Praktikum 7
Mudelid diskreetsele tunnusele II. Üle- ja alahajuvusest

Esimeses üleandes tunneme huvi selle vastu, kas mutatsioonid leiavad aset kõikides geenides sama intensiivsusega või eksisteerib teatavat tüüpi geene, kus mutatsioonid leiavad aset harvemini (või sagedamini) kui teistes geenides. Näidisandmestikuna kasutame inimese 22. kromosoomis paiknevate geenide (DNA-lõigud, millele on antud ensembl-i poolt geeni ID) andmeid. Avatust mutatsioonidele mõõdame antud geenis leiduvate snippide (seni inimpopulatsioonis tuvastatud snippide) arvu abil (NB! Kuna kasutatud snippide andmebaas osutus weebi-teekonna jooksul kergelt kahjustatatuks, siis ei tasuks tulemusi kohe avaldama tõtata vaid võiks enne sarnased arvutused korrektsete andmetega uuesti üle teha).

Loeme sisse andmestiku:

```
load(url("http://www.ms.ut.ee/mart/biomeetria2015/geenid22.RData"))
head(geenid22)
attach(geenid22)
```

Geenide tüübid ja nende esinemissagedused inimese 22. kromosoomis:

```
table(tyyp)
```

Esimene mudel:

```
mudel1=glm(snippe~factor(tyyp), family=poisson())
```

Mis on häda esimesel mudelil? Kuidas saaks mudelit parandada?

Kas mõtlesid ise välja?

Muidugi võiks 2 korda pikemas geenis esineda ka kaks korda enam snippe…
Leiame iga geeni jaoks tema pikkuse:

```
pikkus=l6pp-algus
```

ja kasutame saadud tunnust mudelis offset-ina:

```
mudel2=glm(snippe~factor(tyyp), family=poisson(), offset=log(pikkus))
drop1(mudel2, test="Chisq")
```

Näeme oma meelerõõmuks, et geeni tüüp on statistiliselt ülioluline – tõepoolest esineb eri tüüpi geenides erinevas koguses snippe!

Vaatame hinnatud mudelit veidi lähemalt:

```
summary(mudel2)
```

Kuidas ennustab hinnatud mudel snippide oodatavat arvu nn ”antisense” –geenide korral?

E snippe = …….

Aga nende geenide korral mille tüübiks on ”IG_C_gene” (immunoglobuliini C-domeeni kodeerivad geenid?) ?

E snippe = …….

Kui testime, kas ”antisense” geenidel on keskmiselt rohkem snippe aluspaari kohta kui ”lincRNA” tüüpi geenil, siis milline tuleb selle testi p-väärtus:

p-väärtus: ……..

Muuseas, kuidas on lood üle- või alahajuvusega? Mida vaadates võiks tekkida kahtlus, et andmete hajuvus on liiga suur (tegemist võiks olla ülehajuvusega) ?

Kui kahtlustame ülehajuvust, siis võiksime proovida hinnata näiteks kvaasi-Poissoni mudelit:

```
mudel3=glm(snippe~factor(tyyp), family=quasipoisson(),
       offset=log(pikkus))
drop1(mudel3, test="F")
```

Pane tähele: Poissoni regressiooni korral on tavapärane kasutada (tõepärasuhte) hii-ruut testi, kvaasi-poissoni mudelite korral eelistatakse F-testi.

Näeme, et geeni tüüp jääb ikka statistiliselt oluliseks – eri tüüpi geenides esineb snippe erineva sagedusega. Vaatame aga mõnda võrdlust veidi lähemalt. Testime ka kvaasi-Poissoni mudelit kasutades, kas ”antisense” geenidel on keskmiselt rohkem snippe aluspaari kohta kui ”lincRNA” tüüpi geenil:

p-väärtus: ……..

Võrdle saadud tulemust varemsaadudga (Poissoni regressiooni abil saaduga). Mida märkad?

Kuidas on aga võrreldavad mudel3 ja mudel2 parameetrite hinnangud? Mitu korda enam on snippe (aluspaari kohta) ”IG_C_gene” – geenil (võrreldes ”antisense”-tüüpi geenidega) mudeli2 arvates ja mitu korda enam on neid mudeli3 arvates?

Snippide arv kuskil geenis võib muidugi sõltuda paljudest asjadest – näiteks ka sellest, kui palju esineb geenis GC-tähti. Proovime mudeli koostamisel arvestada ka geenide GC-sisaldust (ja uurigem, kuidas võiks GC-sisaldus snippide arvu mõjutada). Kuna võiks ja kuna ei tohiks mudelisse lisada geeni GC-sisaldust?

Muuseas, kas mudelisse on targem sisse pista GC-aluspaaride üldarvu geenis või GC-protsenti (alternatiivina: GC osakaalu)?

Oskad sa teoreetiliselt põhjendada, miks on GC-osakaalu/protsendi (või GC-protsendi logaritmi) kasutamine mõistlikum?

```
mudel4=glm(snippe~factor(tyyp)+GCprotsent, family=quasipoisson(),
      offset=log(pikkus))
drop1(mudel4, test="F")
summary(mudel4)
```

Kuidas tulemused tulid? Kas GC-rikastes piirkondades on rohkem või vähem snippe?

Kui testime, kas ”antisense” geenidel on keskmiselt rohkem snippe aluspaari kohta kui ”lincRNA” tüüpi geenil, siis milline tuleb selle testi p-väärtus (mudel4 korral):

p-väärtus: ……..

Proovime lisada ka informatsiooni geeni paiknemise kohta kromosoomis (lubame mõndades piirkondades rohkemate mutatsioonide olemasolu kui teistes piirkondades):

```
library(splines)
mudel5=glm(snippe~factor(tyyp)+bs(algus, df=100)+ GCprotsent,
      family=quasipoisson(), offset=log(pikkus))
drop1(mudel5, test="F")
```

Näeme, et geeni asukoht on tähtis, teades geeni paiknemist kromosoomis võib aimata ka seda, kas ta asub snipirikkas piirkonnas või snipivaeses kandis. Aga millistes piirkondades on siis palju snippe, millistes vähe? Teeme selle mõistmiseks ühe abistava joonise.

Esmalt joonistame selle, kui palju snippe oleks ühes- või teises genoomi osas snippe proteiini kodeerivas 100 aluspaari pikkuses geenis (GC 50%):

```
xx=seq(min(algus), max(algus), length=10000)
yy=predict(mudel5, data.frame(tyyp="protein_coding", GCprotsent=50,
 algus=xx, pikkus=1000), type="response")

plot(xx, yy, type="l")
rug(algus)
```

Kui palju oleks samas kohas snippe polümorfses psudogeenis? Lisame joonisele teise, punase joone samas kohas paikneva psudogeeni oodatava snippide arvu kirjeldamiseks:

```
y2=predict(mudel5, data.frame(tyyp="polymorphic_pseudogene",
      GCprotsent=50, algus=xx, pikkus=1000), type="response")
lines(xx, y2, col=2)
```

Kus tundub paiknevat palju snippe, kus vähe?

Mis tüüpi geend siis ikkagi erinevad teistest? Vaatleme kõikmõeldavaid võrdluseid, arvestame mitmese testimise probleemiga (NB! Võrdlemine on üsna aeglane):

```
install.packages("multcomp")
library(multcomp)

tyypF=factor(tyyp)
mudel5=glm(snippe~tyypF+bs(algus, df=100)+GCprotsent,
      family=quasipoisson(), offset=log(pikkus))

v6rdlused=glht(mudel5, linfct = mcp(tyypF = "Tukey"))
confint(v6rdlused)
```

Väga aeglane: `summary(v6rdlused)`

Kas negatiivne binoomjaotus või kvaasi-Poisson?

```
mudel5=glm(snippe~factor(tyyp)+bs(algus, df=100)+GCprotsent,
      family=quasipoisson(), offset=log(pikkus))

prog=predict(mudel5, type="response")
```

Jagame saadud prognoosid klassidesse–päris väikesed, veidi suuremad jne:
Kokku teeme 10 enam-vähem samasuurt klassi:

```
klass=cut(prog, breaks=quantile(prog, seq(0,1, length=11)))
table(klass)
```

Igas klassis prognooside keskmine ja uuritava tunnuse (jääkide) dispersioon
```
EY=as.vector(by(prog, klass, mean))
DY=as.vector(by(snippe-prog, klass, var))
plot(EY, DY)
```

Milline on seos dispersiooni ja keskväärtuse vahel Poissoni regressiooni korral
```
abline(coef=c(0,1))
```

kvaasi-Poissoni regressiooni arvates seos EY ja DY vahel
```
summary(mudel5)
abline(coef=c(0, 15.16027), col=2)
```

Hindame negatiivset binoomjaotust kasutades mudeli snippide arvukusele:
```
library(MASS)
mudel5a=glm.nb(snippe~factor(tyyp)+bs(algus,df=100)+GCprotsent+
      offset(log(pikkus)))
summary(mudel5a)
```

Seos negatiivse binoomjaotuse korral keskväärtuse ja dispersiooni vahel:
```
xx=seq(0, 4000)
lines(xx, xx+xx^2/ 13.870, col=3)
```

Antud juhul tundub kvaasi-Poissoni mudel adekvaatsemalt kirjeldavat keskväärtuse ja dispersiooni omavahelist seost…

Lisamaterial huvilistele – näide alahajuvusest

Teine näide seostub lindudega. Seekord vaatame põllulinde üldisemalt, mitte ainult rukkirääke (andmestik on taas pärit Riho Marjalt). Andmestiku kirjelduse leiad siit:

http://www.ms.ut.ee/mart/biomeetria2015pmk.pdf

Andmestiku sisselugemine R'i käib järgmise käsu abil:

```
andmed2=
   read.table(url("http://www.ms.ut.ee/mart/biomeetria2015/pmk.txt"),
   header=TRUE,  sep="", dec=",")

head(andmed2)
```

Antud praktikumis tunneme huvi kahe tunnuse modelleerimise vastu. Üheks huvipakkuvaks tunnuseks oleks pesitsevate liikide arv transektil (tunnus *pl*), ehk mingis mõttes looduskeskkonna mitmekesisus (kvaliteet). Teiseks huvipakkuvaks tunnuseks oleks pesitsevate paaride arv (kvantiteet, tunnus *sum*).

Argumenteeri paari lausega, kas need tunnused võiksid olla Poissoni jaotusega või mitte. Kumma neist tunnustest jaotus võiks olla lähedasem Poissoni jaotusele?

Kuigi meil võib esineda kahtluseid nende tunnuste jaotuse osas, on siiski selge, et mõlema tunnuse puhul oleks absurdne oodata negatiivset keskväärtust. Seega log-seosefunktsiooni kasutamine (nagu Poissoni regressioon seda teeb) on ikkagi asjakohane. Uuritava tunnuse hajuvuse osas peame aga olema valvsad – antud juhul on põhjust kahelda, kas ikka uuritava tunnuse dispersioon on sama mis keskväärtus, vaata ka näiteks:

```
var(pl)
mean(pl)
```

Alustame lihtsaima võimaliku mudeliga liikide arvule.

```
mudel1=glm(pl~1, family=poisson())
summary(mudel1)
```

Kas oskad põhjendada, miks me antud juhul offset-i ei kasutanud? Kirjelda olukorda, kus offset-i kasutamine oleks osutunud vajalikuks! Muuseas, kas märkad väljundit vaadates ka mõnda potentsiaalset probleemi?

Võrdle saadud tulemust kvaasi-Poissoni mudeli tulemusega:

```
mudel2=glm(pl~1, family=quasipoisson())
summary(mudel2)
```

Kas midagi muutus?

Leiame meie mudeli poolt ennustatud keskmise liikide arvu transektil:

```
exp(0.61026)
```

Liigume edasi tõsisemate mudelite poole. Võtame liikide arvu kirjeldusse sisse kõik tunnused, mis võiksid vähegi ala sobivust (linnule või vaatlejale) kirjeldada, esialgu vaid peamõjudena (koosmõjusid proovime lisada hiljem):

```
mudel_uhke = glm( pl ~ factor(aasta) + factor(loendus) +
factor(piirkond) + factor(toetustyyp)+ factor(tuul) +factor(temp) +
      factor(pilv)+ hein+vili+maastik, family=quasipoisson())
```

Testime, milliseid tunnuseid saaksime mudelist eemaldada ilma mudeli prognoosivõimet oluliselt halvendamata. Märka, et kvaasi-Poissoni mudeli korral kasutame testina F-testi, Poissoni mudeli puhul eelistasime hii-ruut testi („Chisq“):

```
drop1(mudel_uhke, test="F")
```

Peaksime saama järgmise väljundi:

```
[...]
                  Df Deviance F value     Pr(F)
<none>                 167.74
factor(aasta)      0   167.74
factor(loendus)    4   174.24  3.6482  0.006244 **
factor(piirkond)   1   191.31 52.9695 1.985e-12 ***
factor(toetustyyp) 2   170.71  3.3345  0.036686 *
factor(tuul)       2   170.07  2.6200  0.074131 .
factor(temp)       2   168.26  0.5864  0.556819
factor(pilv)       3   168.14  0.2986  0.826411
hein               1   168.14  0.8972  0.344148
vili               1   168.93  2.6681  0.103216
maastik            1   172.45 10.5853  0.001243 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

Aasta mõju sellisel viisil testida ei saa, sest tunnuses *loendus* on ajainformatsioon detailsemalt kui tunnuses *aasta* kirjas ning seega aasta teadmine täiendavat informatsiooni mudelile ei anna (kui *loendus*=2, siis on 2010.a., kui *loendus*=5, siis on 2011.a. jne)

Tunnuse loendus mõju testimisel küsitakse, kas lisaks aasta teadmisele on tarvis teada ka täpsemat loenduse aega. Kui tunnust aasta poleks aga mudelis, siis testiks seesama test küsimust kas aastati muutub linnuriikide rohkus või kas aastasiseselt muutub kohatavate linnuriikide rohkus. Antud juhul saame teada, et ka aastasisene linnuriikide arvukuse muutus on täiesti olemas.

Tunnus *pilv* pole statistiliselt oluline, seega võime öelda, et kui me juba teame, et on tuuline ja külm (tunnused *tuul* ja *temp* on mudelis ju sees), siis teadmine, et on ka pilves ei anna loendaja poolt nähtud linnuliikide arvu prognoosimiseks täiendavat informatsiooni.

Eemaldame mudelist mitteolulised peamõjud. Tuleb meeles pidada, et tunnuseid on otstarbekas eemaldada ükshaaval ja peale iga tunnuse eemaldamist tuleb drop1-käsuga uuesti testida tunnuste vajalikkust/eemaldatavust. Mittevajalikke tunnuseid välja loopides võiksime näiteks jõuda järgmise mudelini:

```
mudel_uhke = glm( pl ~  factor(loendus) + factor(piirkond) +
      factor(toetustyyp)+ factor(tuul)+ vili + maastik,
      family=quasipoisson())
```

Järgmises modelleerimise etapis võiksime proovida lisada tunnustevahelisi koosmõjusid (selliseid, mis tunduvad loogilised) ja testida, kas mõne pideva tunnuse puhul ehk seos pole täiesti lineaarne (lisades mudelisse prooviks pidevate tunnuste ruutliikmed):

```
mudel_uhke = glm( pl ~  factor(loendus) + factor(piirkond) +
      factor(toetustyyp)+ factor(tuul)+ vili+maastik+
      factor(loendus)*factor(piirkond)+factor(piirkond)*factor(tuul)+
      factor(piirkond)*vili+I(vili^2)+I(maastik^2),
            family=quasipoisson())
drop1(mudel_uhke, test="F")
```

Proovi ükshaaval eemaldada mitteolulisi liikmeid mudelist. Kas jõuad samuti alltoodud mudelini? Kui jõudsid mõne muu mudelini, anna märku!

```
mudel_uhke = glm( pl ~  factor(loendus) + factor(piirkond) +
      factor(toetustyyp)+ factor(tuul)+ vili+maastik+
      factor(loendus)*factor(piirkond)+factor(piirkond)*factor(tuul),
            family=quasipoisson())
```

Uurimisülesande peamine eesmärk on kirjeldada põllumajandustoetuse mõju linnustikule. Proovime korra vaadata, mida meie poolt valitud mudel ütleb põllumajandustoetuste mõju kohta.

```
drop1(mudel_uhke, test="F")
[...]
                                  Df Deviance F value     Pr(F)
<none>                                 153.06
factor(toetustyyp)                 2   155.90  3.4948 0.0313453 *
[...]
```

Näeme, et erinevate põllumajandustoetuste korral on linnuliikide arvukus (tõestatavalt) erinev (vähemalt olulisuse nivool 0,05). Aga milline on erinevus? Kasuta käsku

```
summary(mudel_uhke)
```

ja proovi vastata järgmistele küsimustele:
a) mitu korda on liike mahetoetust (toetustyyp=“mahe“) saavatel aladel rohkem kui keskkonnasõbraliku majandamise toetust saavatel aladel (toetustyyp=“ksm“), juhul kui räägime samast aastast ja aastaajast (loenduskorrast), samatuulisest ilmast, samast piirkonnast ja maastikuelementide ja viljarikkuselt samaväärsetest aladest?

b) Mitu korda enam on linnuliike üldist pidalatoetust (toetustyyp=“ypt“) saavatel aladel rohkem kui keskkonnasõbraliku majandamise toetust saavatel aladel (sama tuulisi ilma, sama piirkonna jne korral)?

c) Mitu korda on mahetoetusega aladel linde rohkem kui üldist pindalatoetust saavatel aladel? Vihje: küsimusele vastamisel võib olla kasu relevel-käsust!

d) leia vähemalt ühe ülesandes a)..c) väljarehkendatud kordaja jaoks ka 95%-usaldusintervall!

e) miks valisime kvaasi-Poissoni mudeli, mitte aga negatiivsel binoomjaotusel baseeruva mudeli?

Proovime veel mudelit täiendada. Ennem, koosmõjude lisamise ajal, jäi proovimata koosmõju toetustyyp*loendus. Lisame vastava koosmõju nüüd mudelisse ja vaatame, kas vastav koosmõju on statistiliselt oluline:

```
mudel_uhke = glm( pl~factor(loendus)+factor(piirkond)+
      factor(toetustyyp)+ factor(tuul)+ vili+maastik+
      factor(loendus)*factor(piirkond)+factor(piirkond)*factor(tuul)+
      factor(toetustyyp)*factor(loendus), family=quasipoisson())
drop1(mudel_uhke, test="F")
```

Nii. Kas meid huvitav koosmõju oli statistiliselt oluline? Mida me nüüd teeme?

Esmalt testime, kas põllumajandustoetuse tüübil on (pesitsevate) linnuliikide arvule olulist mõju. Selleks hindame kaks mudelit. Üks mudelitest on selline, kus oleme kasutanud tunnust toetustyyp (`+factor(toetustyyp) + factor(toetustyyp) * factor(loendus)`) ja teine mudel on muus osas samasugune, aga toetuse tüüpi sisaldavad liikmed oleme eemaldanud:

```
mudel_uhke = glm( pl~factor(loendus)+factor(piirkond)+
      factor(toetustyyp)+ factor(tuul)+ vili+maastik+
      factor(loendus)*factor(piirkond)+factor(piirkond)*factor(tuul)+
      factor(toetustyyp)*factor(loendus), family=quasipoisson())
mudel_vaike = glm( pl~factor(loendus)+factor(piirkond)+
                         factor(tuul)+ vili+maastik+
      factor(loendus)*factor(piirkond)+factor(piirkond)*factor(tuul),
                                          family=quasipoisson())
```

Testime mudelite erinevust (Poissoni regressiooni puhul eelistatavalt `test="Chisq"`, kvaasi-Poissoni puhul `test="F"`)

```
anova(mudel_uhke, mudel_vaike, test="F")
```

Testi tulemus:

```
[...]

  Resid. Df Resid. Dev  Df Deviance      F   Pr(>F)
1       366     144.85
2       378     155.90 -12  -11.047 2.4152 0.005052 **
```

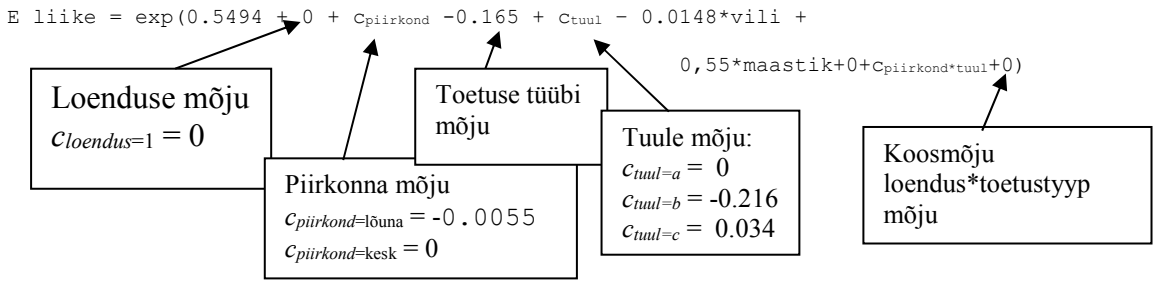
Näeme, et põllumajandustoetuse tüübi arvestamine aitab täpsemalt prognoosida alal pesitsevate lindude arvukust.

Koosmõjudega mudeli puhul on aga mudeli parameetrite tähenduse leidmine veidi keerukam. Mitu korda rohkem liike pesitseb mahetoetust saaval alal sõltub nüüd loenduse korrast. Üritame saadud tulemust siiski kuidagimoodi iseloomustada.

```
> summary(mudel_uhke)

[...]

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               0.549432   0.139761   3.931 0.000101 ***
factor(loendus)2                         -0.265208   0.179851  -1.475 0.141179
factor(loendus)3                         -0.040732   0.174086  -0.234 0.815132
factor(loendus)4                         -0.109985   0.183174  -0.600 0.548583
factor(loendus)5                         -0.222333   0.182913  -1.216 0.224953
factor(loendus)6                         -0.229990   0.241758  -0.951 0.342069
factor(piirkond)louna                    -0.005505   0.135964  -0.040 0.967725
factor(toetustyyp)mahe                   -0.165120   0.151584  -1.089 0.276741
factor(toetustyyp)ypt                    -0.074383   0.149257  -0.498 0.618534
factor(tuul)b                            -0.216350   0.097192  -2.226 0.026622 *
factor(tuul)c                             0.034521   0.198857   0.174 0.862277
vili                                     -0.014805   0.005956  -2.486 0.013368 *
maastik                                   0.550927   0.147095   3.745 0.000209 ***
factor(loendus)2:factor(piirkond)louna    0.691953   0.172490   4.012 7.32e-05 ***
factor(loendus)3:factor(piirkond)louna    0.146454   0.165433   0.885 0.376588
factor(loendus)4:factor(piirkond)louna   -0.184910   0.178556  -1.036 0.301080
factor(loendus)5:factor(piirkond)louna    0.211158   0.188225   1.122 0.262668
factor(loendus)6:factor(piirkond)louna    0.248562   0.260219   0.955 0.340105
factor(piirkond)louna:factor(tuul)b       0.411395   0.130727   3.147 0.001785 **
factor(piirkond)louna:factor(tuul)c       0.345514   0.238383   1.449 0.148081
factor(loendus)2:factor(toetustyyp)mahe   0.223292   0.197164   1.133 0.258157
factor(loendus)3:factor(toetustyyp)mahe   0.497843   0.197646   2.519 0.012199 *
factor(loendus)4:factor(toetustyyp)mahe   0.007851   0.224182   0.035 0.972084
factor(loendus)5:factor(toetustyyp)mahe   0.478809   0.200352   2.390 0.017361 *
factor(loendus)6:factor(toetustyyp)mahe   0.425031   0.195400   2.175 0.030256 *
factor(loendus)2:factor(toetustyyp)ypt    0.003833   0.200749   0.019 0.984776
factor(loendus)3:factor(toetustyyp)ypt   -0.026029   0.208827  -0.125 0.900876
factor(loendus)4:factor(toetustyyp)ypt    0.203006   0.215098   0.944 0.345903
factor(loendus)5:factor(toetustyyp)ypt    0.138678   0.208117   0.666 0.505610
factor(loendus)6:factor(toetustyyp)ypt    0.043501   0.202802   0.215 0.830276

(Dispersion parameter for quasipoisson family taken to be 0.3811735)
```

Proovime aru saada, mitu korda rohkem esineb liike mahetoetusega alal (võrreldes toetustüübiga ksm). Teeme arvutused 1. loenduse (aprill 2010) ja 2. loenduse (mai 2010) jaoks.

Mahetoetus, 1. loendus:



Toetus ksm, 1. loendus:

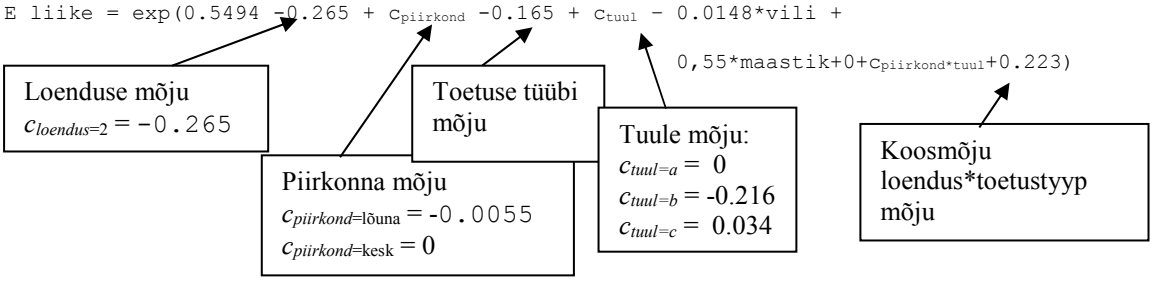
E liike = exp(0.5494 + 0 + cpiirkond + 0 + ctuul – 0.0148*vili+

0,55*maastik+0+cpiirkond*tuul+0)

E(liike | mahe, 1. loendus) / E(liike | ksm, 1. loendus) = exp(-0.165)
= 0.848

Esimesel loenduskorral (2010. aasta aprillis) oli mahetoetust saavatel aladel keskmiselt exp(-0.165)=0.848 korda rohkem liike kui ksm-toetust saavatel aladel (ehk tegelikult siis hoopistükis 15% vähem liike). Tõsi, vastav kordaja polnud statistiliselt oluline (kordaja võis olla ka 0 ehk mahetoetust saavatel aladel võis olla aprillikuus samapalju pesitsevaid liike kui ksm-toetust saavatel aladel).

Mahetoetus, 2. loendus:



Toetus ksm, 2. loendus:

E liike = exp(0.5494 -0.265 + cpiirkond + 0 + ctuul – 0.0148*vili+
0,55*maastik+0+cpiirkond*tuul+0)

E(liike | mahe, 2. loendus) / E(liike | ksm, 2. loendus) = exp(-0.165+0.223)
= 1.0597

Ehk teise loenduskorra ajal (mai 2010) oli juba mahetoetust saavatel aladel liike 1.06 korda rohkem kui ksm-toetust saavatel aladel.

Iseloomustame oma mudeli prognoose ühe konkreetse piirkonna ja keskkonnatingimuste jaoks. Valime välja näiteks lõuna piirkonnas paikneva maalapi, kus loendust on teostatud tuulevaikse ilmaga, kus maastik=0.1 ja vili=5.6 (tunnuste maastik ja vili keskmised väärtused):

Minimalistlik variant:

```
y1=predict(mudel_uhke, data.frame(loendus=1:6, piirkond="louna",
      vili=5.6, maastik=0.1, toetustyyp="mahe", tuul="a"),
      type="response")
y2=predict(mudel_uhke, data.frame(loendus=1:6, piirkond="louna",
      vili=5.6, maastik=0.1, toetustyyp="ypt", tuul="a"),
      type="response")
y3=predict(mudel_uhke, data.frame(loendus=1:6, piirkond="louna",
      vili=5.6, maastik=0.1, toetustyyp="ksm", tuul="a"),
      type="response")

plot(1:6, y1, type="b")
points(1:6, y2, type="b", col=2)
points(1:6, y3, type="b", col=3)
```

Ilusaks tehtud joonisena:

```
y1=predict(mudel_uhke, data.frame(loendus=c(1:3, NA, 4:6),
      piirkond="louna", vili=5.6, maastik=0.1, toetustyyp="mahe",
      tuul="a"), type="response")
y2=predict(mudel_uhke, data.frame(loendus=c(1:3, NA, 4:6),
      piirkond="louna", vili=5.6, maastik=0.1, toetustyyp="ypt",
      tuul="a"), type="response")
y3=predict(mudel_uhke, data.frame(loendus=c(1:3, NA, 4:6),
      piirkond="louna", vili=5.6, maastik=0.1, toetustyyp="ksm",
      tuul="a"), type="response")

x=c(1:3, NA, 4:6)
windows(width=10, height=8)
par(mar=c(8,3,3,1))
plot(x,y1, type="b", lwd=4, col="#d62728", xaxt="n", xlab="",
      ylab="E liike", ylim=c(1,3.5), pch=20)
abline(h=seq(1,3.5, by=0.5), lty=2, col="gray75")
axis(1, at=1:6, c("aprill 2010", "mai 2010", "juuni 2010",
      "aprill 2011", "mai 2011", "juuni 2011"), las=2)
lines(x, y2, lwd=4, col="#fe7f0e", type="b", pch=20)
lines(x, y3, lwd=4, col="#be9b21", type="b", pch=20)
legend(4.4, 3.45, c("mahetoetus", "üldine pindalatoetus",
      "keskonnasõbralik majandamine"), lwd=4, pch=20,
      col=c("#d62728", "#fe7f0e", "#be9b21"))
```

Ülesanne
Proovi iseseisvalt hinnata mudel, mis kirjeldaks pesitsevate paaride üldarvu (tunnus sum). Kas ja kuidas mõjub keskkonnatoetus kvantiteedile?