Biomeetria

2. praktikum

Loeme sisse andmestiku kalad:

```
print(
 load(url("http://www.ms.ut.ee/mart/biomeetria2015/kalamees.RData"))
)
```

Lühikeste tunnuste nimede kasutamiseks anna käsk

```
attach(kalad)
```

Andmestiku lühikirjeldus:
Soomes Tampere lähedal asuvast Laenelmavesi järvest püüti 159 kala. Püütud kalad on pärit 7 liigist.
Mõõdetud tunnuste kirjeldused:

Species on kodeeritud tunnus kalaliikidest:

1 - latikas	2 - siig	3 - särg
4 - linask	5 - tint	6 - haug
7 - ahven		

Weight on kala kaal grammides
Length3 on kala pikkus ninast saba tipuni sentimeetrites.
Height on maksimaalne kõrgus, mis antud protsendina *Length3*-st.
Width on maksimaalne paksus, mis on samuti
Sex on kala sugu: 0-emane; 1-isane

Regressioonanalüüs

Anna-Liisa tegi latikate elu kirjeldavat loodusfilmi. Peale veealuste kaadrite filmimist hakati tavainimese jaoks sobivaid selgitusi lisama. Telekanali esindaja, vana kalamees Kalamees soovis tungivalt, et ühe eriti uhke latika kaalu ka selgitavas tekstis mainitakse. Kuna filmitähest latikas oli juba ammu tont teab kuhu ujunud (või nahka pandud), tuli kala kaal kuidagi kaudsel viisil välja nuputada. Õnneks oli filmitud kaadrite pealt võimalik mõõta latika pikkust. Anna-Liisa otsustaski kaalu prognoosida latika pikkuse järgi, kasutades kaalu prognoosimiseks kasutatava mudeli loomiseks Laenelmavesi järvest püütud latikate andmeid.

Mudeli loomine (latikate jaoks on tunnus *Species* väärtus 1):

Hindame mudeli kasutades vaid latikate andmeid

```
> mudel=lm(Weight~Length3, data=kalad[Species==1,])
```

funktsioontunnus (sõltuv tunnus, *dependent variable*)

argumenttunnus (sõltumatu tunnus, *independent variable*)

Salvestame hinnatud regressioonmudeli

```
> mudel
Call:
lm(formula = Weight ~ Length3, data = kalad[Species == 1, ])
Coefficients:
(Intercept)      Length3
   -1194.40        47.37
```

Kaalu prognoosiv mudel on

Kaal = -1194,4 + 47,37*Pikkus + ε

Seega 30 cm pikkuse latika kaaluks prognoosib mudel -1194,4+47,37*30 = 226,7g, ehk teisisõnu: 30cm pikkuste latikate keskmine kaal on 226,7g.

```
> summary(mudel)

Call:
lm(formula = Weight ~ Length3, data = kalad[Species == 1, ])

Residuals:
     Min       1Q   Median       3Q      Max
-101.671  -29.643   -8.777   28.855  176.486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1194.395     89.815  -13.30 8.29e-15 ***
Length3        47.369      2.328   20.34  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 56.45 on 33 degrees of freedom
Multiple R-Squared: 0.9261,     Adjusted R-squared: 0.9239
F-statistic: 413.8 on 1 and 33 DF,  p-value: < 2.2e-16
```

Prognoosivead:
50% prognoosivigadest ε jääb vahemikku -29,6 ... +28,9

Testitakse, kas kala pikkuse muutudes ikka kala kaal ka muutub (kas regressioonmudelis pikkuse ees olev kordaja erineb nullist)

Parandatud determinatsioonikordaja R_{adj}^2=0,9239. Mudel prognoosib kaalu küllaltki täpselt. Antud järve kalade puhul on pikkuse abil võimalik kirjeldada umbes 92% kaalu varieeruvusest.

Kasutades leitud mudelit võime prognoosida kalade kaale:

Hinnang pikkusega 30cm latikate kaalu keskväärtusele (30cm pika latika kaalu prognoos):

```
> predict(mudel, data.frame(Length3=30))
       1
226.6678
```

30 cm pika latika kaaluks prognoositi 226,7 g

Millise pikkuse jaoks prognoose soovime. Märkus: tunnuse nimi peab langema kokku mudeli hindamisel kasutatud x-tunnuse (prognoosimiseks kasutatava tunnuse) nimega!

Kui täpselt me ikkagi teame 30cm pikkuste latikate kaalu keskväärtust? Leiame 95%-usaldusintervalli 30cm pikkuste latikate kaalude keskväärtusele:

```
>  predict(mudel, data.frame(Length3=30), interval="confidence")
       fit      lwr      upr
1 226.6678 182.5858 270.7497
```

30 cm pika latika kaalu prognoos

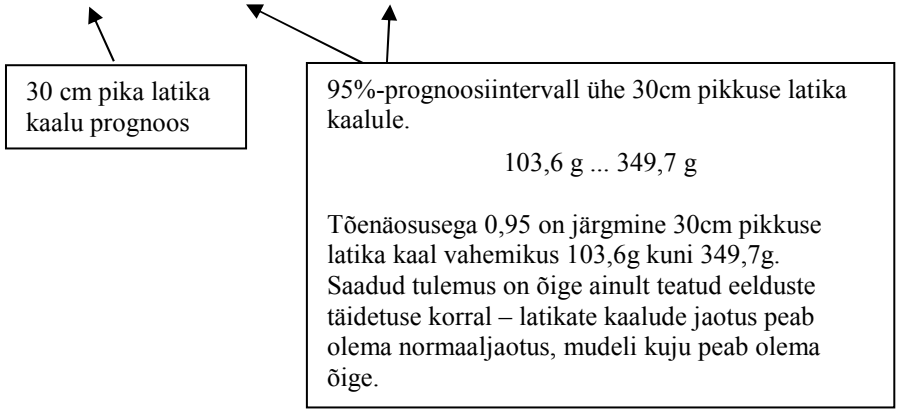
95%-usaldusintervall 30cm pikkuste latikate kaalu keskväärtusele:

182,6 g ... 270,4 g

Eeldab, et mudeli kuju on õige – st. keskväärtus peab sõltuma pikkusest lineaarselt

Kui täpselt me ikkagi suudame prognoosida ühe 30cm pikkuse latika kaalu?

```
> predict(mudel, data.frame(Length3=30), interval="prediction")
       fit      lwr      upr
1 226.6678 103.6458 349.6897
```



Saadud tulemusi – regressioonmudeli sirget, usaldusintervalli ja prognoosiintervalli saab esitada ka graafiliselt.

Järgnevad käsud on tungivalt soovitatav eelnevalt scripti-redaktoris valmis kirjutada ja alles siis käivitada.

Kõigepealt joonistame latikate pikkuste ja kaalude hajuvusgraafiku:

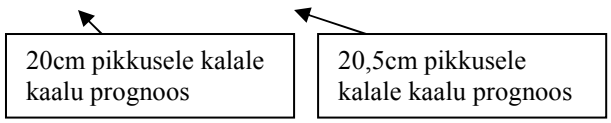
```
plot(Length3[Species==1], Weight[Species==1],
      main="Latika kaalu prognoosimine", xlab="Kala pikkus (cm)",
      ylab="Kala kaal (g)")
```

Lisame saadud joonisele regressioonsirge. Seda on võimalik teha kahel moel. Lihtsaim viis (mida aga ei saa kasutada keerukamate mudelite puhul) oleks käsuga `abline(mudel)`. Teine võimalus (ja ehk veidi paremini üldistatav ka keerukamatele juhtudele) on kasutada `predict` ja `lines` -käske.

Joonisele on kantud kalade pikkused vahemikus 30cm...45cm. Prognoosime veidi laiemas vahemikus (20cm...50cm) hästi erineva pikkustega kalade kaale:

```
xx=seq(20, 50, length=61)
xx
prognoos=predict(mudel, data.frame(Length3=xx))
prognoos
```

```
         1          2    ...
-247.01992 -223.33554    ...
```



Vaadates lühikeste latikate kaalude prognoose peame tõdema – statistiliste mudelite prognoosid on usaldusväärsed vaid nende x-tunnuse väärtuste piirkondades, kust meil on olemas vaatluseid. Käsuga `range(Length3[Species==1])` näeme, et kõige pisem latikas meie andmestikus on 30cm pikk. Mis lühemate kalade kaaludega toimub, seda me ju tegelikult ei tea. Ning ei saa teada ka statistiline mudel (kui kalade-valim oleks esinduslik, siis ei tohiks meile lihtsalt üldsegi ette juhtuda niivõrd väikest latikat...).

Lisame leitud prognoosid varem tehtud joonisele (loodetavasti pole sa graafiku-akent – kuhu oli joonistatud hajuvusdiagramm – vahepeal sulgenud?):

Saadud graafikule võime lisada ka usaldus- ja prognoosiintervalli. Prognooside leidmisel võisime nõuda ka prognoosi- või usaldusintervalli arvutamist:

```
prognoos=predict(mudel, data.frame(Length3=xx),
                                    interval= "prediction")
prognoos
```

```
          fit          lwr          upr
1  -247.01992 -392.375921 -101.663928
2  -223.33554 -367.287154  -79.383926
      …            …            …
```

20,5cm pikkuste latika kaaluks prognoosib meie regressioonmudel olevat -223,3g. Mudeli järgi peaks sellise pikkusega kala kaal tõenäosusega 0,95 olema vahemikus -367g..-79g.
Ilmselgelt on midagi väga valesti.

Oleme saanud iga *xx*-i väärtuse jaoks prognoosi kala kaalule (20cm pikkuste kalade keskmise kaalu, 20,5cm pikkuste kalade keskmise kaalu jne) koos prognoosiintervalliga. Kanname saadud prognoosiintervallid joonisele (alumised prognoosiintervallid on saadud tulemuste maatriksi 2. tulbas, ülemised 3. tulbas):

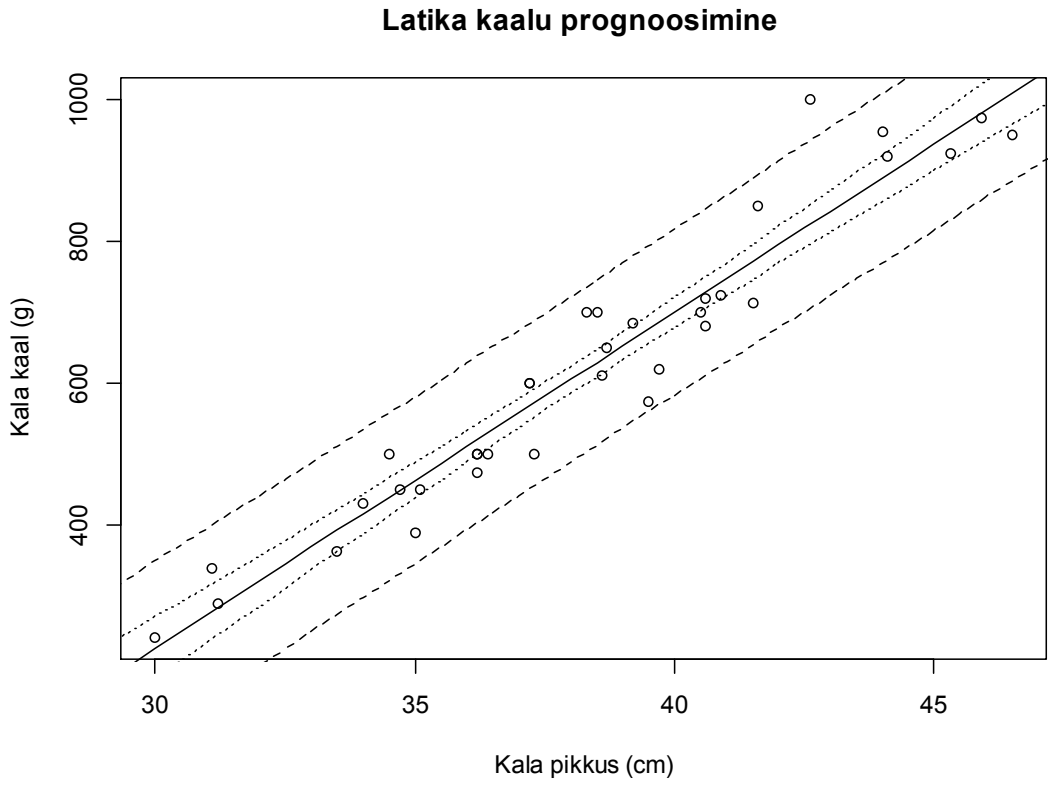
```
lines(xx, prognoos[,2], lty=2)
lines(xx, prognoos[,3], lty=2)
```

Joonte joonistamisel on võimalik määrata joone tüüp (lisaparameetriga `lty=`), näiteks:

lty=1	tavaline joon
lty=2	kriipsjoon (- - - -)
lty=3	punktiirjoon (…..)
….	….

Usaldusintervalli lisamine:

```
prognoos=predict(mudel, data.frame(Length3=xx),
                                    interval="confidence")
lines(xx, prognoos[,2], lty=3)
lines(xx, prognoos[,3], lty=3)
```



Ülesanded

1. Filmilindile jäädvustatud latika pikkus oli 44cm. Prognoosi antud latika kaalu. Leia:

Prognoos latika kaalule: ..... g

Prognoosi täpsust iseloomustav 95%-prognoosiintervall:

.....g - .....g

2. Soovitakse saada sarnast joonist ka ahvenate (*species*=7) jaoks. Hinda lineaarne mudel ahvenate jaoks ja tee joonis. Kommenteeri saadud tulemust.

Eksperiment.

Statistiliste meetodite töö mõistmiseks on vahel hea, kui tõde (tegelikkus, uuritav populatsioon) on teada. Siis võime võtta valimikese ja uurida kuivõrd hästi suudame meie statistika abiga tõele jälile jõuda (või vahel peame paraku ka tunnistama, et õiget vastust olemasolevate andmete põhjal kätte saada ei õnnestu). Järgneva näite teeme olukorras kus tõde – andmeid genereeriv mehhanism – on õppejõule teada. Siis on võimalik ka vaadelda millised tulemused ja kui palju võivad viltu minna...

Anname järgmise mõistatusliku käsu:

```
source(url("http://www.ms.ut.ee/mart/biomeetria2015/katse.R"))
```

Peale selle käsu andmist tekib R-le juurde käsk katse, mis võimaldab meil läbi viia ühte eksperimenti ja näha selle eksperimendi tulemust:

```
> katse(x=3)
  x        y
1 3 98.97949
```

Viime läbi eksperimendi olukorras kus tunnuse *x* väärtus on 3. Saame mingi katsetulemuse – tunnuse y väärtuse. Teie poolt saadav eksperimendi tulemus võib tulla mõnevõrra erinev siintoodust (isegi kui te kõik mulda väetaksite ühtemoodi, siis igaühel teist kasvab ju tulemuseks erineva suurusega taim...)

Selliseid eksperimente võime teostada väga erinevate x-tunnuse väärtustega. Samuti võime ühe ja sellesama x-tunnuse väärtust kasutades läbi viia mitmeid (kordus)eksperimente. Korrakem näiteks eksperimenti 10 korda x-tunnuse väärtusega 1, 10 korda x-tunnuse väärtusega 2 jne:

```
# Milliste x-tunnuse väärtustega kavatseme eksperimente läbi viia
# (kokku kavatseme teostada 100 eksperimenti):

x=rep(1:10, each=10)
x


# Teeme eksperimendid ja vaatame saadud andmestikku:

andmed=katse(x)
andmed
```

Saadud andmestikku võime iseloomustada joonise abil, võime ka hinnata regressioonmudeli

$$y = c_0 + c_1 * x + e$$

ja kirjeldada tunnuste *x* ja *y* vahelise seose tugevust.

```
m1=lm(y~x, data=andmed)
summary(m1)
plot(andmed$x, andmed$y); abline(m1)
```

Mida oskad tulemustest välja lugeda? Kas seos *x*-tunnuse ja *y*-tunnuse vahel eksisteerib (kas seose olemasolu on tõestatav)? Kas tunnuse *x*-abil on võimalik prognoosida tunnuse *y* väärtuseid? Kui hästi? Mida oskad veel tulemustest välja lugeda?

Korda eksperimenti täpselt samal moel 3 korda (iga kord tee 100 katset, täpselt samasuguseid *x*-tunnuse väärtuseid kasutades) ja märgi üles järgmised näitajad (usaldusintervalle parameetritele saad leida `confint`-käsu abil, näiteks mudeli `m1` parameetritele usaldusintervallide leidmiseks anna käsk `confint(m1)`)

Andmestik (eksperiment)	c_1 hinnang	95%-UI c_1-le	c_1 hinnangu standardvea hinnang	R^2 (kohandatud)	jääkide standardhälve
1.	.....	.....	.....	.....	.....
2.	.....	.....	.....	.....	.....
3.	.....	.....	.....	.....	.....

Kuna õppejõud teab andmeid tekitavat protsessi tean ma ka hinnatavate näitajate tegelikke väärtuseid:

Kordaja c_1 tegelik väärtus: 1
c_1 hinnangu standardviga: 0.1435481
Determinatsioonikordaja R^2 : 0.3267327
Mudeli jääkide standardhälve: 4.123

Vaata kui lähedale tõele – tegelikele väärtustele – sa suutsid jõuda? Kas õige kordaja c1 väärtus jäi sul kõigil kolmel katsel usaldusintervalli?

Õigeid väärtuseid (või nendega väga lähedasi numbreid) näeksid sa siis, kui teeksid hiigelpalju vaatluseid/katseid. Näiteks iga x-tunnuse väärtusega kordaksid eksperimenti 100000 korda:

```
x=rep(1:10, each=100000)
andmed=katse(x)
summary(lm(y~x, data=andmed))
```

Vaata kas näed nüüd (väga suure vaatluste arvu korral) ülaltoodud näitajatega väga sarnaseid tulemusi (erandiks standardviga – miks)?

Proovime nüüd aga muuta eksperimendi läbiviimiseks kasutatavate x-tunnuste väärtuseid (x-tunnsue väärtuste jaotust).

Variant 1

```
x=rep(c(1:5, 31:35), each=10)
andmed=katse(x)
summary(lm(y~x, data=andmed))
```

Variant 2

```
x=rep(1:5, each=20)
andmed=katse(x)
summary(lm(y~x, data=andmed))
```

Mis juhtub jääkide standardvea hinnanguga? Mis juhtud determinatsioonikordaja hinnanguga? Miks?