Lineaarsed mudelid

1. praktikum

Lineaarsed mudelid R-is. Käsk lm (***lm***: ***l****inear* ***m****odel*).

Esmalt tutvume lm-käsuga (ja selgitame välja kui häid teadmiseid lineaarsete mudelite kohta olete omandanud eelnenud õpingute käigus).

Loeme sisse Tartu Ülikooli (arstiteaduskonna) tudengite andmestiku:

```
print(load(url("http://www.ms.ut.ee/mart/linmud2020/kysitlus.RData")))
tudengid[1:3,]
attach(tudengid)
```

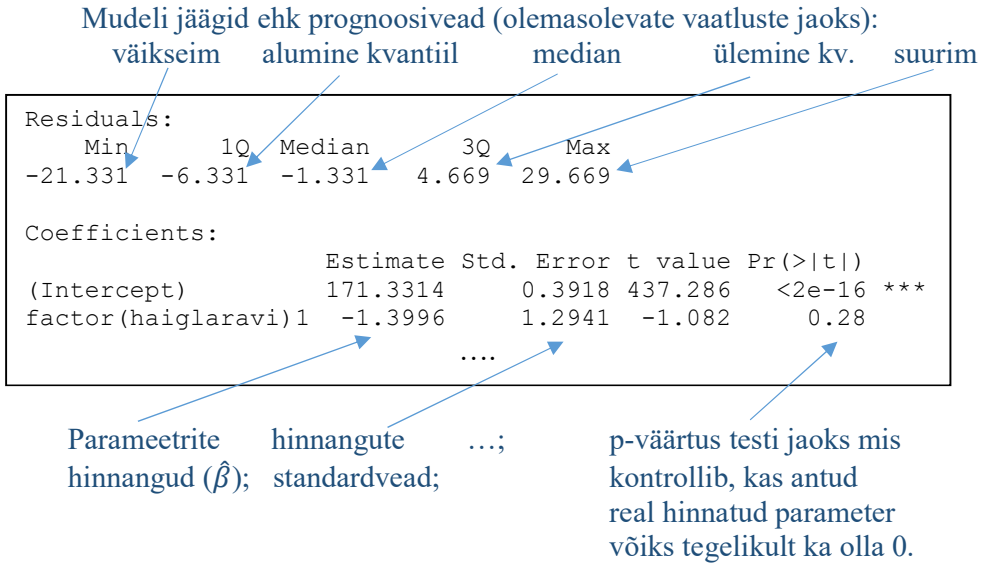
Alljärgnevalt märgime ära mõned andmestikus esinevad tunnused:

pikkus – tudengi pikkus (cm)
sugu – naine või mees
olu – mitu pudelit nädalas tudeng õlut joob
haiglaravi – kas tudeng on viimase kahe aasta jooksul vajanud haiglaravi (1) või mitte (0)

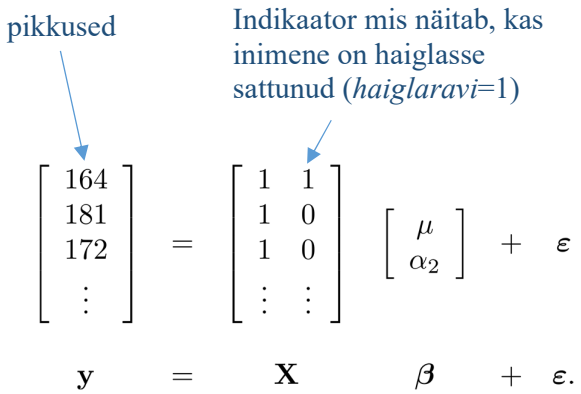
Näide 1

Hindame esmalt ühe lihtsa dispersioonanalüüsi mudeli:

```
mudel1 = lm(pikkus ~ factor(haiglaravi))
summary(mudel1)
```



Seega hinnatud mudel maatrikskujul näeb välja selline:



Ehk hinnatud mudeli võib kirja panna ka nii:

$$pikkus = 171{,}3 - 1{,}4\ I(haiglaravi{=}1) + \varepsilon$$

Ehk haiglaravi mittevajanud tudengite keskmiseks pikkuseks on hinnatud 171,3-1,4*0 = 171,3 ja haiglarvai vajanud tudengite keskmiseks pikkuseks 171,3-1,4*1=169,9.

Pane tähele, et haiglaravi mittesaanutele vastava mõju on R otsustanud võtta nulliks ($\alpha_1 = 0$). Dispersioonanalüüsi mudel on üleparametriseeritud ja parameetrite üheseks määramiseks tuleb kasutada lisatingimusi ehk reparametriseerimistingimusi. Võttes haiglaravi mittevajavate tudengitele vastava mõju nulliks määrame selle grupi sisuliselt referentsgrupiks – kõiki teisi gruppe hakatakse võrdlema referentsgrupi suhtes, kõigi teiste gruppidele vastavad mõjud iseloomustavad nonde teiste gruppide keskmiste erinevust referentsgrupist (ehk siis praegu haiglaravi mittevajanud inimeste keskmisest pikkusest)

Kas haiglaravi vajamise ja tudengi pikkuse vahel eksisteerib seos? Seda näeme vaadates haiglaravile sattumise indikaatori ees kordaja olulisust vaadates. Kuna p-väärtus on suur siis kordaja võib olla ka null ja seega ei pruugi pikkuste keskväärtus olla haiglaravi vajanud ja mittevajanud tudengite vahel olla erinev. Seega tõestada seose olemasolu nende tunnuste vahel ei saa.

Keskväärtuste võrdlemist võiksime aga praegu teha ka t-testi abil (sest antud juhul on ju vaid kaks gruppi: haiglarvil viibinud/mitteviibinud):

```
t.test(pikkus ~ factor(haiglaravi))
```

Võrdle näiteks mõlemal meetodil saadud p-väärtuseid!

Miks erinevad tulemuseks saadud p-väärtused teineteisest (tulemused on üsna sarnased, aga siiski mitte täpselt samad)? Vihjeks: erinevus tuleneb ühest eeldusest, mida lineaarse mudeli puhul on tehtud ja mida t-testi tehes praegu ei tehta. Mis eeldus see võiks olla?

Kuidas saada t.test-käsu abil täpselt samasugune p-väärtus kui lm-käsku kasutades?

Vihjeks: vaata ka abiinformatsiooni käsu t.test kohta: ?t.test

Ülesanne

Uurime, kas tudengitele rohkem õlut jootes saaksime pikemaid tudengeid.

```
mudel2 = lm(pikkus ~ factor(olu))
summary(mudel2)
table(olu)
```

Vasta järgmistele küsimustele:

1. Pane kirja enda poolt hinnatud dispersioonanalüüsi mudel:

   Pikkus = ……

2. Milline on õlut mittejoovate tudengite keskmine pikkus: ……..

3. Mida näitab nn 1-5 pudeli mõju (6,7734)? Millise kahe grupi keskmiste pikkuste erinevust ta iseloomustab?

   …….

4. Kas õlletarbimise ja tudengi pikkuse vahel on seos? Milline see seos on? Mis võiks nähtud seose põhjuseks olla?

   ……...

Üks vägagi kasulik käsk, mida töös lineaarsete mudelitega vaja läheb, on funktsioon estimable (lisamoodulist gmodels). Selle funktsiooni kaheks peamiseks argumendiks on lm-käsu abil hinnatud mudel ja vektor, mis näitab millist mudeli parameetrit mitu korda arvutustes arvesse võtta (täiendavalt võib nõuda aga ka muud, näiteks usalduspiire arvutuste tulemustele):

```
install.packages("gmodels")
library(gmodels)
estimable(mudel2, c(1, 0, 0, 0, 1), conf.int=0.95)
```

```
           Estimate Std. Error  t value  DF Pr(>|t|) Lower.CI Upper.CI
(1 0 0 0 1) 180.2143   2.993941 60.19299 655        0 174.3354 186.0932
```

Mida iseloomustab (millise grupi keskmist) lahtris *Estimate* antud number (180,2…)?

Võrdle estimable-käsu tulemust ka järgmise käsu tulemusega:

```
predict(mudel2, data.frame(olu=">12"), interval="confidence")
```

Kas saad nüüd aru, mida teeb estimate-käsk? Näita seda vastates järgmisele küsimusele!

Ülesanne

Proovi interpreteerida järgmise käsu väljundit – mida siin on hinnatud (ja testitud)?

```
estimable(mudel2, c(0, 0, 0, 1, -1), conf.int=0.95)
```

Vastus: ……..

Näide 3

Uurime jätkuvalt pikkuse ja õlletarbimise vahelist seost, lisame mudelisse ka tudengi soo:

```
mudel3 = lm( pikkus ~ factor(olu)+factor(sugu) )
summary(mudel3)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      168.02792    0.37514 447.904   <2e-16 ***
factor(olu)<1     -0.19703    0.52584  -0.375    0.708
factor(olu)1-5     0.58049    0.78367   0.741    0.459
factor(olu)5-12   -1.76034    1.25775  -1.400    0.162
factor(olu)>12     0.08445    2.36404   0.036    0.972
factor(sugu)mees  14.11890    0.64873  21.764   <2e-16 ***
```

Mida näitab (kuidas on interpreteeritav) 1-5 pudeli õlle tarbimisele vastav mõju (0,58) käesolevas mudelis? See mõju ei kirjelda enam seda, kuidas erinevad 1-5 pudelit nädalas õlut joovate tudengite keskmine pikkus ja õlut mittejoovate tudengite keskmine pikkus. Milliste gruppide erinevust siis antud mõju kirjeldab?

Milline on antud mudeli prognoos 1-5 pudelit õlut nädalas joova naistudengi pikkusele:

168 – 0,19*0 + 0,58*1 -1,76*0 0,084*0 + 14,12 *0 = **168 + 0,58**

Milline on antud mudeli prognoos õlut mittejoova naistudengi pikkusele:

168 – 0,19*0 + 0,58*0 -1,76*0 0,084*0 + 14,12 *0 = **168**

Seega iseloomustab 0,58 seda, kui palju 1-5 pudelit õlut tarbiv naistudeng võiks olla pikem õlut mittejoovast naistudengist.

Samas võime kirja panna ka antud mudeli prognoosid 1-5 pudelit õlut tarbiva meestudengi pikkuse jaoks ja õlut mittetarbiva meestudengi pikkusele:

1-5 pudelit õlut, mees:

168 – 0,19*0 + 0,58*1 -1,76*0 0,084*0 + 14,12 *1 = **168 + 14,12 + 0,58**

õlut ei joo, mees:

168 – 0,19*0 + 0,58*0 -1,76*0 0,084*0 + 14,12 *1 = **168 + 14,12**

Seega näitab see ka erinevust õlut mittejoovate meeste ja 1-5 pudelit nädalas õlut joovate meeste keskmiste pikkuste vahel.

Seega täpsem interpretatsioon:

Kui võrdleksime samast soost tudengeid, ühed neist ei joo õlut ja teised tarbivad 1-5 pudelit õlut nädalas, siis nende tudengite keskmiste pikkuste erinevus on mudeli arvates 0,58cm (andmete lisandumisel võib aga selguda, et tegelikku erinevust ka pole, sest vastav mõju pole statistiliselt oluline). See erinevus ei iseloomusta aga erinevust 1-5 pudelit õlut nädalas joovate tudengite (kes on enamasti mehed) ja õlut mittejoovate tudengite (enamasti naised) keskmiste pikkuste vahel

Ülesanne

Interpreteeri mida näitab antud mudelis ”mehe” mõju!

……

Näide 4 -regressioonanalüüs

Uurime, miks mõnes riigis on inimesed õnnelikumad kui teises. Loeme sisse õnneandmestiku:

```
andmed=read.csv(
  url("http://www.ms.ut.ee/mart/linmud2020/Onn_kokaiin_majandus_2006.csv"),
  header=TRUE)

andmed[1:3,]
```

Viskame andmestikust välja mõned riigid, kus meid huvitavate tunnuste väärtused pole teada:

```
andmed_vaike=andmed[!is.na(andmed$onn) & !is.na(andmed$kokaiin),]
attach(andmed_vaike)
```

Tunnuste tähendused (enamike tunnuste väärtused on 2006. aasta seisuga):

riik – riigi nimi

piirkond – õppejõu suva järgi määratud geograafiline piirkond, kus vastav riik asub.

onn - Kui õnnelikud on inimesed (keskmiselt) mingis riigis. Mõõdetud skaalal 0-10. Suuremad numbrid näitavad õnnelikumaid inimesi. Andmed pärit õnneandmestikust: http://worlddatabaseofhappiness.eur.nl/

kokaiin, kanep, amfetamiin, oopium – vastava narkootikumi tarvitajate protsent täiskasvanud elanikkonnast (2009. aasta andmed, andmed pärit: http://www.guardian.co.uk/news/datablog/2009/jun/24/drugs-trade-drugs)

Ülejäänud andmed pärinevad maailmapangast (http://data.worldbank.org/)

haridusraha – haridusele kulutatud vahenite protsent SKP-st

laenuintress – pankade poolt väljastatud laenude keskmine intress

....

Vaatame, kas inimeste õnn sõltub kokaiinist:

```
mudel = lm(onn~kokaiin)
summary(mudel)
```

Kirjuta välja saadud regressioonimudel:

Õnn = ...

Tore, et R meile midagi arvutab. Kas ta aga teeb oma arvutused korrektselt?

Kas mudeli parameetrid on ikka õieti arvutatud? Kordame arvutusi „käsitsi“. Mäletatavasti saime parameetrite vektori β hindamiseks järgmise valemi (kui lähtusime vähimruutude printsiibis): $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$ Moodustame kõigepealt edaspidistes arvutustes kasutamiseks mudeli maatriksi $\mathbf{X}$:

```
X=cbind(1, kokaiin)
X
```

arvutame hinnangu parameetervektorile (%*% on maatrikskorrutis; funktsioon t() transponeerib maatriksi või vektori; solve-leiab pöördmaatriksi):

```
beeta = solve(t(X)%*%X)%*%t(X)%*% onn
beeta
```

Kas said samad hinnangud kui lm-käsu abil?

kas mudeli prognoosid vaatlustele tulevad ka samasugused? Võrdle:

```
p1=X%*%beeta
p1
p2=predict(mudel)
p2
data.frame(p1, p2)
```

Võid võrrelda ka Sinu ja R'i poolt kasutatud mudelimaatrikseid:

```
X
model.matrix(mudel)
```

Või prognoosime, kui õnnelikud ollakse keskmiselt riigis, kus 4% elanikest tarvitab kokaiini:

```
predict(mudel, data.frame(kokaiin=4))

lambda=c(1, 4)
t(lambda)%*%beeta

estimable(mudel,lambda)
```

Vaatame, milline näeb välja hinnatud regressioonisirge:

```
windows(width=8, height=6)
plot(kokaiin, onn)

x=seq(0,4, length=100)
y=predict(mudel, data.frame(kokaiin=x))
lines(x,y, lwd=2)

x1=kokaiin[riik=="Estonia"]; y1=onn[riik=="Estonia"];

points(x1, y1, pch=20, col="red", cex=2)
text(x1+0.15, y1-0.2, "Eesti", adj=c(0,1), col="red")

arrows(x1+0.15, y1-0.2, x1, y1, col=2, length=0.15)
```

Lisame ka ühe teise kõvera joonisele:

```
X2 = cbind(1, onn)
P2 = X2 %*% solve(t(X2)%*%X2) %*% t(X2)
lines(P2%*%kokaiin, onn, lwd=2, col=2)
```

Ülesanne

Mida võiks iseloomustada see teine joonisele lisatud sirge? Proovi interpreteerida antud sirge tähendust!

Ülesanne

Ühes maailmale suletud kommunistlikus riigis saavad kõik inimesed riigi käest palka. Luureorganisatsioon, mille heaks töötad, soovib hinnata antud riigi keskmist palka. Salakavala satelliidi abil saab igal kuul määrata ühe juhuslikult valitud inimese palga suuruse. Sel viisil on juba kogutud andmeid päris mitme inimese kohta. Ühtäkki aga otsustab riigi juhtkond, et järgmisest kuust saavad inimesed teatud summa võrra rohkem raha. Kõigi inimeste palgale lisatakse samasugune rahasumma (tegemist on ju kommunistliku riigiga) ja pealt kuulatud vestluse järgi on lisatud summa valitud spetsiaalselt nii, et riigi keskmine palk kahekordistuks.

Hinda luureandmete põhjal, milline on uus keskmine palk. Kasuta selleks kõiki kogutud andmeid (vihje: peale palkade muutmist on keskmine palk on ju esialgsest kaks korda suurem – seega on meil kõigest üks tundmatu parameeter, mida peaksime hindama!).

Luureandmed:

Inimene	palk	mõõtmine tehtud enne/pärast palgatõusu
1	123	enne
2	145	enne
3	155	enne
4	119	enne
5	190	pärast
6	260	pärast
7	245	pärast
8	280	pärast
9	310	pärast

Millist mudelit kasutad küsimusele vastamiseks? Milline näeb välja mudelimaatriks? Millise hinnanguni jõuad?

Muuseas, miks ei tohiks kasutada sama lähenemist siis, kui riik oleks iga inimese palganumbri kahekordistanud? Paku välja, kuidas võiks samu andmeid analüüsida siis, kui palgatõus oleks läbi viidud iga inimese palganumbrit kahekordistades?