Meenutuseks

```
> mm=glm(y~factor(SNP), family=poisson())
> summary(mm)
Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
A/T (Intercept)    3.1678374  0.0005043 6281.844  < 2e-16 ***
factor(SNP)A/C -0.0085273  0.0006900  -12.359  < 2e-16 ***
factor(SNP)A/G -0.0105455  0.0005553  -18.991  < 2e-16 ***
factor(SNP)C/G -0.0193019  0.0006872  -28.088  < 2e-16 ***
factor(SNP)C/T -0.0093731  0.0005553  -16.878  < 2e-16 ***
factor(SNP)G/T -0.0055447  0.0006886   -8.053 8.11e-16 ***
```

E(Y| SNP=„A/T“) = exp(3.16..) =23,7…

E(Y| SNP=„C/G“) = exp(3.16..)*exp(-0.019..)

= 23,7..*0,98.. =23,3..

1

Rukkirääkude mudel ...

```
> m2=glm(raak~factor(toetustyyp), offset=log(pindala),
                                          family=poisson())
> summary(m2)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.94497    0.11868 -33.241  < 2e-16 ***
factor(toetustyyp)ksm  -0.95392    0.15760  -6.053 1.42e-09 ***
factor(toetustyyp)mahe -0.02797    0.18805  -0.149    0.882
factor(toetustyyp)ypt   0.19152    0.13556   1.413    0.158
factor(toetustyyp) ei   0           .         .        .
```

Kui toetustüüp=“ksm”

log(E(*raak*)) = -3.94497 -0.95392 + log(pindala)

E(*raak*) = exp(-3.94497 -0.95392 + log(pindala))

= 0.007454.. * pindala

```
> predict(m2, data.frame(toetustyyp="ksm", pindala=1), type="response")
        1
0.007454874
```

Täiendame mudelit...

```
> m3=glm(raak~factor(toetustyyp)+factor(hairing), offset=log(pindala),
             family=poisson())

> summary(m3)

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 ***
```

```
> table(hairing)
hairing
      hairing hairingutpole
          267          3041
```

Exp(1.042..) = 2.836..
Häiringuteta aladel on 2,8 korda enam rukkirääke kui sama toetustüübiga samasuurtel häiringuga aladel...

Koosmõjust I. Peamõjud *(Main effects)*

Tavaline lineaarne mudel

E *rääk/pindala* = $\mu + \alpha_{toetustüüp} + \beta_{häiring}$

μ = -0.0026

α_{ksm} = -0.0032

α_{mahe} = 0.0124

α_{ypt} = 0.0112

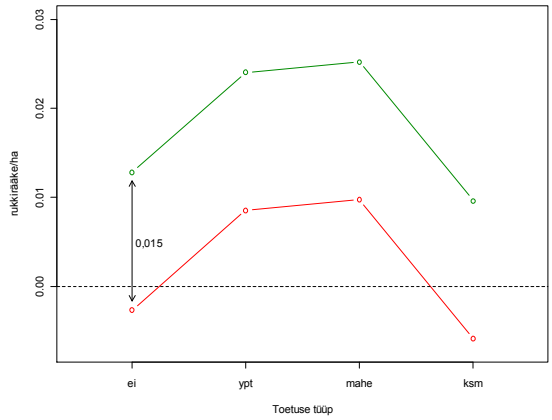
α_{ei} = 0 (võrdlustase)

$\beta_{häiringut\ pole}$ = 0.015

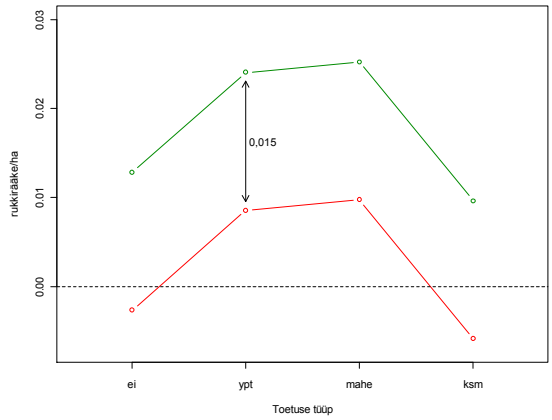
$\beta_{häiring}$ = 0 (võrdlustase)

```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.002640   0.004826  -0.547 0.584415
factor(toetustyyp)ksm       -0.003194   0.003163  -1.010 0.312726
factor(toetustyyp)mahe       0.012404   0.004831   2.567 0.010294 *
factor(toetustyyp)ypt        0.011198   0.002996   3.738 0.000189 ***
factor(hairing)hairingutpole 0.015486   0.004476   3.460 0.000547 ***
```

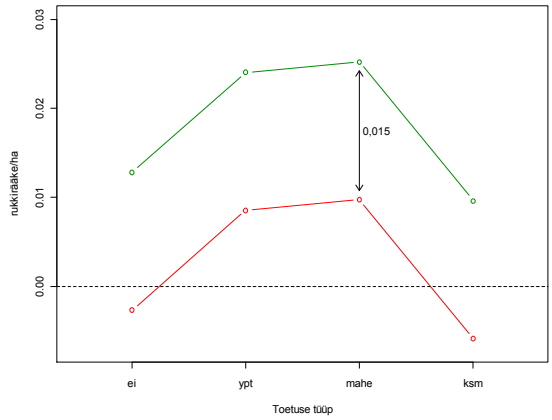
Peamõjudega tavaline lineaarne mudel (ANOVA)



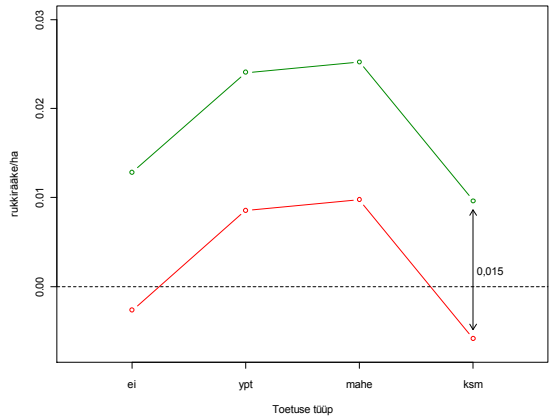
Peamõjudega tavaline lineaarne mudel (ANOVA)



Peamõjudega tavaline lineaarne mudel (ANOVA)



Peamõjudega tavaline lineaarne mudel (ANOVA)



Koosmõjust I. Koosmõjud *(interaction)*

Tavaline lineaarne mudel

$$E\ rääk/pindala = \mu + \alpha_{toetustüüp} + \beta_{häiring} + (\alpha\ \beta)_{toetustüüp,häiring}$$

μ = −0.0026

α_{ksm} = −0.0032 $\quad \beta_{häiringut\ pole}$ = 0.015

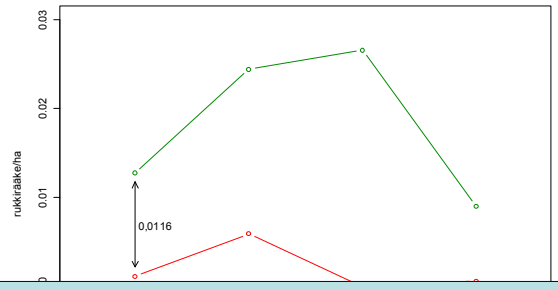
α_{mahe} = 0.0124 $\quad \beta_{häiring}$ = 0 (võrdlustase)

α_{ypt} = 0.0112

α_{ei} = 0 (võrdlustase)

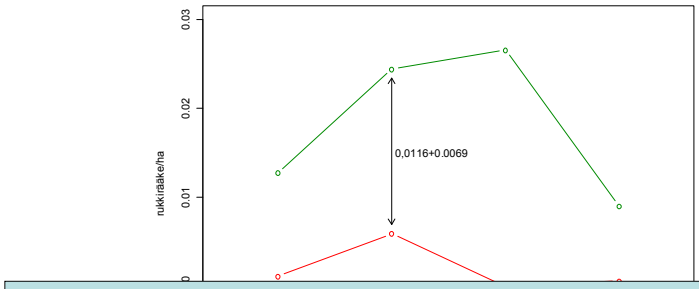
$(\alpha\beta)_{ksm,\ pole}$	= −0.00317	$(\alpha\beta)_{ksm,\ häiring}$	= 0
$(\alpha\beta)_{mahe,\ pole}$	= 0.01497	$(\alpha\beta)_{mahe,\ häiring}$	= 0
$(\alpha\beta)_{ypt,\ pole}$	= 0.00069	$(\alpha\beta)_{ypt,\ häiring}$	= 0
$(\alpha\beta)_{ei,\ pole}$	= 0	$(\alpha\beta)_{ei,\ häiring}$	= 0

Koosmõjudega tavaline lineaarne mudel (ANOVA)



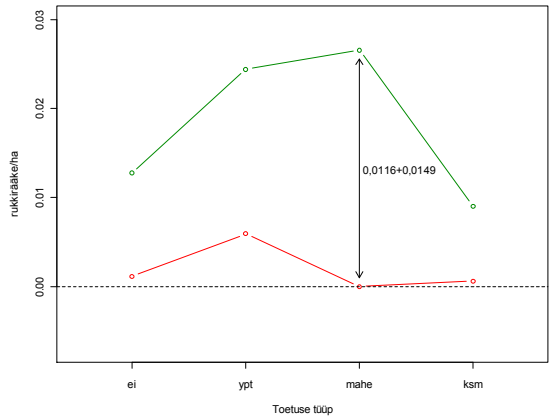

```
(Intercept)                              0.0011620340
factor(toetustyyp)ksm                   -0.0005550818
factor(toetustyyp)mahe                  -0.0011620340
factor(toetustyyp)ypt                    0.0047687493
factor(hairing)hairingutpole             0.0115739846
factor(toetustyyp)ypt:factor(hairing)hairingutpole    0.0068889103
factor(toetustyyp)mahe:factor(hairing)hairingutpole   0.0149714617
factor(toetustyyp)ksm:factor(hairing)hairingutpole   -0.0031756087
```

Koosmõjudega tavaline lineaarne mudel (ANOVA)

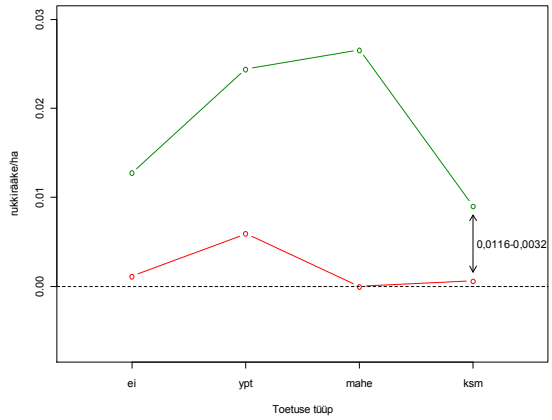



```
(Intercept)                                    0.0011620340
factor(toetustyyp)ksm                         -0.0005550818
factor(toetustyyp)mahe                        -0.0011620340
factor(toetustyyp)ypt                          0.0047687493
factor(hairing)hairingutpole                   0.0115739846
factor(toetustyyp)ypt:factor(hairing)hairingutpole   0.0068889103
factor(toetustyyp)mahe:factor(hairing)hairingutpole  0.0149714617
factor(toetustyyp)ksm:factor(hairing)hairingutpole  -0.0031756087
```

Koosmõjudega tavaline lineaarne mudel (ANOVA)



Koosmõjudega tavaline lineaarne mudel (ANOVA)

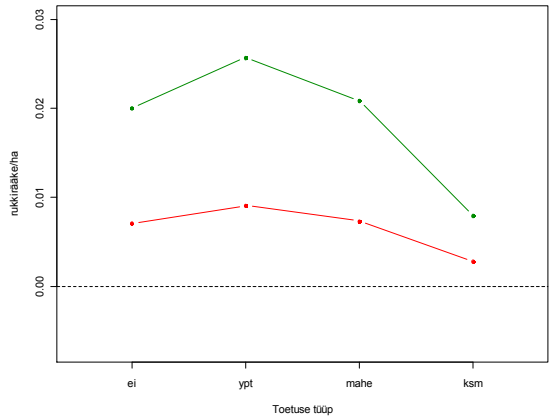


Koosmõjudeta Poissoni regressioon

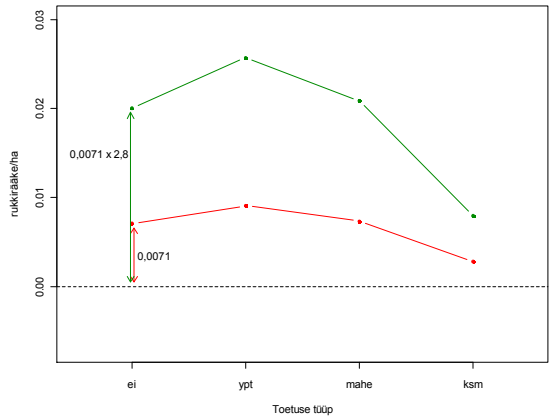
```
> m3=glm(raak~factor(toetustyyp)+factor(hairing), offset=log(pindala),
                family=poisson())

> summary(m3)

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 ***
```

Exp(1.042..) = 2.836..
Häiringuteta aladel on 2,8 korda enam rukkirääke kui sama toetustüübiga samasuurtel häiringuga aladel...

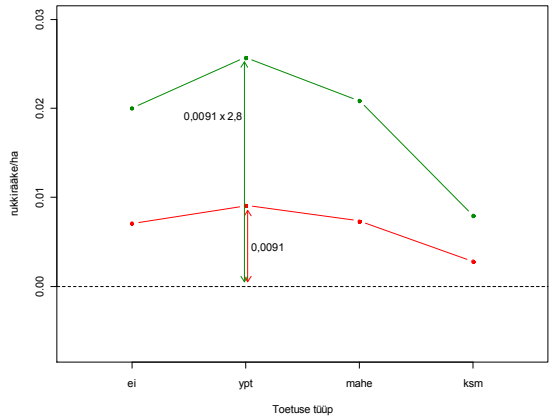
Koosmõjudeta Poissoni regressioon



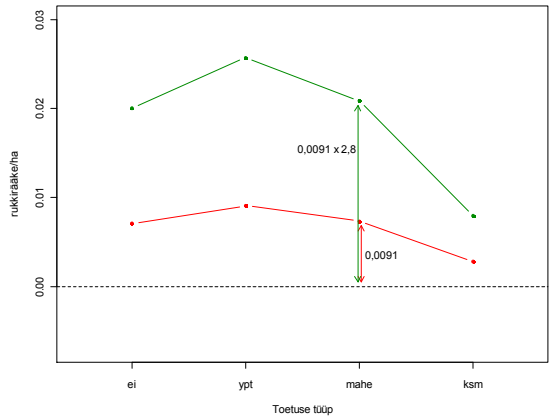
Koosmõjudeta Poissoni regressioon



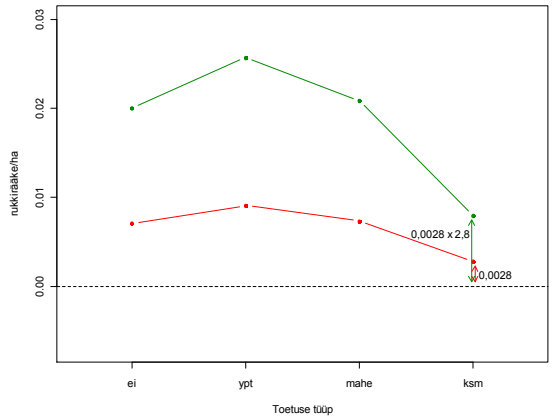
Koosmõjudeta Poissoni regressioon



Koosmõjudeta Poissoni regressioon



Koosmõjudeta Poissoni regressioon



Koosmõjudega Poissoni regressioon

```
> m4=glm(raak~factor(toetustyyp)+factor(hairing)+
                          factor(toetustyyp)*factor(hairing),
                                  offset=log(pindala), family=poisson())
> summary(m4)


Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -5.2807     1.0000  -5.281 1.29e-07
factor(toetustyyp)ksm                             -0.7445     1.1547  -0.645    0.519
factor(toetustyyp)mahe                           -13.2456   328.7088  -0.040    0.968
factor(toetustyyp)ypt                              0.8372     1.0308   0.812    0.417
factor(hairing)hairingutpole                       1.3766     1.0071   1.367    0.172
factor(toetustyyp)ksm:factor(hairing)hairingutpole -0.1783     1.1656  -0.153    0.878
factor(toetustyyp)mahe:factor(hairing)hairingutpole 13.3411   328.7088   0.041    0.968
factor(toetustyyp)ypt:factor(hairing)hairingutpole -0.6104     1.0399  -0.587    0.557
```

Toetust ei saa, häiring, 1 hektarine maalapp:

E rääk = exp(-5.28 + 0 + 0 + 0 +log(1)) = exp(-5.28)

Toetust ei saa, häiringut pole, 1 hektarine maalapp:

E rääk = exp(-5.28 + 0 + 1.3766 + 0 +log(1))
= exp(-5.28)*exp(1.3766) = exp(-5.28)*3,96

Koosmõjudega Poissoni regressioon

```
> m4=glm(raak~factor(toetustyyp)+factor(hairing)+
                          factor(toetustyyp)*factor(hairing),
                                  offset=log(pindala), family=poisson())
> summary(m4)
```

```
Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -5.2807     1.0000  -5.281 1.29e-07
factor(toetustyyp)ksm                             -0.7445     1.1547  -0.645    0.519
factor(toetustyyp)mahe                           -13.2456   328.7088  -0.040    0.968
factor(toetustyyp)ypt                              0.8372     1.0308   0.812    0.417
factor(hairing)hairingutpole                       1.3766     1.0071   1.367    0.172
factor(toetustyyp)ksm:factor(hairing)hairingutpole -0.1783     1.1656  -0.153    0.878
factor(toetustyyp)mahe:factor(hairing)hairingutpole 13.3411   328.7088   0.041    0.968
factor(toetustyyp)ypt:factor(hairing)hairingutpole -0.6104     1.0399  -0.587    0.557
```

Mahetoetus, häiring, 1 hektarine maalapp:
E rääk = exp(-5.28 -13.25 + 0 + 0 +log(1)) = exp(-5.28-13.25)
= 0,00000008...

Mahetoetus, häiringut pole, 1 hektarine maalapp:
E rääk = exp(-5.28 -13.25 + 1.3766 + 13.34 +log(1))
= exp(-5.28-13.25)*exp(14.72) = exp(-5.28) * 2470670

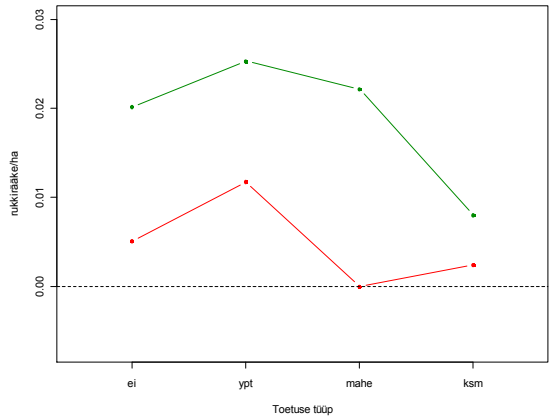
Kui keskväärtus peaks olema 0...

```
> m4=glm(raak~factor(toetustyyp)+factor(hairing)+
    factor(toetustyyp)*factor(hairing), offset=log(pindala),
    family=poisson(), epsilon=1e-14, maxit=100)
```

```
> coef(summary(m4))
                                                  Estimate  Std. Error      z value     Pr(>|z|)
(Intercept)                                     -5.2807133   -5.2807 ) -5.280713e+00 1.286819e-07
factor(toetustyyp)ksm                           -0.7444680   -0.7445 ) -6.447282e-01 5.191033e-01
factor(toetustyyp)mahe                         -27.2455135  -13.2456 ) -7.558289e-05 9.999397e-01
factor(toetustyyp)ypt                            0.8371654    0.8372 )  8.121697e-01 4.166942e-01
factor(hairing)hairingutpole                     1.3766102    1.3766 )  1.366881e+00 1.716625e-01
factor(toetustyyp)ksm:factor(hairing)hairingutpole -0.1783113  -0.1783 ) -1.529721e-01 8.784203e-01
factor(toetustyyp)mahe:factor(hairing)hairingutpole 27.3409750 13.3411 )  7.584771e-05 9.999395e-01
factor(toetustyyp)ypt:factor(hairing)hairingutpole -0.6104007  -0.6104 ) -5.869797e-01 5.572174e-01
```

In addition: Warning messages:
1: glm.fit: fitted rates numerically 0 occurred

Koosmõjudega Poissoni regressioon

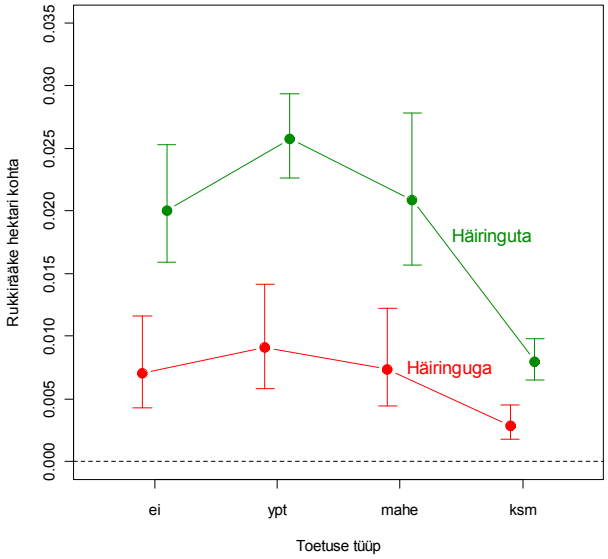


Koosmõjust võib õnneks antud näite puhul loobuda...

```
> m4=glm(raak~factor(toetustyyp)+factor(hairing)+
   factor(toetustyyp)*factor(hairing), family=poisson(),
   offset=log(pindala))
> drop1(m4, test="Chisq")
Model:
raak ~ factor(toetustyyp) + factor(hairing) + factor(toetustyyp) *
    factor(hairing)
                                Df Deviance    AIC    LRT Pr(Chi)
<none>                               1492.2 2233.9
factor(toetustyyp):factor(hairing)  3   1499.2 2234.8 6.9876 0.07229 .

> m4v=glm(raak~factor(to
   offset=log(pindala))
> drop1(m4v, test="Chisq
Model:
raak ~ factor(toetustyyp
                  Df De
<none>
factor(toetustyyp)  3
factor(hairing)     1   1527.8 2261.5  28.657 8.642e-08 ***
```

```
> table(hairing, toetustyyp)
                toetustyyp
hairing            ei  ksm mahe  ypt
  hairing          31   79   30  127
  hairingutpole 1073  787  226  955
```

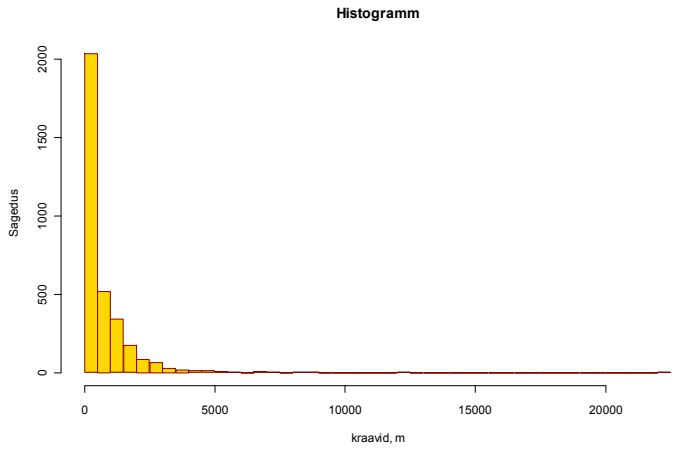
Vahetulemus



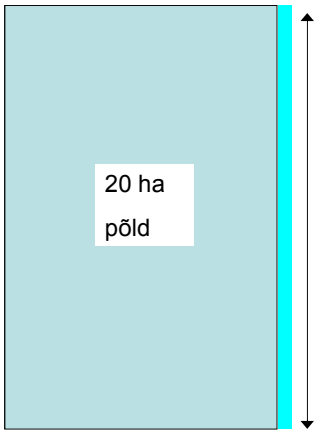
Vahetulemuse joonistanud programm

```
m4v=glm(raak~factor(toetustyyp)+factor(hairing), family=poisson(), offset=log(pindala))

abi_hairinguta=predict(m4v, data.frame(toetustyyp=c("ei","ypt","mahe","ksm"),
    hairing=rep("hairingutpole",4), pindala=1), type="link", se.fit=TRUE)
abi_hairinguga=predict(m4v, data.frame(toetustyyp=c("ei","ypt","mahe","ksm"),
    hairing=rep("hairing",4), pindala=1), type="link", se.fit=TRUE)

prognoos_hairinguta=exp(abi_hairinguta$fit)
prognoos_hairinguga=exp(abi_hairinguga$fit)

UI_hairinguta_alumine=exp(abi_hairinguta$fit-1.96*abi_hairinguta$se.fit)
UI_hairinguta_ylemine=exp(abi_hairinguta$fit+1.96*abi_hairinguta$se.fit)
UI_hairinguga_alumine=exp(abi_hairinguga$fit-1.96*abi_hairinguga$se.fit)
UI_hairinguga_ylemine=exp(abi_hairinguga$fit+1.96*abi_hairinguga$se.fit)

plot(1:4+0.1, prognoos_hairinguta, col="green4", pch=20, cex=2, xlab="Toetuse tüüp",
    ylab="Rukkirääke hektari kohta", xaxt="n", ylim=c(0, 0.035), xlim=c(0.5,4.5),
    type="b")
points(1:4-0.1, prognoos_hairinguga, col="red", pch=20, cex=2, type="b")
axis(1, at=1:4, c("ei","ypt","mahe","ksm"))
abline(h=0, lty=2)
arrows(1:4+0.1, UI_hairinguta_alumine, 1:4+0.1, UI_hairinguta_ylemine, code=3,
    angle=90, length=0.1, col="green4")
arrows(1:4-0.1, UI_hairinguga_alumine, 1:4-0.1, UI_hairinguga_ylemine, code=3,
    angle=90, length=0.1, col="red")
text(3.75, 0.018, "Häiringuta", col="green4", cex=1.2)
text(3.4, 0.0075, "Häiringuga", col="red", cex=1.2)
```

Pideva tunnuse lisamine



Millises skaalas lisada?

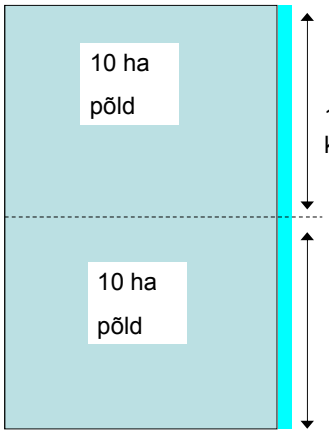


E rääke = $\exp(\beta_0+\beta_1*kraav+\log(pindala))$

$= c_0*\exp(\beta_1*kraav)*pindala$

$= c_0*\exp(\beta_1*200)*20$

Millises skaalas lisada?



Enne kaheks jagamist:

E rääke = $\exp(\beta_0+\beta_1*kraav+\log(pindala))$

= $c_0*\exp(\beta_1*kraav)*pindala$

= $c_0*\exp(\beta_1*200)*20$

Pärast kaheks jagamist:

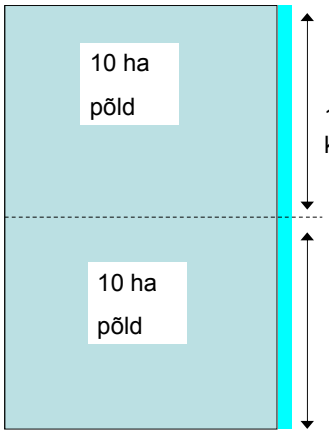
E rääke = $\exp(\beta_0+\beta_1*kraav+\log(pindala))$

= $c_0*\exp(\beta_1*kraav)*pindala$

= $c_0*\exp(\beta_1*100)*10$

$2* c_0*\exp(\beta_1*100)*10 \neq c_0*\exp(\beta_1*200)*20$

Millises skaalas lisada?



Enne kaheks jagamist:

E rääke = $\exp(\beta_0+\beta_1*(kraav/pindala)+\log(pindala))$

= $c_0*\exp(\beta_1*kraav/pindala)*pindala$

= $c_0*\exp(\beta_1*200/20)*20$

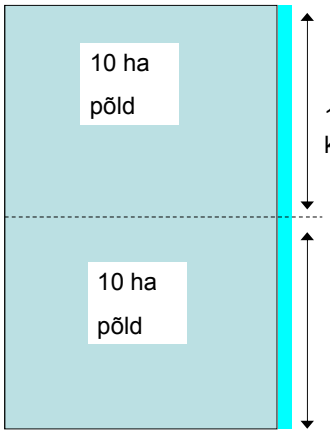
Pärast kaheks jagamist:

E rääke = $c_0*\exp(\beta_1*kraav/pindala)*pindala$

= $c_0*\exp(\beta_1*100/10)*10$

$2* c_0*\exp(\beta_1*100/10)*10 = c_0*\exp(\beta_1*200/20)*20$

Millises skaalas lisada?



Enne kaheks jagamist:

E rääke = $\exp(\beta_0+\beta_1*\log(kraav/pindala)$
$+\log(pindala))$
$= c_0*(\text{kraav/pindala})^{\beta 1}*pindala$
$= c_0*(200/20)^{\beta 1}*20$

Pärast kaheks jagamist:

E rääke = $c_0*(\text{kraav/pindala})^{\beta 1}*pindala$
$= c_0*(100/10)^{\beta 1}*10$

$2*\ c_0*(100/10)^{\beta 1}*10 =$
$c_0*(200/20)^{\beta 1}*20$

Lisame kraavid...

```
> mudel1=glm(raak~factor(toetustyyp)+factor(hairing)+I(kraavid2/pindala)+
                offset(log(pindala)), family=poisson()); summary(mudel1)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -4.9283311  0.2603598 -18.929  < 2e-16 ***
factor(toetustyyp)ksm        -0.9357706  0.1606201  -5.826 5.68e-09 ***
factor(toetustyyp)mahe        0.0364291  0.1884282   0.193   0.8467
factor(toetustyyp)ypt         0.2401695  0.1373258   1.749   0.0803 .
factor(hairing)hairingutpole  1.0417357  0.2292775   4.544 5.53e-06 ***
I(kraavid2/pindala)          -0.0001896  0.0004568  -0.415   0.6781

AIC: 2236.7

> drop1(mudel1, test="Chisq")

                    Df Deviance    AIC     LRT   Pr(Chi)
<none>                   1499.0 2236.7
factor(toetustyyp)   3   1606.7 2338.3 107.665 < 2.2e-16 ***
factor(hairing)      1   1527.6 2263.3  28.601 8.894e-08 ***
I(kraavid2/pindala)  1   1499.2 2234.8   0.180    0.6717
```

Lisame kraavid...

```
> mudel1=glm(raak~factor(toetustyyp)+factor(hairing)+
        I(log(kraavid2/pindala))+offset(log(pindala)), family=poisson())
> summary(mudel1)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -5.04603    0.27111 -18.613  < 2e-16 ***
factor(toetustyyp)ksm      -0.89704    0.15963  -5.620 1.91e-08 ***
factor(toetustyyp)mahe      0.02807    0.18859   0.149   0.8817
factor(toetustyyp)ypt       0.25243    0.13583   1.858   0.0631 .
factor(hairing)hairingutpole  1.05070    0.22940   4.580 4.65e-06 ***
I(log(kraavid2/pindala))    0.02412    0.02396   1.007   0.3139

AIC: 2235.8

> drop1(mudel1, test="Chisq")
                      Df Deviance    AIC    LRT   Pr(Chi)
<none>                     1498.1 2235.8
factor(toetustyyp)     3   1597.2 2328.9 99.038 < 2.2e-16 ***
factor(hairing)        1   1527.3 2262.9 29.117 6.812e-08 ***
I(log(kraavid2/pindala))  1   1499.2 2234.8  1.032    0.3098
```

Biomeetria bioloogidele

loeng 7

Üldistatud lineaarne mudel II
(*Generalized Linear Model – GLM*)

Üle- ja alahajuvus

Poissoni regressioon (Log-lineaarne mudel)

- Kui uuritav tunnus on saadud millegi loendamisel (diskreetne tunnus), siis peab uuritava tunnuse keskväärtus olema mittenegatiivne
  (-> kasutame log-seosefunktsiooni)
- Poissoni jaotuse puhul oli uuritava tunnuse dispersioon võrdne keskväärtusega, seega arvestamegi mudeli parameetrite hindamisel neid vaatluseid, mille puhul uuritava tunnuse keskväärtus on suurem, väiksema kaaluga

35

Kuna tekib Poissoni jaotus?

- Sündmused toimuvad sõltumatult (ühe sündmuse toimumine ei suurenda ega kahanda teiste sündmuste asetleidmise tõenäosust lähikonnas…);
- Samade sõltumatute tunnuste (*X*-tunnuste) väärtuste korral peab sündmuste asetleidmise „võimalused“ olema samasugused…

36

Kuna tekib Poissoni jaotus?

- Sündmused toimuvad sõltumatult (ühe sündmuse toimumine ei suurenda ega kahanda teiste sündmuste asetleidmise tõenäosust lähikonnas…);

+ Uute leukeemiajuhtude arv: aasta jooksul linnakeses aset leidvad uued leukeemiajuhud (kui üks inimene haigestub vähki, siis see iseenesest ei suurenda ega vähenda teiste inimeste vähkihaigestumise riski…)

+ Rikki läinud lambipirnide arv: lambi läbipõlemine õppehoones ei kutsu teistes lambipirnides esile surmasoovi…

37

Kuna tekib Poissoni jaotus?

- Sündmused toimuvad sõltumatult (ühe sündmuse toimumine ei suurenda ega kahanda teiste sündmuste asetleidmise tõenäosust lähikonnas…);

Lindude arv???
linnud kaitsevad koduterritooriumi…

38

Kuna tekib Poissoni jaotus?

- Sündmused toimuvad sõltumatult (ühe sündmuse toimumine ei suurenda ega kahanda teiste sündmuste asetleidmise tõenäosust lähikonnas…);

Lindude arv???

linnud otsivad elukaaslast…

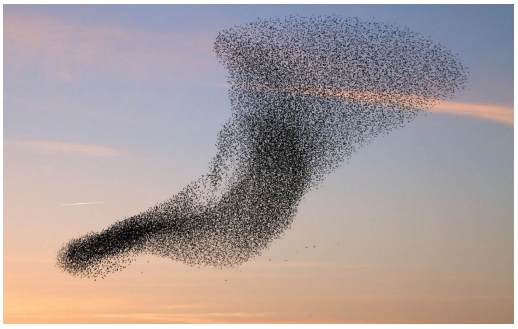
39

Kuna tekib Poissoni jaotus?

- Sündmused toimuvad sõltumatult (ühe sündmuse toimumine ei suurenda ega kahanda teiste sündmuste asetleidmise tõenäosust lähikonnas…);

Lindude arv???

Või võivad hoopis eelistada viibida seltskonnas…



40

Kuna tekib Poissoni jaotus?

- Samade sõltumatute tunnuste (*X*-tunnuste) väärtuste korral peab sündmuste asetleidmise „võimalused“ olema samasugused…

+ Uute leukeemiajuhtude arv (100000 elanikuga linnakeses): kui elanike vanuseline struktuur püsib ligikaudu sama ja naabruses ei ole hiljuti plahvatanud tuumajaama…

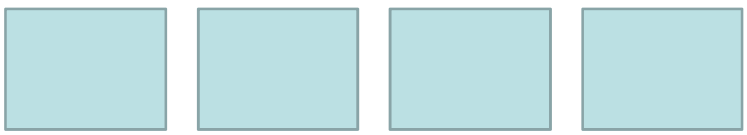
+ Nädala jooksul rikki läinud lambipirnide arv: Ühe töönädala jooksul kasutatakse/kurnatakse lambipirne ligilähedaselt sarnaselt, seega on eri nädalate „riskisus“ üsna hästi võrreldavad (ehkki jõulunädal võib vist veidi vähemriskantne olla tavalistele lambipirnidele?)

41

Kuna tekib Poissoni jaotus?

- Samade sõltumatute tunnuste (*X*-tunnuste) väärtuste korral peab sündmuste asetleidmise „võimalused“ olema samasugused…

- Linnukesi maatükil: Erinevad maatükid mis meile tunduvad (mõõdetud tunnuste järgi) sarnased…



42

Kuna tekib Poissoni jaotus?

- Samade sõltumatute tunnuste (*X*-tunnuste) väärtuste korral peab sündmuste asetleidmise „võimalused“ olema samasugused…

- Linnukesi maatükil: Erinevad maatükid mis meile tunduvad (mõõdetud tunnuste järgi) sarnased…

  … võivad lindude meelest olla vägagi erinevad…




43

Kõhklused...

Õnnetused ratsaväe ohvitseridega	**Rukkirääkgud**
Õnnetused toimuvad teineteisest sõltumatult	Kas linnud / loomad elavad teineteist segamata? Tundmata tõmmet teineteise vastu?
Õnnetusrisk samasugune	Vaatamata meie pingutustele jäävad osad põllud rukkirääkudele sobivamaks kui teised...

44

Probleem

Poissoni regressiooni kasutades oleme eeldanud, et vaatluste dispersioon (samade x-tunnuse väärtuste korral) on võrdne keskväärtusega:

$$D(Y|X) = E(Y|X)$$

Kui aga uuritav tunnus pole päriselt Poissoni jaotusega…???

45

Probleem

Poissoni jaotuse (Poissoni regressiooni) korral:

$$D(Y|X) = E(Y|X)$$

Sageli aga:

$D(Y|X) << E(Y|X)$ Alahajuvus *(underdispersion)*

või

$D(Y|X) >> E(Y|X)$ Ülehajuvus *(overdispersion)*

46

Kuidas märgata probleeme?

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

Call:
glm(formula = raak ~ 1, family = poisson(), offset = log(pindala))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7763  -0.4916  -0.2858  -0.1557   3.2826

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.16475    0.04746  -87.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1632.2  on 3307  degrees of freedom
Residual deviance: 1632.2  on 3307  degrees of freedom
AIC: 2359.9

Number of Fisher Scoring iterations: 6
```

Rusikareegel: kui mudeliga kõik korras, siis

Residual deviance ≈ df

47

Kuidas märgata probleeme?

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

Call:
glm(formula = raak ~ 1, family = poisson(), offset = log(pindala))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7763  -0.4916  -0.2858  -0.1557   3.2826

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.16475    0.04746  -87.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1632.2  on 3307  degrees of freedom
Residual deviance: 1632.2  on 3307  degrees of freedom
AIC: 2359.9

Number of Fisher Scoring iterations: 6
```

Kui

Residual deviance << df

siis alahajuvus??

Residual deviance >> df

siis ülehajuvus??

48

Mis siis, kui *Residual deviance ≠ df*?

Alahajuvus (*underdispersion*)
Residual deviance << df

- “Sündmused negatiivselt korreleeritud”, näiteks koduterritooriumi tõttu elab 1 hektaril maksimaalselt vaid 1 isasloom. Sellisel juhul 10ha maatükil võime näha kõige enam 10 isast loomakest, aga mitte 13 või 16 (mis Poissoni jaotuse puhul oleks täiesti võimalik). Poissoni regressioon ei sobi, tarvis muuta analüüsi.
- Nullide mittelugemine (vahel ei panda andmestikku kirja juhtumeid, kus esines 0 “sündmust”). Poissoni regressioon ei sobi, tarvis muuta analüüsi.
- On vaatluseid, mille puhul Poissoni regressioon hindab sündmuse esinemist võimatuks (näiteks maalappe, kus Poissoni regressiooni arvates kohtame ootuspäraselt ≈0 rukkirääku...). Sellisel juhul tutvustatud rusikareegel ei tööta...

```
 > sum(predict(mudel1, type="response")<0.1)
[1] 2090
```

49

Mis siis, kui *Residual deviance ≠ df*?

Ülehajuvus (*overdispersion*)
Residual deviance >> df

- Mudel pole täielik, kokku on pandud väga erinevate tingimustega (väga erineva keskväärtusega) alad. Vaja kas mudelit täiendada täiendavate tunnustega või tuleks kasutada Poissoni regressiooni asemel midagi muud...
- “Sündmused” on “positiivselt korreleeritud”, st kui näeme juba ühte linnukest, siis on arvatavasti sellel samal maalapil kümneid kui mitte sadu linnukesi (sest linnuparv on maandunud justnimelt sellele maalapikesele). Poissoni regressiooni asmel tuleks arvatavasti kasutada midagi muud...
- Andmestikus on liigseid nulle, näiteks tänu sellele, et oleme lugenud rukkirääke kokku ka maadelt, mis põhimõtteliselt sobiksid rukkiräägule, aga kuhu rukkiräägu leviala ei ulatu (näiteks need põllud Põhja-Ameerikas, mis said andmestikku lisatud...). Poissoni regressiooni ei tohiks kasutada...

50

pscl

Mida teha, kui esineb ala- või ülehajuvus?

- Kasutada teste, mis taluvad vigu hajuvuse modelleerimisel;
- Poissoni regressiooni asemel hinda mõni teine mudel, mis lubab keerukamat seost dispersiooni ja keskväärtuse vahel (näiteks kvaasi-Poissoni mudel või negatiivsel binoomjaotusel baseeruv mudel);
- Kasutada mudelit, mis lubab liigseid nulle (*zero-inflated Poisson regressioon*) või puuduvaid nulle (*zero-truncated Poisson regressioon*). R-is kättesaadavad näiteks lisamoodulis *pscl*.

51

Kvaasi-Poisson ja negatiivne binoomjaotus

- Kvaasi-Poissoni mudel (quasi-Poisson)

$$D(Y) = c\,E(Y)$$

konstandi c väärtus hinnatakse.

Võimaldab modelleerida nii ala- kui ka ülehajuvust.

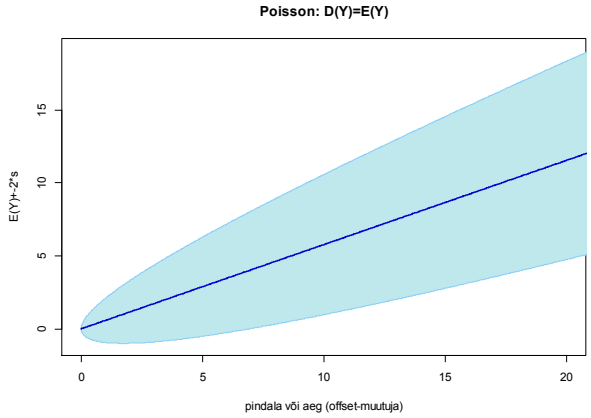
- Negatiivsel binoomjaotusel baseeruv mudel

$$D(Y) = E(Y) + (E(Y))^2/\theta$$

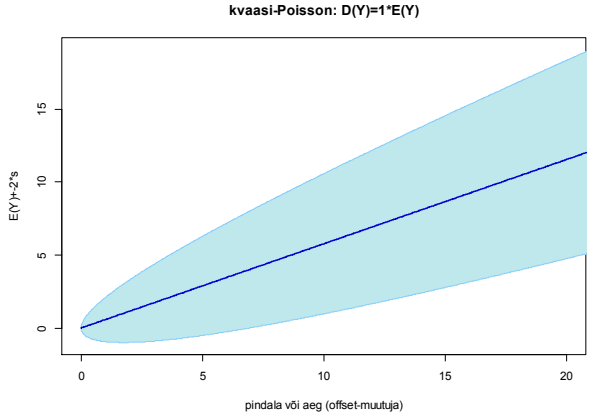
konstandi θ väärtus hinnatakse.

Võimaldab modelleerida ainult ülehajuvust

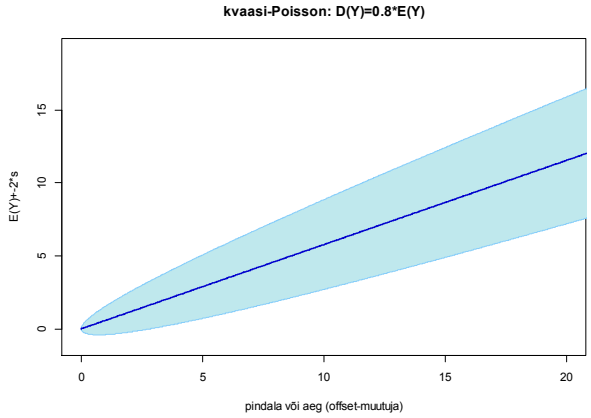
52


Poisson: D(Y)=E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

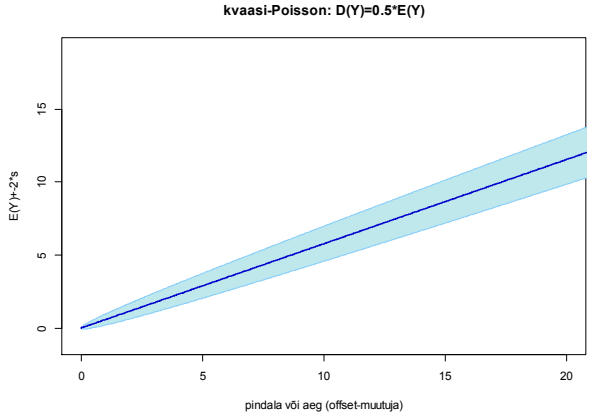
53


kvaasi-Poisson: D(Y)=1*E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

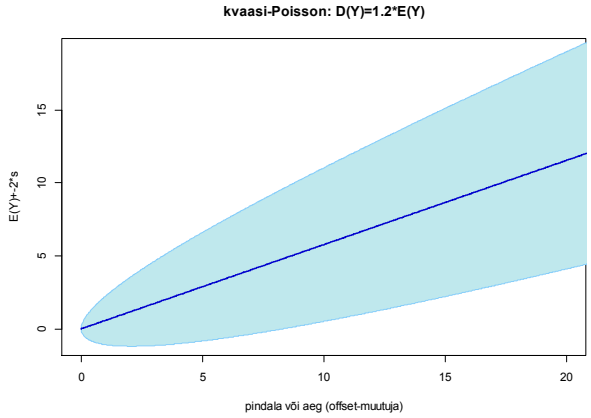
54


kvaasi-Poisson: D(Y)=0.8*E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

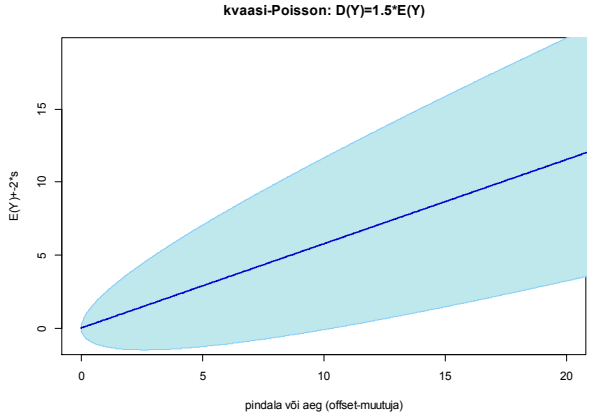
55


kvaasi-Poisson: D(Y)=0.5*E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

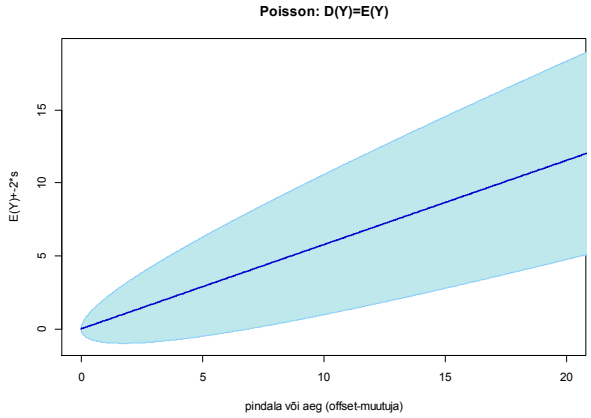
56


kvaasi-Poisson: D(Y)=1.2*E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

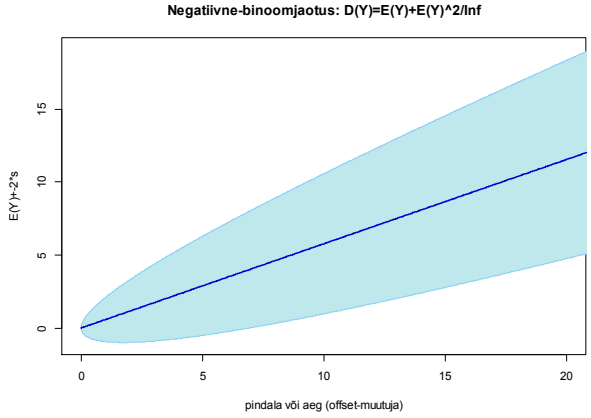
57


kvaasi-Poisson: D(Y)=1.5*E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

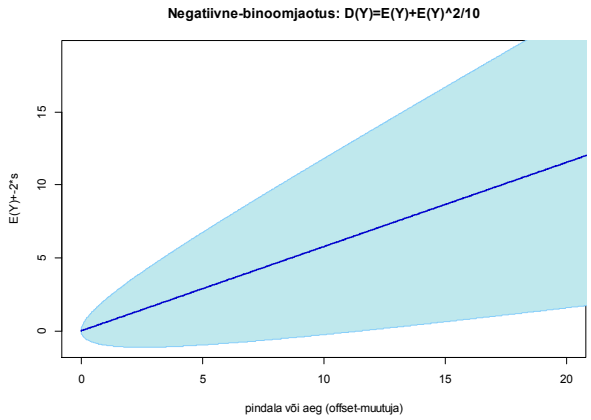
58


Poisson: D(Y)=E(Y)
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

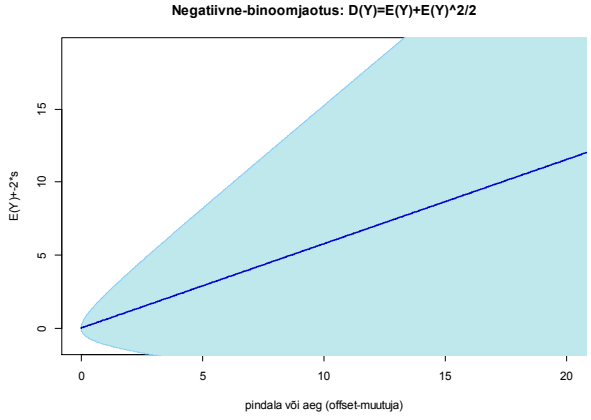
59


Negatiivne-binoomjaotus: D(Y)=E(Y)+E(Y)^2/Inf
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

60

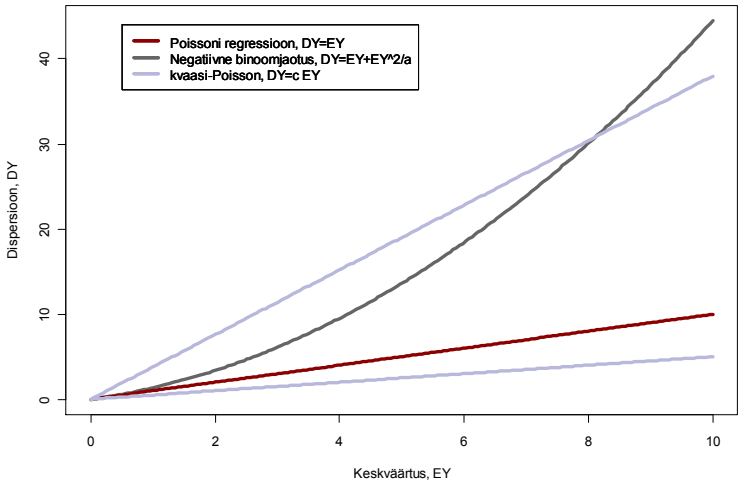

Negatiivne-binoomjaotus: D(Y)=E(Y)+E(Y)^2/10
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

61


Negatiivne-binoomjaotus: D(Y)=E(Y)+E(Y)^2/2
E(Y)+-2*s
0
5
10
15
0
5
10
15
20
pindala või aeg (offset-muutuja)

62

Paar levinud alternatiivi Poissoni regressioonile



63

Negatiivne binoomjaotus või kvaasi-Poissoni mudel

Näide olukorrast, kus lindude arvu jaotus võiks olla negatiivse binoomjaotusega:

Oletame, et mõned põllud sobivad rukkirääkudele märksa enam kui teised põllud. Miks see nii on, seda me ei mõista, oluline põllu sobivust kirjeldav tunnus on puudu (jäänud mõõtmata). Sellisel juhul on meie mudeli seisukohast võrdväärsed alad tegelikult väga erineva sobivusega, väga heterogeensed. Kui uuritava tunnuse jaotuseks on tegelikult väga erinevate Poissoni jaotuste segu (teatud viisil moodustatud segu), siis on lindude arvu jaotuseks negatiivne binoomjaotus.

64

Negatiivne binoomjaotus või kvaasi-Poissoni mudel

Näide olukorrast, kus lindude arvu jaotust võiks modelleerida kasutades kvaasi-Poissoni mudelit:

Kvaasi-Poissoni mudel võiks sobida lindude arvu kirjeldamiseks näiteks siis, kui linnupesade arv mingil alal on küll Poissoni jaotusega, aga iga linnupesaga kaasneb 2 täiskasvanud lindu (ema ja isa). Sellisel juhul on lindude arvu jaotuse hajuvus suurem kui Poissoni jaotus seda ette näeb, kuid sõltub siiski lineaarselt keskväärtusest (DY= 2 EY).

Märkus: kvaasi-Poissoni jaotust kui sellist ei eksisteeri. Kvaasi-Poissoni mudel ei esita uuritava tunnuse jaotusele mingeid nõudmisi. Ainsaks eelduseks on, et uuritava tunnuse dispersioon on proportsionaalne keskväärtusega, D*Y* = c * E*Y.*

65

Mudelite hindamine

Kvaasi-Poissoni mudel (quasi-Poisson regression)

```
mudel_kvaasiPoisson=(glm(raak~1,
  offset=log(pindala), family=quasipoisson()))
```

Või, alternatiivina:

```
mudel_kvaasiPoisson=(glm(raak~1,offset=log(pindala),
  family=quasi(link="log", var="mu")))
```

Negative Binomial Regression

```
library(MASS)
mudel_nb=glm.nb(raak~1+offset(log(pindala)))
```

66

Tulemused

Poisson

hinnang	std.viga
-4.16475	0.04746

kvaasi-Poisson

hinnang	std.viga
-4.16475	0.04973

Ülehajuvusparameeter: 1.098

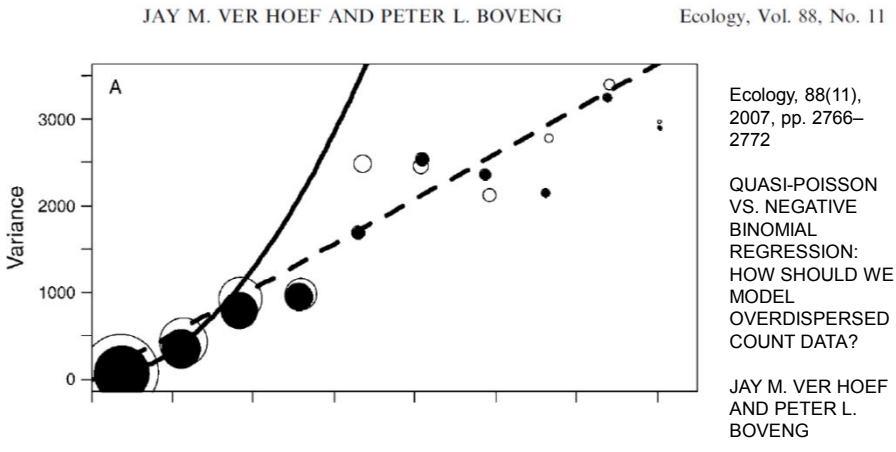
```
> sqrt(0.04746**2*1.098)
[1] 0.0497312
```

neg. binoom

hinnang	std.viga
-4.12553	0.05637

67

Negatiivne binoomjaotus või kvaasi-Poissoni mudel?



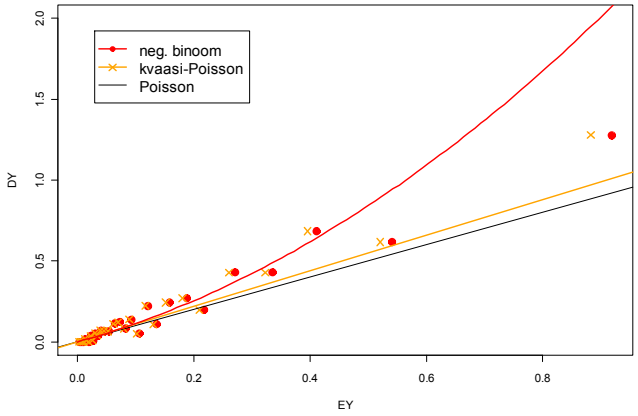
Ecology, 88(11), 2007, pp. 2766–2772

QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA?

JAY M. VER HOEF AND PETER L. BOVENG

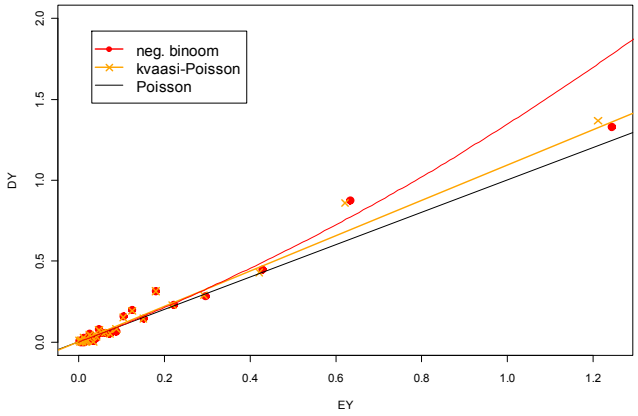
68

Rukkiräägud – lihtne mudel



69

Rukkiräägud – keerukam mudel



70

Tulemuste võrdlus keerukama mudeli korral

	Poisson			kvaasi-Poisson			neg. binoom		
	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
b0	-7.343	1.039	0.000	-7.343	1.086	0.000	-7.421	1.047	0.000
b1	-0.701	0.212	0.001	-0.701	0.222	0.002	-0.679	0.226	0.003
b2	-0.535	0.424	0.207	-0.535	0.443	0.227	-0.456	0.427	0.285
b3	-0.163	0.196	0.404	-0.163	0.205	0.425	-0.150	0.209	0.474
b4	1.413	0.231	0.000	1.413	0.241	0.000	1.463	0.242	0.000
b5	1.085	1.159	0.349	1.085	1.212	0.371	1.105	1.166	0.343
b6	2.646	1.002	0.008	2.646	1.048	0.012	2.666	1.006	0.008
b7	2.358	1.006	0.019	2.358	1.052	0.025	2.355	1.010	0.020
b8	-0.023	1.417	0.987	-0.023	1.481	0.988	-0.019	1.422	0.989
b9	0.225	1.100	0.838	0.225	1.150	0.845	0.228	1.106	0.836
b10	0.255	1.064	0.810	0.255	1.112	0.818	0.256	1.069	0.811
b11	0.574	1.074	0.593	0.574	1.122	0.609	0.575	1.080	0.594
b12	-0.116	0.040	0.003	-0.116	0.042	0.005	-0.112	0.042	0.008
b13	0.177	0.054	0.001	0.177	0.056	0.002	0.172	0.057	0.002
b14	0.135	0.100	0.180	0.135	0.105	0.199	0.118	0.101	0.245
b15	0.137	0.051	0.007	0.137	0.053	0.010	0.140	0.054	0.009

71

Programm eeltoodud jooniste tegemiseks

```
m_qp=glm(raak~1, offset=log(pindala), family=quasipoisson())
m_nb=glm.nb(raak~1+offset(log(pindala)))
jaak_qp=residuals(m_qp, type="response")
jaak_nb=residuals(m_nb, type="response")
prognoos_qp=predict(m_qp, type="response")
prognoos_nb=predict(m_nb, type="response")

progklass_qp=cut(prognoos_qp, quantile(prognoos_qp, seq(0,1, length=30)))
progklass_nb=cut(prognoos_nb, quantile(prognoos_nb, seq(0,1, length=30)))

y_qp=as.vector(by(jaak_qp, progklass_qp, var))
y_nb=as.vector(by(jaak_nb, progklass_nb, var))

x_qp=as.vector(by(prognoos_qp, progklass_qp, mean))
x_nb=as.vector(by(prognoos_nb, progklass_nb, mean))

plot(x_nb, y_nb, type="p", pch=20, col="red", cex=2, ylim=c(0,2), xlab="EY", ylab="DY")
points(x_qp, y_qp, pch=4, col="orange", cex=1.3, lwd=2)

abline(coef=c(0,1))
abline(coef=c(0,1.098123), lwd=2, col="orange")

xx=seq(0,2, length=200)
yy=xx+xx^2/0.732

lines(xx,yy, col="red", lwd=2)

legend(0.03, 1.92, c("neg. binoom", "kvaasi-Poisson", "Poisson"),pch=c(20,4,NA), lty=c(1,1,1),
    col=c("red","orange","black"), lwd=c(2,2,1), cex=1.3)
```

72

Mõned detailid I

```
abline(coef=c(0,1.098123), lwd=2, col="orange")

> summary(m_qp)
[...]
(Dispersion parameter for quasipoisson family taken
  to be 1.098123)
```

73

Mõned detailid II

```
xx=seq(0,2, length=200)
yy=xx+xx^2/0.732
```

$D(Y)=E(Y)+E(Y)^2/\theta$

```
> summary(m_nb)
[...]

              Theta:  0.732
          Std. Err.:  0.142
```

74

Praktilisi märkusi

- Negatiivset binoomjaotust kasutav mudel võib olla raskemini hinnatav (hindamine ei pruugi alati õnnestuda) võrreldes kvaasi-Poissoni mudeliga
- Negatiivse binoomjaotuse abil ei saa modelleerida alahajuvust.
- Kvaasi-Poissoni mudel on ~jaotusvaba. Võivad tekkida probleemid näitajatega, mis eeldavad uuritava tunnuse jaotuse teadmist (raskused AIC, tõepära leidmisel)

75

Negatiivsel binoomjatusel baseeruva mudeli ja Poissoni regressiooni võrdlus

```
> m_nb=glm.nb(raak~factor(toetustyyp)+...+
          offset(log(pindala)))
> m_p=glm(raak~factor(toetustyyp)+...,
     offset=log(pindala), family=poisson())

> AIC(m_p)
[1] 2062.781
> AIC(m_nb)
[1] 2054.912
```

Negatiivset binoomjaotust kasutava mudeli AIC-väärtus on väiksem (väiksem on parem), seega on vastava mudeli (uute vaatluste) prognoositäpsus (arvatavasti) suurem – eelista negatiivset binoomjaotust tavalisele Poissoni regressioonile.

76

Testid (eeldavad lisamoodulit)

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

Poisson vs kvaasi-Poisson

```
> dispersiontest(m_p,trafo=1)

        Overdispersion test

data:  m_p
z = 3.4725, p-value = 0.0002578
alternative hypothesis: true alpha is greater
than 0
sample estimates:
    alpha
0.1049018
```

77

Testid (eeldavad lisamoodulit)

```
> dispersiontest(m_p,trafo=2)

        Overdispersion test

data:  m_p
z = 5.0688, p-value = 2.002e-07
alternative hypothesis: true alpha is greater
than 0
sample estimates:
    alpha
0.6217027
```

Poisson vs negatiivne binoomjaotus

78

Kvaasi-Poissoni ja Poissoni regressiooni võrdlus

Märkus:

Kui kvaasi-Poissoni mudel hindab dispersiooniparameetri (*Dispersion parameter for quasipoisson family taken to be…*) väärtuseks 4, siis kõigi parameetrite standardhälvete hinnangud on sqrt(4)=2 korda suuremad Poissoni regressiooni abil saadud hinnangutest.

Kui dispersiooniparameeter on ligikaudu 1, siis langevad kvaasi-Poissoni ja Poissoni mudeli baasil leitud testid/usaldusintervallid enam-vähem kokku.

79

Märkus mudeli headuse kirjeldamise kohta

Head alternatiivi lineaarsetest mudelitest tuntud determinatsioonikordajale R^2-le pole. Soovi korral võib kasutada näiteks McFadden'i R^2-tu:

```
> summary(m1)
[...]
Null deviance: 504.12  on 395  degrees of freedom
Residual deviance: 397.94  on 392  degrees of freedom
```

McFadden'i R^2 saab leida järgmise arvutuse abil:

```
1-397.94/504.12 = 0,2106...
```

McFadden'i R^2 on küll vahemikus 0 (kasutu mudel) kuni 1 (ajalooliste andmete perfektne prognoos), kuid ühtegi väga ilusat ja kergesti hoomatavat interpretatsiooni antud statistikule anda ma ei oska.

80