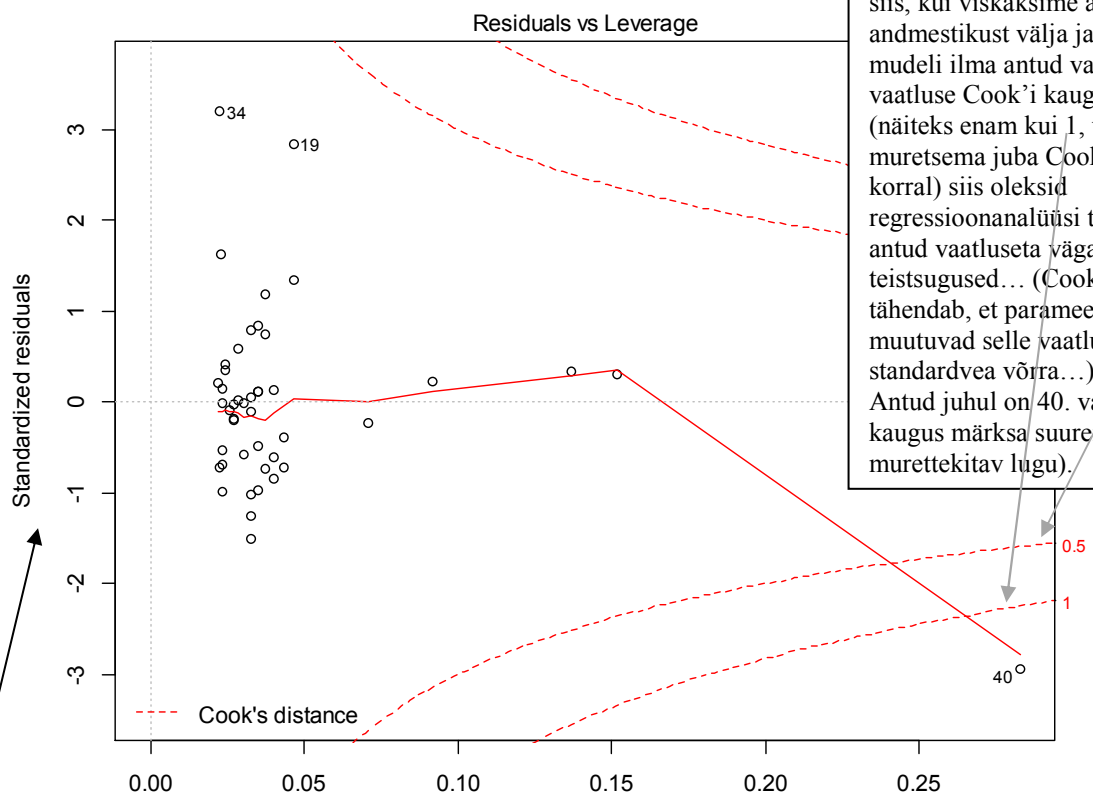
Kanna mudeli m2 poolt kirjeldatud regressioonijoon hajuvusgraafikule.
Mis torkab silma?

Mudeli eelduste kontrollimiseks pakub R erinevaid diagnostilisi graafikuid. Ühega neist – mudeli kuju kontrollimiseks mõeldud graafikuga oleme juba tuttavad:

```
# Mudeli (keskväärtusele) õigsuse kontrollimiseks mõeldud graafik
plot(m1, 1)
```

Üks teine diagnostiline graafik on mõeldud aga eelkõige erindite uurimiseks – vaatluste leidmiseks mis teistest erinevad (ja mis võivad analüüsi tulemusi tugevalt mõjutada).

```
plot(m1, 5)
```



Cooki kaugus (Cook's distance) – iseloomustab kuivõrd erinevad tuleksid mudeli hinnatud parameetrid siis, kui viskaksime antud vaatluse andmestikust välja ja hindaksime mudeli ilma antud vaatluseta. Kui vaatluse Cook'i kaugus on suur (näiteks enam kui 1, vahel hakatakse muretsema juba Cook'i kauguse 0,5 korral) siis oleksid regressioonanalüüsi tulemused ilma antud vaatluseta vägagi teistsugused… (Cook'i kaugus 1 tähendab, et parameetrite hinnangud muutuvad selle vaatluse tõttu ~1 standardvea võrra…)
Antud juhul on 40. vaatluse Cooki kaugus märksa suurem kui 1 (väga murettekitav lugu).

Jääk jagatud oma standardhälbega. Enamik standardiseeritud jääke peaks jääma vahemikku -2…2, Kui siit vahemikust jääb välja ohtrasti jääke võib see viidata normaaljaotuse eelduse rikutusele või mõnele muule probleemile.
Absoluutväärtuselt väga suured (-3 või väiksemad; 3 või suuremad) jäägid võivad viidata erinditele/sisestusvigadele või kasutatud mudeli valele kujule.

Vaatluse mõjukus (leverage). Suure mõjuga vaatlustel on võime regressioonjoont tuntavalt mõjutada, väikese mõjukusega vaatluse võime regressioonanalüüsi tulemusi tuksi keerata on väike.
Ligikaudne interpretatsioon: Kui mitme parameetri väärtuse määrab ära antud vaatlus….

Võrdle mõlema mudeli (m1, m2) korral erindite leidmiseks mõeldud graafikuid:

```
plot(m1, 5, main="Mudel 1")
plot(m2, 5, main="Mudel 2")
```

Vaatlus nr 40 on eripärane, tegemist on erindiga. Väärib edasist uurimist, miks see patsient on teistest erinev. Seda teie teha ei saa, sest teie pole neid andmeid korjanud. Küll aga päriselus uuriti antud patsiendi eripära põhjuseid hoolega. Küsi õppejõult selle uurimise tagajärgede 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)
```

Kas peale erindi eemaldamist peaksime kasutama lihtsamat või keerulisemat mudelit?

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

AIC(m1a, m2a)
anova(m1a, m2a)
```

Millisele otsusele jõuad?

Julgematele – simulatsioon (kus tõde teada…)

Järgnev programm genereerib 50 vaatluse andmed. Mõõdetud on 6-t sõltumatut tunnust (x-tunnust, tunnust mille abil üritatakse prognoosi leida) ja lisaks ka y-tunnuse vaatlused (mis sõltub x-tunnustest):

```
set.seed(1)

n=50
x1=rnorm(n)
x2=rnorm(n)
x3=rnorm(n)
x4=rnorm(n)
x5=rnorm(n)
x6=rnorm(n)

y=10+8*x1+0.5*x2-0.01*x3+0.01*x4+0.001*x5+0.01*x6+rnorm(n)
```

Hindame saadud andmete pealt kolm mudelit:

```
m1=lm(y~x1)
m2=lm(y~x1+x2)
m3=lm(y~x1+x2+x3+x4+x5+x6)
```

Pane tähele: Milline mudelitest tegelikult õige on? Õige mudel on m3 (sest y-tunnuse väärtuste moodustamisel on kasutatud nii tunnuseid x1, x2, x3, x4, x5 kui ka x6-te).

Vaata, millist nendest mudelitest peab AIC tulevasi vaatluseid kõige täpsemini prognoosivaks mudeliks:

```
AIC(m1, m2, m3)
```

Proovime järgi!

Laseme täpselt samal andmeid tekitaval mehhanismil luua veel 1000 vaatluse jaoks x-tunnuste ja y-tunnuse väärtused:

```
n=1000
x1=rnorm(n)
x2=rnorm(n)
x3=rnorm(n)
x4=rnorm(n)
x5=rnorm(n)
x6=rnorm(n)

y_uus=10+8*x1+0.5*x2-0.01*x3+0.01*x4+0.001*x5+0.01*x6+rnorm(n)
```

Leiame nende uute vaatluste y-tunnuse väärtustele prognoosid kõigi kaalumisel oleva kolme mudeli abil:

```
prognoos1=predict(m1, data.frame(x1,x2,x3,x4,x5,x6))
prognoos2=predict(m2, data.frame(x1,x2,x3,x4,x5,x6))
prognoos3=predict(m3, data.frame(x1,x2,x3,x4,x5,x6))
```

Vaatame, millise mudeli prognoosid tulid kõige lähemale õigetele (tegelikele) y-tunnuse väärtustele. Selleks leiame iga mudeli prognooside keskmise ruutvea:

```
mean((y_uus-prognoos1)**2)
mean((y_uus-prognoos2)**2)
mean((y_uus-prognoos3)**2)
```

Millise mudeli prognoosid olid kõige täpsemad? Kas õige mudeli omad? Või AIC-kriteeriumi järgi väljavalitud mudeli omad?