Lineaarsed mudelid
4. praktikum

Paremini saab alati!

Agronoom Ahti eksperimenteeris põllul – lisas põllu kündmise ajal mullale praegu veel salastatud lisaainet *x* (erinevates kogustes) ja uuris, kas aine lisamine mõjutab saagikust (tunnus *y*). Kuna pärisandmed on jätkuvalt veel salastatud (seniks kuni lisaaine *x* patenteerimisprotsess käib) peame hakkama saama ja analüüse tegema kasutades kunstlikku abivahendit ehk genereeritud andmestikku:

```
set.seed(19)
x=1:10
y=2+0.85*x+rnorm(10, sd=7)
m1=lm(y~x)
summary(m1)
```

Näeme oma suureks kurvastuseks, et olemasolevate andmete põhjal lisaaine *x* mõju tõestada ei õnnestu. Õnneks tuli Ahtile hea mõte enne regressioonanalüüsi tegemist vaatlusandmeid siluda. Ta asendab vaatluse y_i vaatlusega y_i* mis on leitud järgmisel viisil:

$$y_i^* = (y_{i-1} + y_i + y_{i+1})/3$$

Antud valemi rakendamisel tekivad muidugi raskused esimese ja viimase vaatlusega (peaksime ju y_1^* leidmiseks teadma vaatlust y_0, mida meil pole kusagilt võtta ja vaatluse y_n^* leidmiseks läheks tarvis vaatlust y_{n+1} millega on sama häda...). No jätame siis esimese ja viimase vaatluse muutmata.

Uuri ja proovi aru saada mida teevad järgmised käsud:

```
abi=data.frame(vaatlus=y, jargmine=c(y[-1], NA), eelmine=c(NA,y[-10]))
yuus=rowSums(abi)/3

yuus[1]=y[1]; yuus[10]=y[10]

data.frame(abi, yuus)
```

Kasutades transformeeritud tunnust `yuus` proovis Ahti uuesti hinnata regressioonmudelit:

```
m2=lm(yuus~x)
summary(m2)
```

Oma rõõmuks leidis Ahti järgmist:

a) regressioonsirge tõusu hinnang oli muutunud täpsemaks (standardviga oli pisem)
b) tunnuse *x* mõju osutus nüüd statistiliselt oluliseks (on väiksem kui 0,05)!
c) Mudeli headust näitav determinatsioonikordaja R^2 on muutunud märksa suuremaks – uus mudel on võimeline kirjeldama ära hoopis suurema osa uuritava tunnuse hajuvusest!

Selgita, kuidas on see võimalik Gauss-Markovi teoreemi valguses?

- Kas Gauss-Markovi teoreemi saab antud näite puhul kasutada või ei saa?
- Miks BLUE-hinnang ei osutunud täpsemaks?

Mõtiskle nähtud tulemuste valguses järgmiste küsimuste üle:

Kumb meetod andis regressioonsirge tõusule ikkagi täpsema hinnangu?

Vaata mõlema meetodi abil hinnatud regressioonsirge tõuse.

Kõik mudeli parameetrite hinnangud (ühe pika vektorina) saad kätte käsuga:

```
coef(m1)
```

Regressioonsirge tõusu (teise parameetri) hinnangu saab välja noppida käsuga

```
coef(m1)[2]
```

Hinnangu standardvea hinnangu saad soovi korral kätte aga nii:

```
coef(summary(m1))[2,2]
```

Simuleeri 1000 eksperimentaatori tulemusi (ilma set.seed-käsku kasutamata!) ja salvesta kõigi simuleeritud andmestike korral mõlemal viisil leitud hinnangud regressioonsirge tõusule. Lisaks salvesta ka hinnangute standardvigade hinnangud. Hinda nende 1000 eksperimendi põhjal järgmiste näitajate väärtused.

1. Kontrolli, kas mõlema mudeli poolt antavad hinnangud regressioonsirge tõusule on nihketa hinnangud. Milline on sinu otsus:

   ……

2. Leia esimese mudeli korral hinnangu (regressioonsirge tõusu) dispersioon

   .....

3. Leia teise mudeli korral hinnangu (regressioonsirge tõusu) dispersioon

   .....

Kumb mudelitest siis ikkagi andis täpsema hinnangu sind huvitavale parameetrile?

……..……..

4. Leia esimese mudeli korral dispersiooni hinnangute keskmine (sirge tõusu hinnangu dispersiooni hinnangute keskmine üle kõigi eksperimentide):

   ……

5. Leia teise mudeli korral dispersiooni hinnangute keskmine

   ……

Kommenteeri saadud tulemusi. Milles on probleem?

Jätkame tegelemist põllumajandusega. Loeme sisse järgmise pisiandmestiku:

```
andmed=read.table(url("http://www.ms.ut.ee/mart/linmud2021/saagikus.csv"),
      header=TRUE)
head(andmed)
```

Tegemist on eri sorti põlluviljade saagikusega erinevatel aastatel (saak pindalaühiku kohta).

```
m1=lm(saak~factor(sort))
summary(m1)
```

Mida näitab vabaliige, mida factor(sort)D järel olev parameeter?
Võrdle:

```
mean(saak[sort=="A"])
mean(saak[sort=="D"])
```

Milline sortidest annab kõige halvemini saaki?

Lisame mudelile veidi „uhkust“:

```
m2=lm(saak~factor(sort)+factor(aasta))
summary(m2)
```

Mida näitab vabaliige, mida `factor(sort)D` järel olev parameeter?
Milline sortidest annab (hinnanguliselt) kõige halvemini saaki?

Mida annavad (sisuliselt) meile järgmised käsud:

```
library(gmodels)
estimable( m2,c(1, 0,0,1, 0,0,0), conf.int=0.95)

Vastus: ……..

estimable( m2,c(1, 0,0,1, 1/4,1/4,1/4))

Vastus: ……..

estimable( m2,c(1, 0,0,0, 1/4,1/4,1/4))

Vastus: ……..
```

Millise estimable-käsu abil saaksime mudelit m2 kasutades kätte sellesama sort-A saagikuste keskmise mida näeme kas mean käsu või t.test- käsu abil:

```
mean(saak[sort=="A"])
t.test(saak[sort=="A"])
```

Vastus: sama keskmiseni jõuaksin kasutades järgmist estimable-käsku:

estimable(m2, ……..)

Milline kirjeldatud variantidest (mean-käsk või mõni ülaltoodud estimable-käskudest) võiks kõige paremini iseloomustada sort A saagikust? Miks just see variant? Miks nad üldse annavad erinevaid hinnanguid sort A saagikusele?

……..
……..
……..
……..

Lisame mudelile keerukust. Kas üks sort on kõigil aastatel teisest sordist ühtemoodi parem? Või mõni aasta sobib ühele sordile, teine aasta pigem teisele sordile?

```
m3=lm(saak~factor(sort)+factor(aasta)+factor(sort)*factor(aasta))
summary(m3)
drop1(m3, test="F")
```

Märkus: drop1 –käsk proovib mudelist ükshaaval mõjusid eemaldada. Kontrollitavaks nullhüpoteesiks on väide, et ilma mingi mõjuta mudel prognoosib uuritavat tunnust sama hästi kui keerukam, vastavat mõju sisaldav mudel. Suur p-väärtus näitab seega, et me võiksime vastava mõju mudelist eemaldada, ilma et meie mudeli prognoosivõime (tõestatavalt) kehvemaks muutus.

Milline on sinu otsus ülaltoodud küsimusel? Kas mõni aasta sobib paremini mõnele sordile ja mõni aasta hoopis mõnele teisele sordile?

Põllumees soovib võrrelda sorte A ja B. Kumba neist sortidest oleks targem kasvatada ja kui palju võiks ta saagikuses võita? Millisel viisil peaksime soovitud võrdlust läbi viima? Leia põllumehe jaoks kõige mõistlikum võrdlus kasutades estimable-käsku.

```
estimable( m3, c(……,……,……,……,……,……,……, ……,……,……,……,……,……,……,…… ))
```