Biomeetria bioloogidele

6. praktikum

I osa. Mudeli valik, põhjuslik vs prognoosiv mudel.

Järgnevalt proovime aru saada, kui kerge või keeruline on vaatlusandmete (või kontrollitud eksperimendi abil) kirjeldada põhjuslikke mõjusid – mis ikkagi mida mõjutab, kui tugevasti ja mis suunas.

Loeme sisse abifunktsiooni mis meie eest eksperimente teeb ja andmeid kogub:

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

Peale ülaltoodud käsu andmist R-le tekib R-i juurde uus funktsioon nimega v6tavalim.

Võtamegi selle uue funktsiooni abil valimi, uurime 1000-t inimest (või uurimisobjekti):

```
andmestik=v6tavalim(1000)
```

Saadud andmestikus on 6 tunnust. Meid huvitab eelkõige see, kuidas tunnus X võiks mõjutada tunnuse Y väärtust (teame lisaks, et kui mõju eksisteerib, siis X-tunnuse mõju Y-le on lineaarne).

Proovime seda mõju kirjeldada mitme mudeli abil:

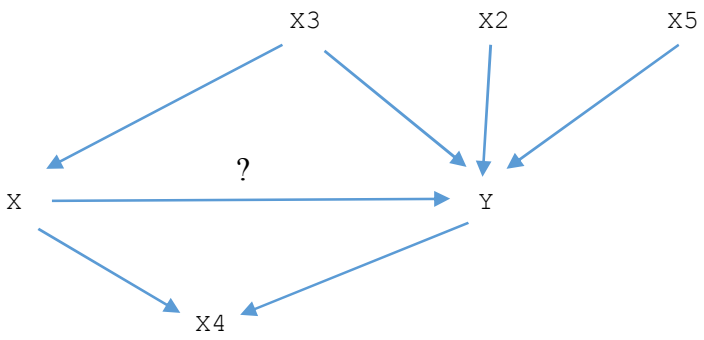
```
mudel1=lm(Y~X, data=andmestik)
mudel2=lm(Y~X+X2, data=andmestik)
mudel3=lm(Y~X+X3, data=andmestik)
mudel4=lm(Y~X+X4, data=andmestik)
mudel5=lm(Y~X+X5, data=andmestik)
mudel6=lm(Y~X+X3+X5, data=andmestik)
mudel7=lm(Y~X+X2+X3+X5, data=andmestik)
mudel8=lm(Y~X+X2+X3+X4+X5, data=andmestik)

AIC(mudel1, mudel2, mudel3, mudel4,
       mudel5, mudel6, mudel7, mudel8)
```

Mudel	X-tunnuse ees olev kordaja
mudel1	…….
mudel2	…….
mudel3	…….
mudel4	…….
mudel5	…….
mudel6	…….
mudel7	…….
mudel8	…….

Millised on hinnangud X-tunnuse mõjule? Kas need tulevad kõigis mudelites samasugused või erinevad? Millise mudeli peaksime kuulutama parimaks mudeliks AIC-väärtuse põhjal otsustades? Millist mudelit usaldaksid sina?

Kas sinu otsus muutuks, kui teaksid kuidas tunnused üksteist mõjutavad? Millisest mudelist saaksid hinnata *X*-tunnuse põhjuslikku mõju siis, kui tunnustevahelisi põhjuslikke mõjusid kirjeldaks järgmine skeem:



Võime ka veenduda, et AIC-järgi välja valitud parim mudel (enamikel teist peaks parimaks mudeliks osutuma *mudel8*) prognoosib uusi vaatluseid täpsemalt kui õige mudel (mis kirjeldab tunnuse Y väärtuste tekkemehhanismi). Selles veendumiseks võime võtta uue valimi ja näha, et uute vaatluste korral on ajalooliste andmete põhjal hinnatud mudeli *mudel8* prognoosid täpsemad kui tunnuste tegelikku tekkemehhanismi õigesti kirjeldaval mudelil *mudel7*:

```
uued=v6tavalim(2000)

# Leiame prognoosid mõlema mudeliga
prognoos_valemudel=predict(mudel8, uued)
prognoos_6igemudel=predict(mudel7, uued)

# Leiame prognoosivigade ruutude keskmise
#   (mida suurem, seda ebatäpsemad prognoosid)
mean((uued$Y-prognoos_valemudel)**2)
mean((uued$Y-prognoos_6igemudel)**2)
```

Tõsi, tunnuse väärtuste tekkemehhanismi kirjeldaval nn põhjuslikul mudelil on siiski omad eelised. Eelkõige siis, kui soovime kirjeldada sekkumise võimalikku mõju.

Vaatame milliseid andmeid (eksperimenditulemusi) me näeksime, kui kunstlikult muudaksime X-tunnuse väärtuse 10-ks:

```
uued2=v6tavalim(X=rep(10, 10000))
mean(uued2$Y)
```

Milline tuleks Y-tunnuse keskmine siis, kui muudaksime X-tunnuse väärtuse hoopis 20-ks?

```
uued3=v6tavalim(X=rep(20, 10000))
mean(uued3$Y)
```

Märksa lihtsam oleks põhjusliku mõju tuvastada ja kirjeldada eksperimendi abil, kus tunnuse *X*-väärtused on meie kontrolli all (tunnust *X* ei mõjuta ükski uurimisalustest tunnustest):

```
andmed2=v6tavalim(X=seq(30, 80, 0.5))
andmed2
attach(andmed2)
```

Vaata nüüd erinevaid mudeleid – ka lihtsaim mudel annab õige ettekujutuse tunnuse *X* mõjust. Ainult mudelid, mis sisaldavad tunnust X4 annavad eksitava väärtuse X-tunnuse ees olevale kordajale – kuna tunnust *X*4 mõjutab nii tunnus *X* kui ka tunnus *Y*:

```
summary(lm(Y~X))
summary(lm(Y~X+X2))
summary(lm(Y~X+X4))
summary(lm(Y~X+X2+X3+X4+X5))
```

II osa. Mudelid diskreetsele tunnusele I. Poissoni regressioon

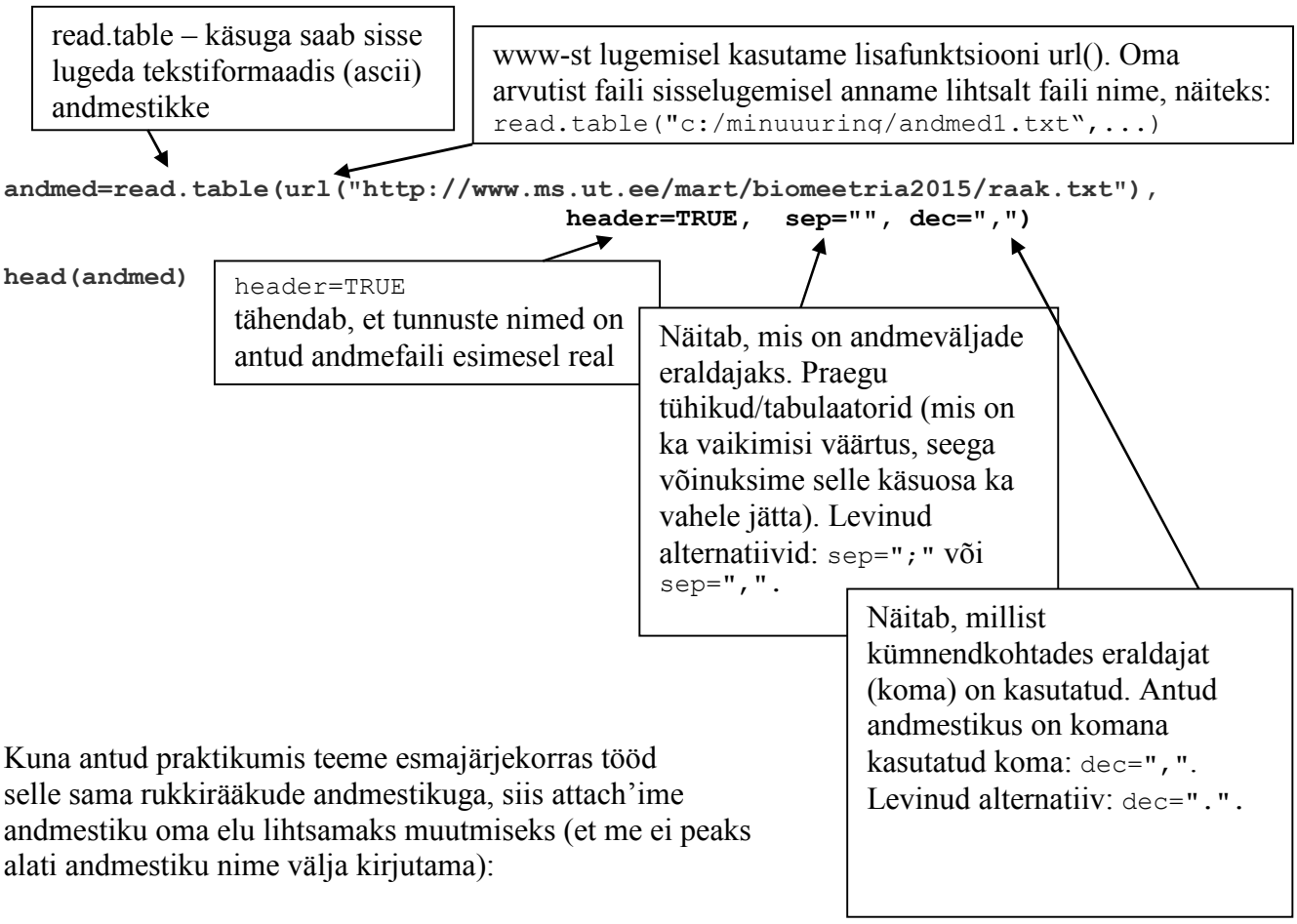
Praktikumis kasutame rukkirääga loendusandmeid (andmed pärinevad Riho Marja uurimistööst). Andmestiku kirjelduse võid leida aadressilt:

http://www.ms.ut.ee/mart/biomeetria2015/rukkiraak.pdf

Andmestik ise paikneb aadressil

http://www.ms.ut.ee/mart/biomeetria2015/raak.txt

Andmestik on tavalise tekstifaili kujul ja tuleb R-i importida. Andmestiku sisselugemiseks võime anda järgmise käsu:



Kuna antud praktikumis teeme esmajärjekorras tööd selle sama rukkirääkude andmestikuga, siis attach'ime andmestiku oma elu lihtsamaks muutmiseks (et me ei peaks alati andmestiku nime välja kirjutama):

```
attach(andmed)
```

Meid huvitab eelkõige mudel rukkirääkude arvule, soovime kirjeldada, millest sõltub tunnus raak. Vaatame korra seda tunnust:

```
table(raak)
```

ja joonistame ka graafiku (muuseas, mis on neil kahel barplot-käsul vahet ja miks tuleks ühte neist teisele eelistada?)

```
barplot(table(raak), col="gold", border="gold4")

barplot(table(factor(raak, levels=0:7)),
      col="gold", border="gold4")
```

Mudeli loomist võiksime alustada ühe kõige olulisema näitaja sissetoomisega – mida suurem on maatüki pindala, seda rohkem me võiksime antud maatükil rukkirääke näha:

```
plot(pindala, raak)
```

Loome oma esimese mudeli:

```
mudel1=glm(raak~1, family=poisson(), offset=log(pindala))
summary(mudel1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.16475    0.04746  -87.76   <2e-16 ***
```

Hinnatud mudel on kujul (kõik kirjaviisid samaväärsed):

$$\log(\mathrm{E}(raak)) = -4.16475 + \log(pindala)$$
$$\mathrm{E}(raak) = \exp(-4.16475 + \log(pindala))$$
$$\mathrm{E}(raak) = \exp(-4.16475)\ pindala$$
$$\mathrm{E}(raak) = 0.0155336\ pindala$$

Hinnatud mudel väidab, et keskmiselt on meil

```
exp(-4.16475)= 0.0155336
```

rukkirääku hektari maa kohta. Soovides leida oodatavat rukkirääkude arvu 100 hektarilisel põllul, võime seda teha kas ise arvutades:

```
exp(-4.16475)*100 =  1.55336
```

või võime vastava arvutuse lasta teha R'il:

```
> predict(mudel1, data.frame(pindala=100), type="response")
[1] 1.553361
```

Joonistame ka huvi pärast graafiku, mis iseloomustaks seost pindala ja rukkirääkude arvu vahel:

```
x=seq(0.01, 110, length=200)
y=predict(mudel1, data.frame(pindala=x), type="response")

plot(pindala, raak, col=rgb(0.1, 0.1, 0.1, 0.2), pch=20, cex=2)
lines(x,y, col="red", lwd=2)
```

Leidsime eelnevalt, et hinnanguliselt tuleb 0.0155 rukkirääku hektari kohta. Hinnangutele on aga sobilik lisada usalduspiirid.

1. Leiame usalduspiirid hinnatud parameetri (parameetrite) tegelikule äärtusele (ligikaudne 95%-usaldusvahemik vabaliikme tegelikule väärtusele):

```
confint(mudel1)
```

```
                    2.5 %    97.5 %
-4.259232 -4.073156
```

Oleme leidnud usalduspiirid mudeli vabaliikmele ehk suurusele log(EY), 95%-kindlusega asub mainitud suurus vahemikus (-4.259...-4.073).

2. Leidmaks usalduspiire keskmisele rukkirääkude arvule hektari kohta ehk EY-le, tuleb varemtoodud vahemiku piire teisendada eksponentfunktsiooni abil:

```
> exp(confint(mudel1))
                                2.5 %     97.5 %
                            0.01413316 0.01702359
```

Seega 95%-usaldusintervall keskmisele rukkirääkude arvule hektari kohta oleks (0.0141...0.0170).

Võid võrrelda leitud usalduspiire klassikaliste usalduspiiridega keskväärtusele:

```
t.test(raak/pindala)
```

Milline meetod võimaldab rukkirääkude arvu keskväärtust täpsemalt määrata?

Liigume kiiresti edasi keerukama mudeli suunas. Toome mängu ka tunnuse hairing (kas rukkirääkude lugemise ajal toimus midagi põllul või põllu lähiümbruses (kas traktor mürises põllul vms):

```
table(hairing)
```

Hindame mudeli, kus keskmine rukkirääkude arv põllul võib sõltuda häiringu olemasolust:

```
mudel2 = glm(raak~factor(hairing), offset=log(pindala), family=poisson())
summary(mudel2)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -5.0681     0.2236 -22.665  < 2e-16 ***
factor(hairing)hairingutpole   0.9750     0.2288   4.261 2.03e-05 ***
```

Hinnatud mudel:

$$\log(E(raak)) = -5.0681+0.975*I_{hairingutpole} + \log(pindala)$$
$$E(raak) = \exp(-5.0681+0.975*I_{hairingutpole} + \log(pindala))$$
$$E(raak) = 0.006294368..* 2.65^{I_{hairingutpole}} * pindala$$

Häiringuga ala (võrdlusnivoo): keskmiselt 0,00629 rukkirääku hektari kohta
Häiringuta ala: keskmiselt 0,00629*2,65 rukkirääku hektari kohta ehk 2,65 korda rohkem linnukesi.

Ülesanne

Hinda mudel (mudel3), kus rukkirääkude arvukus võib sõltuda põllule makstavast toetuse tüübist (tunnus toetustyyp). Interpreteeri mudeli parameetreid.

Proovime järgnevalt hinnata mudeli, kus rukkirääkude arvukus võiks sõltuda nii häiringu olemasolust/mitteolemasolust kui ka põllumajandustoetuse tüübist samal ajal. Heidame esmalt pilgu peale rukkirääkude arvukuse prognoosimisel kasutatavatele tunnustele:

```
table(hairing, toetustyyp)
```

või, sama informatsioon veidi "uhkemalt" esitatul (tulpade laius iseloomustab antud toetustüübi rohkust meie vaatlusandmetes):

```
windows(width=10, height=6)
barplot(100*prop.table(table(hairing, toetustyyp),2),
      width=prop.table(table(toetustyyp)), col=c("darkred","green4"),
      xlab="Põllumajandustoetuse tüüp", ylab="Häirinuga alade %",
      xlim=c(0,1.5))
legend(1.25, 90, c("rahulik", "häiring"), fill=c("green4","darkred"))
```

Lisame need kaks tunnust rukkirääkude arvukust prognoosivasse mudelisse:

```
mudel4=glm(raak~factor(toetustyyp)+factor(hairing),
                        offset=log(pindala), family=poisson())

drop1(mudel4, test="Chisq")
```

```
                    Df Deviance    AIC     LRT   Pr(Chi)
<none>                   1499.2 2234.8
factor(toetustyyp)   3   1607.5 2337.2 108.304 < 2.2e-16 ***
factor(hairing)      1   1527.8 2261.5  28.657 8.642e-08 ***
```

Näeme, et samasuurtel aladel, samas häiringuolekus (näiteks häiringuta olekus) on erinevate toetustüüpide korral rukkirääkude arvukus (tõestatavalt) erinev

Näeme, et samasuurtel aladel, sama toetustüübi korral (näiteks mahetoetust saaval alal) on häiringul tuvastatav mõju rukkirääkude arvukusele.

drop1 – käsk üritab mudelist kordamööda kõiki tunnuseid eemaldada ja testib, kas saame ilma ühe või teise tunnuseta ka hakkama. Näiteks drop1-käsu väljundis real factor(hairing) toodud p-väärtus on saadud alltoodud kahe mudeli võrdlemise abil:

```
mudel4=glm(raak~factor(toetustyyp)+factor(hairing),
                        offset=log(pindala), family=poisson())
mudel2=glm(raak~factor(toetustyyp),
                        offset=log(pindala), family=poisson())

anova(mudel4, mudel2, test="Chisq")
```

Märkus: kui kasutame drop1-käsku tavaliste lineaarsete mudelite korral (lm käsuga hinnatud), siis on targem kasutada p-väärtuse arvutamiseks hii-ruut testi asemel F-testi: drop1(lineaarnemudel, test="F").

Ülesanne:
Proovi, millise kahe mudeli võrdlemine anova-käsu abil annab sama tulemuse (olulisustõenäosuse ehk p-väärtuse), kui

Mudeli parameetrite hinnanguid näeme summary-käsu abil:

```
summary(mudel4)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -4.95225    0.25419 -19.482  < 2e-16 ***
factor(toetustyyp)ksm         -0.92263    0.15765  -5.852 4.85e-09 ***
factor(toetustyyp)mahe         0.03998    0.18827   0.212   0.8318
factor(toetustyyp)ypt          0.24906    0.13578   1.834   0.0666 .
factor(hairing)hairingutpole   1.04258    0.22927   4.547 5.43e-06 ***
```

Kui võrdleme kahte sama suurusega põldu, sama häiritusega (mõlemad kas häiritud olekus või mõlemad ilma häiringuta) siis mahepõllul on võrreldes kontrollgrupiga (toetustyyp="ei") rukkirääke keskmiselt exp(0.03998)= 1.04079 korda rohkem kui toetust mittesaaval põllul.

Näeme, et mahetoetusega aladel ja ilma keskkonnatoetuseta aladel võib rukkirääkude arvukus olla ka sama (kui räägime sama suurest maa-alast ja samast häirituse olekust)

Oodatav rukkirääkude arv 10-hektarilisel mahetoetust saaval maa-alal, kus pole äsja põllutöid/niitmist tehtud, oleks leitav järgmiselt:

```
exp(-4.95225+0.03998+1.04258+log(10))
```

või, predict-käsku kasutades:

```
predict(mudel4, data.frame(toetustyyp="mahe",
      hairing="hairingutpole", pindala=10), type="response")
```

Ülesanne

Leia mudeli prognoosid (oodatavad rukkirääkude arvud) ka järgmiste juhtude tarvis:

1. Toetustüüp="ksm", häiringuta, 10 ha: …….…….
2. Toetustüüp="ei", häiringuta, 10 ha: …….…….
3. Toetustüüp="ei", häiringuta, 20 ha: …….…….
4. Toetustüüp="mahe", häiringuga, 10 ha: …….…….

Usalduspiiride leidmise võimalustest

Mudeli parameetritele võisime usalduspiire leida käsu confint-abil:

```
confint(mudel4)
```

Paraku on saadud usaldusintervallidest vähe kasu, kui soovime teada seda, kui täpselt me siis ikkagi ühte- või teisttüüpi põldude keskmist rukkirääguriikkust teame. Sellisteks arvutusteks kasutasime varem (lineaarsete mudelite korral) predict-käsku koos lisaparameetriga interval=”confidence”, kuid paraku üldistatud lineaarsete mudelite korral ei ole hetkel veel R-is mainitud lisaparameetrit võimalik kasutada. Küll aga on võimalik predict-käsu abil tellida hinnangu standardvea hinnangut, mida teades on võimalik leida ligikaudset usaldusintervalli meid huvitavale keskväärtusele.

Leiame näiteks ligikaudse 95%-usaldusintervalli rukkirääkude keskmisele arvukusele häiringuta mahetoetust saavate 100ha suuruste põldude jaoks:

```
prog=predict(mudel4, data.frame(hairing="hairingutpole", toetustyyp="mahe",
            pindala=100), se.fit=TRUE, type="link")
```

```
prog
```

Palume standardvea hinnangut

Hinnatakse suurust log(E(raake))

Tegelik log(E(raak)) väärtus asub enamasti kahe standardvea (1,96 standardvea) kaugusel hinnangust. Seega aga paikneb tegelik keskväärtus E(raak) aga 95%-kindlusega vahemikus

```
exp(prog$fit-1.96*prog$se.fit)
exp(prog$fit+1.96*prog$se.fit)
```

Ülesanne

Leia ligikaudne 95%-usaldusintervall rukkirääkude arvu keskväärtusele häiringuta ksm-toetust saavate 50ha suuruste põldude jaoks:

Leitud 95%-usaldusintervall on:

…….