Leitav aadressilt
http://www-1.ms.ut.ee/mart/linmud2020/

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 paar lihtsat dispersioonanalüüsi mudelit:

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

Kas haiglaravi vajamise ja tudengi pikkuse vahel eksisteerib seos? Kas tulemus tuleb sama kui t.test-käsku kasutades:

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

Miks erinevad tulemuseks saadud p-väärtused teineteisest (tulemused on sarnased, aga siiski mitte täpselt samad)? Kuidas saada t.test-käsu abil täpselt samasugune p-väärtus kui lm-käsku kasutades?

Näide 2

Uurime, kas tudengitele rohkem õlut jootes saaksime pikemaid tudengeid.

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

Vasta järgmistele küsimustele:

1. Pane hinnatud dispersioonanalüüsi mudel kirja:

   Pikkus = ……

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

3. Mida näitab nn 1-5 pudeli mõju (6,7734)?

   …….

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):

```
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 lahtris *Estimate* antud number (180,2…)?

Soovi korral võrdle saadud tulemust ka järgmise käsu tulemusega:

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

Kas saad nüüd aru, mida teeb estimate-käsk? Proovi siis interpreteerida järgmise käsu väljundit – mida siin on hinnatud (ja testitud)?

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

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 käesolevas mudelis? Kuidas interpreteerida antud mudelis mõju ”mees”?

Arvuta ise, milline on (ligikaudu) antud mudeli prognoos 1-5 pudelit õlut nädalas joova naistudengi pikkusele:

…….

Lisame mudelisse ka õlletarbimise ja soo koosmõju (vastava koosmõju järgi pole praegu küll vajadust, aga proovime ka koosmõjudega mudeli mõjusid interpreteerimida…)

```
mudel4 = lm(pikkus ~ factor(olu)+factor(sugu)+
                                 factor(olu)*factor(sugu))
summary(mudel4)
```

Arvuta ise (käsitsi) mudel4 hinnanguid kasutades pikkuse prognoosid järgmistele tudengitele:

a) Mees, joob 1-5 pudelit õlut nädalas ……

b) Mees, ei joo õlut ……

c) Naine, ei joo õlut ……

d) Naine, joob 1-5 pudelit õlut nädalas; ……

Mida näitavad järgmiste mõjude hinnangud (mida nad sisuliselt näitavad):

”1-5 pudelit nädalas” …….

”mees” …….

”1-5 pudelit nädalas : mees …….

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)
```

Mida võiks iseloomustada see teine joonisele lisatud sirge?

Joonisele vaadates – kas märkad, mis on viga mudelil, mis õnnelikkust kokaiini tarbimise järgi prognoosib?

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?