Lineaarsed mudelid
Eeldused II

Näide 1. Pärisandmed.
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/linmud2020/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)
```

Ülesanne 1: Kommenteeri diagnostilisi graafikuid! Kas mudeliga on kõik korras, kui mitte, siis milles võiks olla probleem?

Vaata, mida soovitab Box-Cox-i transformatsioon!

```
library(MASS)

boxcox(m1)
```

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

Ülesanne 2: Pane kirja kasutatud transformatsioon, kommenteeri transformeeritud y-tunnust kasutava mudeli eelduseid. Interpreteeri tulemuseks saadud mudelit!

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, 7))
plot(m1)
```

Mida oskad nendelt graafikutelt välja lugeda?
Vaata, millist transformatsiooni soovitab Box-Cox'i meetod?

```
boxcox(m1)
```

Kas soovitatud transformatsioon on ootuspärane (sina ju praegu tead andmeid genereerivat mehhanismi)?

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

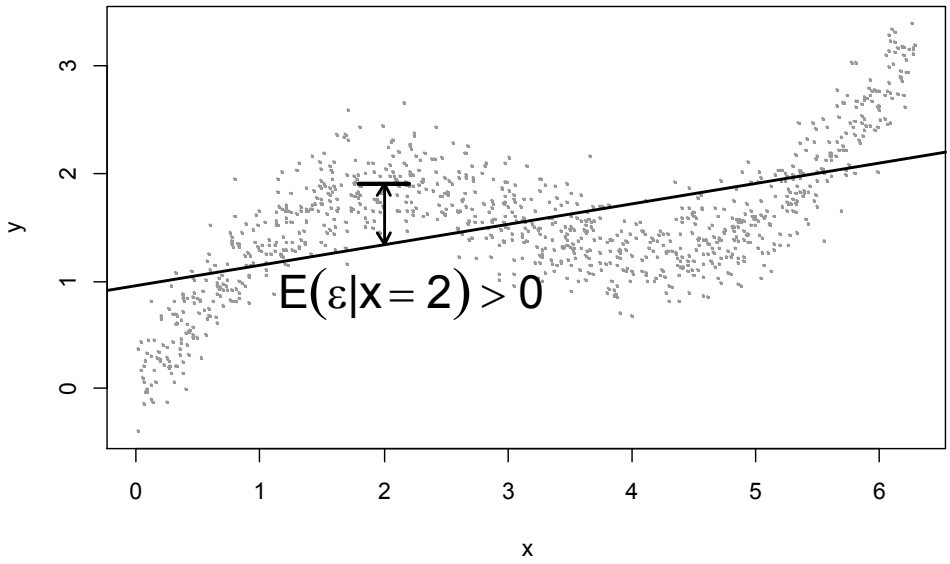
```
plot(x,y, pch=20, col="gray80")
xx=seq(4,21, length=1000)
lines(xx, f(xx), col=2, lwd=2)
y1=predict(m1, data.frame(x=xx))
lines(xx, y1, col="pink", lwd=3)

# Täida lüngad (…) enne käivitamist!
m2=lm(…~poly(x, 7))
y2=predict(m2, data.frame(x=xx), interval="prediction")
lines(xx, …, col="gold2", lwd=3)
lines(xx, …, col="gold")
lines(xx, …, col="gold")
```

Võrdle saadud prognoose (mudelid m1 ja m2) ja vaata prognoosiintervalli – märka, et prognoosiintervall on muutunud peale uuritava tunnuse transformeerimist asümmeetriliseks!

Mudeli kuju

Lineaarse mudeli oluliseks eelduseks on nõue, et jääkide keskväärtus 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=6)
```

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 (graafik 1 või graafik 2) pealt on kergem märgata, et meie mudel pole täiuslik?

Sageli eelistatakse hoopis prognoos vs jääk graafikut:

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

# või, alternatiivina:

plot(m1,1)
```

Kui meie hädine mudel prognoosib *y*-tunnuse väärtuseks 70 (x1=7,56), siis milline võiks prognoos olla tegelikult?

Vastus: jääkide jooniselt näeme, et parem prognoos oleks umbes 10 ühikut suurem (prognoosi 70 korral on jäägid süstemaatiliselt positiivsed, keskmine jääk on umbes +10). Seega kui meie mudel prognoosib väärtust 70 siis oleks tark prognoosi tõsta 80-le.

Kontrolli: milline on tegelik y-tunnuse tinglik keskväärtus, E(y| x1=7,56)? Kuna praegu tead andmeid genereerivat mehhanismi, saad sa sellele küsimusele vastata. Kas tulemus on lähemal 70-le või 80-le?

Kui mudeli prognoosi saab kerge vaevaga parandada, siis on kasutatav mudel vale. Paranda mudel vastuvõetavaks (kasutades polünoomi või splaine) – ilma mudeli tegelikku kuju piilumata – ja vaata siis jääkide graafikuid! Kas diagnostilised graafikud tunduvad nüüd vastuvõetavad?

Kanna hajuvusgraafikule nii oma mudeli prognoosid kui ka tegelik seos – kas jõudsid tõele lähedale? Millist väärtust prognoosib parandatud mudel väärtuse x1=7,56 korral?

Ülesanne 3

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/linmud2020/sacracoronaunita.csv"),
  header=TRUE);

plot(andmed$x1, andmed$y)
plot(andmed$x2, andmed$y)

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 mõtekas ka suuremate mudelite korral mudeli jääke iga sõltumatu tunnusega 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(andmed$x1, residuals(m1), pch=20, cex=0.5)
abline(v=1120, col=2, lwd=2)
```

Antud jääkide graafikut vaadates ei saa küll kindlalt öelda, et x_1=1120 ümbruses meie mudeli jäägid oleksid süstemaatiliselt ülevalpool 0-i või allpool nulli. Pigem paistab olevat usutav, et jääkide keskmine x_1=1120 korral võikski olla ligikaudu null.

Tee sarnane graafik ka x_2 vs jäägid jaoks. Kas meie mudel tõenäolisemalt üle- või alahindab uuritava tunnuse väärtust (x_1=1120; x_2=90 korral)? Kui palju?