

Mudelid II

Log-lineaarne mudel

Poissoni regressioon
Kvalitatiivsete andmete analüüs
7. loeng

Meenutuseks

- Lineaarne mudel

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

- Üldistatud lineaarne mudel

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Kui teame, et keskvärtus ei saa olla negatiivne (uuritavaks tunnuseks on sagedused):

$$\log(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Kui teame, et keskvärtus on vahemikus 0...1:

$$\log(E(Y)/(1 - E(Y))) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Meenutuseks

- Lineaarne mudel

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

- Üldistatud lineaarne mudel

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Kui teame, et keskvärtus ei saa olla negatiivne (uuritavaks tunnuseks on sagedused):

$$\log(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Kui teame, et keskvärtus on vahemikus 0...1:

$$\log(p/(1 - p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Näide (2020.a. andmed)

					Maakond	juhte
Põlva	Tartu	Valga	Viljandi	Võru	Põlva maakond	80
					Tartu maakond	442
					Valga maakond	87
					Viljandi maakond	164
					Võru maakond	119

Kas *i*. maakonna esmajuhtude arvu
kirjeldab mudel 1 või mudel 2:

$$\log(E(juhte_i)) = \beta_0$$

```
m0=glm(juhte~1, family=poisson())
```

$$E(juhte_i) = \exp(\beta_0) = konst$$

$$\log(E(juhte_i)) = \beta_0 + \beta_i$$

```
m1=glm(juhte~factor(Maakond),  
family=poisson())
```

```
> anova(m0, m1, test="LRT")  
Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1         4      424.79  
2         0         0.00  4    424.79 < 2.2e-16 ***
```

Näide (2020.a. andmed)

					Maakond juhte elanikke	
					Põlva maakond	80 23496
Põlva	Tartu	Valga	Viljandi	Võru	Tartu maakond	442 128643
80	442	87	164	119	Valga maakond	87 22649
					Viljandi maakond	164 44044
					Võru maakond	119 33553

Kas i. maakonna esmajuhtude arvu kirjeldab mudel 1 või mudel 2:

E(juhte_i) = konst · elanikke

E(juhte_i) = konst_i · elanikke

Näide (2020.a. andmed)

					Maakond juhte elanikke	
					Põlva maakond	80 23496
Põlva	Tartu	Valga	Viljandi	Võru	Tartu maakond	442 128643
80	442	87	164	119	Valga maakond	87 22649
					Viljandi maakond	164 44044
					Võru maakond	119 33553

Kas i. maakonna esmajuhtude arvu kirjeldab mudel 1 või mudel 2:

log(E(juhte_i)) = beta_0 + log(konst) + log(elanikke)

E(juhte_i) = konst · elanikke

log(E(juhte_i)) = beta_0 + beta_i + log(konst_i) + log(elanikke)

E(juhte_i) = konst_i · elanikke

```
> anova(m0, m1, test="Chisq")
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4      1.4958
2      0      0.0000 4      1.4958 0.8274
```

Offset - näiteid

Uuritav tunnus	offset
Võilillede arv aasal	log(aasa pindala)
katkiläinud toodete arv	log(müüdnud toodete arv)
paanikahoogude arv patsiendil	log(jälgimisaeg)
liiklusõnnetuste arv teel	log(tee pikkus)+log(jälgimisaeg)

Muud jaotused

Sageduste modelleerimiseks (parameetrite hindamiseks) võib kasutada ka teisi jaotuseid – aga seosefunktsioon on alati log, sest keskmine sagedus (üle erinevate valimite) ei saa kunagi olla negatiivne:

log(E(juhte_i)) = beta_0 + beta_1 I(vana) + log(Inimesi)

```
m1=glm(juhte~factor(vanus)+offset(log(Inimesi)), family=poisson())
m1a=glm(juhte~factor(vanus)+offset(log(Inimesi)), family=quasipoisson())
m1b=glm.nb(juhte~factor(vanus)+offset(log(Inimesi)))
```

Poisson	Estimate	Std. Error	z value	Pr(> z)
m1	(Intercept)	-7.46064	0.06757	-110.41 <2e-16
	factor(grupp) vana	2.63581	0.06930	38.04 <2e-16

Kvaasi-Poisson	Estimate	Std. Error	t value	Pr(> t)
m1a	(Intercept)	-7.4606	0.7184	-10.385 4.15e-11
	factor(grupp) vana	2.6358	0.7367	3.578 0.00129

neg. binoom.	Estimate	Std. Error	z value	Pr(> z)
m1b	(Intercept)	-7.4337	0.1835	-40.51 <2e-16
	factor(grupp) vana	2.7068	0.2420	11.19 <2e-16

Kahemõõtmeline sagedustabel.

Y ↓ \ X →	A	B	C
on	12	25	36
pole	23	50	97

Kas õhtuks väsinud

Kolm erinevat hommikusöögi

Kahemõõtmeline sagedustabel.

Y ↓ \ X →	A	B	C
on	12 $\approx n\pi_{11}$	25 $\approx n\pi_{21}$	36 $\approx n\pi_{31}$
pole	23 $\approx n\pi_{12}$	50 $\approx n\pi_{22}$	97 $\approx n\pi_{32}$

oodatavad sagedused

Kombinatsiooni $X = x_i$ ja $Y = y_j$ esinemistõenäosus:

$$P(X = x_i \cap Y = y_j) = \pi_{ij}.$$

Ootuspärane vaatluste arv i . veerus ja j . reas:

$$E(sagedus_{ij}) = n\pi_{ij}.$$

vaatluste koguarv

Kahemõõtmeline sagedustabel.

Y ↓ \ X →	A	B	C
on	12 $\approx n\pi_{11}$	25 $\approx n\pi_{21}$	36 $\approx n\pi_{31}$
pole	23 $\approx n\pi_{12}$	50 $\approx n\pi_{22}$	97 $\approx n\pi_{32}$

Šanss olla väsinud hommikusöögi A korral:

$$\frac{P(on|X=A)}{P(pole|X=A)} = \frac{P(on \cap X=A)/P(X=A)}{P(pole \cap X=A)/P(X=A)} = \frac{\pi_{11}}{\pi_{12}} = \frac{n \cdot \pi_{11}}{n \cdot \pi_{12}} = \frac{\mu_{11}}{\mu_{12}}$$

Šanssi saab leida oodatavate sageduste kaudu!

Kahemõõtmeline sagedustabel. Sõltumatus.

Y ↓ \ X →	A	B	C
on	12 $\approx n\pi_{11}$	25 $\approx n\pi_{21}$	36 $\approx n\pi_{31}$
pole	23 $\approx n\pi_{12}$	50 $\approx n\pi_{22}$	97 $\approx n\pi_{32}$

$$\pi_{ij} = \pi_{i+} \pi_{+j} \quad \mu_{ij} = n\pi_{ij} = n\pi_{i+} \pi_{+j}$$

Tõenäosus näha i . hommikusööki sõõnud inimest

Tõenäosus näha Y tunnuse j . väärtust

Šanss olla väsinud i . hommikusöögi korral:

$$\text{šanss}_i = \mu_{i1}/\mu_{i2} \approx 12/23$$

lihtne hinnang šansile

Šansside suhe, i . vs i' .:

$$OR = \frac{\text{šanss}_i}{\text{šanss}_{i'}} = \frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}$$

Šanss olla väsinud i . hommikusöögi korral:

$$\text{šanss}_i = \mu_{i1}/\mu_{i2}$$

Šansside suhe, i . vs i' .:

$$OR = \frac{\text{šanss}_i}{\text{šanss}_{i'}} = \frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}$$

Kahemõõtmeline sagedustabel. Sõltumatus.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y$

$$\pi_{ij} = \pi_i + \pi_{+j}$$

Tõenäosus näha *i*.
hommikusöögi
söönud inimest

Tõenäosus näha *Y*
tunnuse *j*. väärtust

$$\mu_{ij} = n\pi_{ij} = n\pi_i + \pi_{+j}$$

$$\log(\mu_{ij}) = \log(n) + \log(\pi_{i+}) + \log(\pi_{+j})$$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Šanss olla väsinud *i*. hommikusöögi korral:

$$\text{šanss}_i = \mu_{i1} / \mu_{i2}$$

Šansside suhe, *i*. vs *i'* .:

$$OR = \frac{\text{šanss}_i}{\text{šanss}_{i'}} = \frac{\mu_{i1} / \mu_{i2}}{\mu_{i'1} / \mu_{i'2}}$$

Kahemõõtmeline sagedustabel. Sõltumatus.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y$

$$\pi_{ij} = \pi_i + \pi_{+j}$$

Tõenäosus näha *i*.
hommikusöögi
söönud inimest

Tõenäosus näha *Y*
tunnuse *j*. väärtust

$$\mu_{ij} = n\pi_{ij} = n\pi_i + \pi_{+j}$$

$$\log(\mu_{ij}) = \log(n) + \log(\pi_{i+}) + \log(\pi_{+j})$$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

$\log(\text{Šanss})$ olla väsinud *i*. hommikusöögi korral:

$$\log(\text{šanss}_i) = \log(\mu_{i1} / \mu_{i2}) = \log(\mu_{i1}) - \log(\mu_{i2})$$

$$\log(\text{šanss}_i) = \lambda + \lambda_i^X + \lambda_1^Y - (\lambda + \lambda_i^X + \lambda_2^Y)$$

$$\log(\text{šanss}_i) = \lambda_1^Y - \lambda_2^Y$$

Šansside suhte logaritm:

$$\log(OR) = \log\left(\frac{\text{šanss}_i}{\text{šanss}_{i'}}\right) = \log\left(\frac{\mu_{i1} / \mu_{i2}}{\mu_{i'1} / \mu_{i'2}}\right)$$

$$\log(OR) = \lambda_1^Y - \lambda_2^Y - (\lambda_1^Y - \lambda_2^Y) = 0$$

Kahemõõtmeline sagedustabel. Sõltumatus.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y$

$$\pi_{ij} = \pi_i + \pi_{+j}$$

Tõenäosus näha *i*.
hommikusöögi
söönud inimest

Tõenäosus näha *Y*
tunnuse *j*. väärtust

$$\mu_{ij} = n\pi_{ij} = n\pi_i + \pi_{+j}$$

$$\log(\mu_{ij}) = \log(n) + \log(\pi_{i+}) + \log(\pi_{+j})$$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

$$= (\lambda + c) + (\lambda_i^X - c) + \lambda_j^Y$$

$$= (\lambda + \lambda_1^X) + (\lambda_i^X - \lambda_1^X) + \lambda_j^Y$$

$$= (\lambda + \lambda_1^X + \lambda_1^Y) + (\lambda_i^X - \lambda_1^X) + (\lambda_j^Y - \lambda_1^Y)$$

$$= \lambda^* + \lambda_i^{*X} + \lambda_j^{*Y}$$

$\log(\text{Šanss})$ olla väsinud *i*. hommikusöögi korral:

$$\log(\text{šanss}_i) = \log(\mu_{i1} / \mu_{i2}) = \log(\mu_{i1}) - \log(\mu_{i2})$$

$$\log(\text{šanss}_i) = \lambda + \lambda_i^X + \lambda_1^Y - (\lambda + \lambda_i^X + \lambda_2^Y)$$

$$\log(\text{šanss}_i) = \lambda_1^Y - \lambda_2^Y$$

Kahemõõtmeline sagedustabel. Sõltumatus.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

mõjud mida
hinnangute
tabelis pole
toodud on
nullid

$$\lambda_{pole}^Y = 0,8453$$

$$\lambda_{on}^Y = 0$$

$\log(\text{Šanss})$ olla väsinud *i*. hommikusöögi korral:

$$\log(\text{šanss}_i) = \lambda_1^Y - \lambda_2^Y = 0 - 0,8453$$

```
ml=glm(sagedused~factor(Y)+factor(X),
family=poisson())
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.3527	0.1953	12.045	< 2e-16
factor(Y) pole	0.8453	0.1399	6.041	1.53e-09
factor(X) B	0.7621	0.2047	3.723	0.000197
factor(X) C	1.3350	0.1900	7.027	2.11e-12

Hinnang väsinud olemise šansile leitud üle kõigi
hommikusööjate – kui hommikusöögil pole
mõju, siis võime kõiki inimesi kasutada šansi
hindamiseks

$$(12+25+36) / (23+50+97) = 0.4294...$$

$$\text{šanss}_i = \exp(-0.8453) = 0.4294...$$

Kahemõõtmeline sagedustabel. Sõltuvad tunnused.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y + \lambda_{31}^{XY}$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y + \lambda_{32}^{XY}$

$\pi_{ij} \neq \pi_i + \pi_j$ Parandusliige

$\pi_{ij} = \pi_i + \pi_j + p_{ij}$

$\mu_{ij} = n\pi_{ij} = n\pi_i + \pi_j p_{ij}$

$\log(\mu_{ij}) = \log(n) + \log(\pi_i) + \log(\pi_j) + \log(p_{ij})$

$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{i,j}^{X,Y}$

$\log(\mu_{i1}/\mu_{i2}) = \lambda + \lambda_i^X + \lambda_1^Y + \lambda_{i,1}^{XY} - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_{i,2}^{XY})$

$= \lambda_1^Y + \lambda_{i,1}^{XY} - \lambda_2^Y - \lambda_{i,2}^{XY}$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \log\left(\frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}\right)$

Kahemõõtmeline sagedustabel. Sõltuvad tunnused.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y + \lambda_{31}^{XY}$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y + \lambda_{32}^{XY}$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \lambda_1^Y + \lambda_{i,1}^{XY} - \lambda_2^Y - \lambda_{i,2}^{XY} - (\lambda_1^Y + \lambda_{i',1}^{XY} - \lambda_2^Y - \lambda_{i',2}^{XY})$

$= \lambda_{i,1}^{XY} - \lambda_{i,2}^{XY} - \lambda_{i',1}^{XY} + \lambda_{i',2}^{XY}$

$\log(\mu_{i'1}/\mu_{i'2}) = \lambda_1^Y + \lambda_{i',1}^{XY} - \lambda_2^Y - \lambda_{i',2}^{XY}$

$\log(\mu_{i1}/\mu_{i2}) = \lambda + \lambda_i^X + \lambda_1^Y + \lambda_{i,1}^{XY} - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_{i,2}^{XY})$

$= \lambda_1^Y + \lambda_{i,1}^{XY} - \lambda_2^Y - \lambda_{i,2}^{XY}$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \log\left(\frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}\right)$

Kahemõõtmeline sagedustabel. Sõltuvad tunnused.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y + \lambda_{31}^{XY}$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y + \lambda_{32}^{XY}$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \lambda_1^Y + \lambda_{i,1}^{XY} - \lambda_2^Y - \lambda_{i,2}^{XY} - (\lambda_1^Y + \lambda_{i',1}^{XY} - \lambda_2^Y - \lambda_{i',2}^{XY})$

$= \lambda_{i,1}^{XY} - \lambda_{i,2}^{XY} - \lambda_{i',1}^{XY} + \lambda_{i',2}^{XY}$

$\log(\mu_{i'1}/\mu_{i'2}) = \lambda_1^Y + \lambda_{i',1}^{XY} - \lambda_2^Y - \lambda_{i',2}^{XY}$

$\log(\mu_{i1}/\mu_{i2}) = \lambda + \lambda_i^X + \lambda_1^Y + \lambda_{i,1}^{XY} - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_{i,2}^{XY})$

$= \lambda_1^Y + \lambda_{i,1}^{XY} - \lambda_2^Y - \lambda_{i,2}^{XY}$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \log\left(\frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}\right) = \lambda_{i,1}^{XY} - \lambda_{i,2}^{XY} - \lambda_{i',1}^{XY} + \lambda_{i',2}^{XY}$

Kahemõõtmeline sagedustabel. Sõltuvad tunnused.

Y ↓ \ X →	A	B	C
on	$\log(12) \approx \lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$	$\log(25) \approx \lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}$	$\log(36) \approx \lambda + \lambda_3^X + \lambda_1^Y + \lambda_{31}^{XY}$
pole	$\log(23) \approx \lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}$	$\log(50) \approx \lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}$	$\log(97) \approx \lambda + \lambda_3^X + \lambda_2^Y + \lambda_{32}^{XY}$

```
m1=glm(s~Y+X+Y*X, family=poisson())

(25/50) / (12/23)= 0.95833

OR(B vs A) =
= exp(-0.04256)
= 0,95833
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.48491	0.28868	8.608	< 2e-16 ***
Ypole	0.65059	0.35611	1.827	0.067708 .
XB	0.73397	0.35119	2.090	0.036622 *
XC	1.09861	0.33333	3.296	0.000981 ***
Ypole:XB	0.04256	0.43222	0.098	0.921561
Ypole:XC	0.34060	0.40608	0.839	0.401600

$\log(OR(B vs A)) = 0 - 0.04256 - 0 + 0$

$\log(OR) = \log\left(\frac{\check{\text{san}}s_i}{\check{\text{san}}s_{i'}}\right) = \log\left(\frac{\mu_{i1}/\mu_{i2}}{\mu_{i'1}/\mu_{i'2}}\right) = \lambda_{i,1}^{XY} - \lambda_{i,2}^{XY} - \lambda_{i',1}^{XY} + \lambda_{i',2}^{XY}$

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

Dayton, Ohio, USA.
keskkooli lõpetajad
(abituriendid),
maapiirkond.

Alkohol	Suits	Marihuaana	n
+	+	+	911
+	+	-	583
+	-	+	44
+	-	-	456
-	+	+	3
-	+	-	43
-	-	+	2
-	-	-	279

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m0=glm(n~factor(alko)+factor(suits)+factor(mari),
        family=poisson())
summary(m0)
```

Coefficients:

	Estimate	Std. Error	z	Pr(> z)
(Intercept)	4.173	0.065	64.2	< 2e-16
factor(alko)+	1.785	0.060	29.9	< 2e-16
factor(suits)+	0.649	0.044	14.7	< 2e-16
factor(mari)+	-0.315	0.042	-7.4	1.08e-13

	alko	suits	mari	n
1	+	+	+	911
2	+	+	-	538
3	+	-	+	44
4	+	-	-	456
5	-	+	+	3
6	-	+	-	43
7	-	-	+	2
8	-	-	-	279

Alkoholi tarvitajaid on
 $\exp(1.785) = 5,96$ korda rohkem
kui mittetarvitajaid (alkoholi
tarvitamise shanss on 5,96)

Antud mudeli järgi on alkoholi tarvitamise shanss alati 5,96
korda suurem. Nii suitsetajate kui ka mittesuitsetajate seas
on alkoholi tarbijaid 5,96 korda rohkem kui mittetarbijaid.

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m1=glm(n~factor(mari)+factor(suits)*factor(alko), family=poisson())
summary(m1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.09053	0.06228	81.732	< 2e-16 ***
factor(mari)+	-0.31542	0.04244	-7.431	1.08e-13 ***
factor(suits)+	-1.80971	0.15905	-11.378	< 2e-16 ***
factor(alko)+	0.57625	0.07456	7.729	1.08e-14 ***
factor(suits)+:factor(alko)+	2.87373	0.16730	17.178	< 2e-16 ***

Alkoholi tarvitajaid on mittesuitsetajate seas
 $\exp(0.576) = 1,78$ korda rohkem kui
mittetarvitajaid

Alkoholi tarvitajaid on suitsetajate seas
 $\exp(0.57625+2.87373)=31,5$ korda rohkem kui
mittetarvitajaid

Alkoholi tarvitamise šansid on suitsetajatel
 $\exp(2,87373)=17,7$ korda suuremad kui
mittesuitsetajatel.

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m1=glm(n~factor(mari)+factor(suits)*factor(alko), family=poisson())
summary(m1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.09053	0.06228	81.732	< 2e-16 ***
factor(mari)+	-0.31542	0.04244	-7.431	1.08e-13 ***
factor(suits)+	-1.80971	0.15905	-11.378	< 2e-16 ***
factor(alko)+	0.57625	0.07456	7.729	1.08e-14 ***
factor(suits)+:factor(alko)+	2.87373	0.16730	17.178	< 2e-16 ***

```
> anova(m0, m1, test="Chisq")
```

Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	4	1286.02			
2	3	843.83	1	442.19	< 2.2e-16 ***

Koosmõjuga mudel on
parem mudel.

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m2=glm(n~factor(mari)*factor(suits)+factor(alko)*factor(mari)
+factor(suits)*factor(alko), family=poisson())
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.63342	0.05970	94.361	< 2e-16 ***
factor(mari)+	-5.30904	0.47520	-11.172	< 2e-16 ***
factor(suits)+	-1.88667	0.16270	-11.596	< 2e-16 ***
factor(alko)+	0.48772	0.07577	6.437	1.22e-10 ***
factor(mari)+:factor(suits)+	2.84789	0.16384	17.382	< 2e-16 ***
factor(mari)+:factor(alko)+	2.98601	0.46468	6.426	1.31e-10 ***
factor(suits)+:factor(alko)+	2.05453	0.17406	11.803	< 2e-16 ***

Kui vaatame marihukaanat mittetarbivaid ja mittesuitsetavaid tudengeid, siis on alkoholi tarvitajaid selliste tudengite seas $\exp(0.4877) = 1,63$ korda rohkem kui mittetarvitajaid (alkoholi tarvitamise šanss on 1,63).

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m2=glm(n~factor(mari)*factor(suits)+factor(alko)*factor(mari)
+factor(suits)*factor(alko), family=poisson())
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.63342	0.05970	94.361	< 2e-16 ***
factor(mari)+	-5.30904	0.47520	-11.172	< 2e-16 ***
factor(suits)+	-1.88667	0.16270	-11.596	< 2e-16 ***
factor(alko)+	0.48772	0.07577	6.437	1.22e-10 ***
factor(mari)+:factor(suits)+	2.84789	0.16384	17.382	< 2e-16 ***
factor(mari)+:factor(alko)+	2.98601	0.46468	6.426	1.31e-10 ***
factor(suits)+:factor(alko)+	2.05453	0.17406	11.803	< 2e-16 ***

Kui vaatame suitsetavaid kuid marihukaanat mittetarbivaid tudengeid, siis on alkoholi tarvitajaid selliste tudengite seas $\exp(0.4877+2.0545) = 12,7$ korda rohkem kui mittetarvitajaid.

OR (šansid tarbida alkohooli, suitsetajad vs mittesuitsetajad) on marihukaanat mittetarbivate tudengite jaoks on $\exp(2.054)=7,8$

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m2=glm(n~factor(mari)*factor(suits)+factor(alko)*factor(mari)
+factor(suits)*factor(alko), family=poisson())
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.63342	0.05970	94.361	< 2e-16 ***
factor(mari)+	-5.30904	0.47520	-11.172	< 2e-16 ***
factor(suits)+	-1.88667	0.16270	-11.596	< 2e-16 ***
factor(alko)+	0.48772	0.07577	6.437	1.22e-10 ***
factor(mari)+:factor(suits)+	2.84789	0.16384	17.382	< 2e-16 ***
factor(mari)+:factor(alko)+	2.98601	0.46468	6.426	1.31e-10 ***
factor(suits)+:factor(alko)+	2.05453	0.17406	11.803	< 2e-16 ***

Kui vaatame marihukaanat tarbivaid ja mittesuitsetavaid tudengeid, siis on alkoholi tarvitamise šanss $\exp(0.4877+2.986) = 32,3$

Kui vaatame marihukaanat tarbivaid ja suitsetavaid tudengeid, siis on alkoholi tarvitamise šanss $\exp(0.4877+2.986+2.054) = 251,5$

OR (šansid tarbida alkohooli, suitsetajad vs mittesuitsetajad) on marihukaanat tarbivate tudengite jaoks $\exp(2.054)=7,8$

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
> m3=glm(n~factor(mari)*factor(suits)*factor(alko), family=poisson())
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.63121	0.05987	94.060	< 2e-16 ***
factor(mari)+	-4.93806	0.70964	-6.959	3.44e-12 ***
factor(suits)+	-1.87001	0.16383	-11.414	< 2e-16 ***
factor(alko)+	0.49128	0.07601	6.464	1.02e-10 ***
factor(mari)+:factor(suits)+	2.27548	0.92746	2.453	0.014149 *
factor(mari)+:factor(alko)+	2.59976	0.72698	3.576	0.000349 ***
factor(suits)+:factor(alko)+	2.03538	0.17576	11.580	< 2e-16 ***
factor(mari)+:factor(suits)+:factor(alko)+	0.58951	0.94236	0.626	0.531600

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
> m3=glm(n~factor(mari)*factor(suits)*factor(alko), family=poisson())
Coefficients:
(Intercept) 5.63121 0.05987 94.060 < 2e-16 ***
factor(mari)+ -4.93806 0.70964 -6.959 3.44e-12 ***
factor(suits)+ -1.87001 0.16383 -11.414 < 2e-16 ***
factor(alko)+ 0.49128 0.07601 6.464 1.02e-10 ***
factor(mari)+:factor(suits)+ 2.27548 0.92746 2.453 0.014149 *
factor(mari)+:factor(alko)+ 2.59976 0.72698 3.576 0.000349 ***
factor(suits)+:factor(alko)+ 2.03538 0.17576 11.580 < 2e-16 ***
factor(mari)+:factor(suits)+:factor(alko)+ 0.58951 0.94236 0.626 0.531600

Marihuaanat mittetarbivate tudengite korral on
alkoholi tarbimise šansside suhe (suitsetajad vs
mittesuitsetajad) exp(2,035) = 7,65
```

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
> m3=glm(n~factor(mari)*factor(suits)*factor(alko), family=poisson())
Coefficients:
(Intercept) 5.63121 0.05987 94.060 < 2e-16 ***
factor(mari)+ -4.93806 0.70964 -6.959 3.44e-12 ***
factor(suits)+ -1.87001 0.16383 -11.414 < 2e-16 ***
factor(alko)+ 0.49128 0.07601 6.464 1.02e-10 ***
factor(mari)+:factor(suits)+ 2.27548 0.92746 2.453 0.014149 *
factor(mari)+:factor(alko)+ 2.59976 0.72698 3.576 0.000349 ***
factor(suits)+:factor(alko)+ 2.03538 0.17576 11.580 < 2e-16 ***
factor(mari)+:factor(suits)+:factor(alko)+ 0.58951 0.94236 0.626 0.531600

Marihuaanat mittetarbivate tudengite korral on
alkoholi tarbimise šansside suhe (suitsetajad vs
mittesuitsetajad) exp(2,035) = 7,65

Marihuaanat tarbivate tudengite korral on
alkoholi tarbimise šansside suhe (suitsetajad vs
mittesuitsetajad) exp(2,035+0.5895) = 13.8
```

Näide 2

Allan Agresti, An Introduction to Categorical Data Analysis (1996), lk 152

```
m0=glm(n~factor(alko)+factor(suits)+factor(mari), family=poisson())
m1=glm(n~factor(mari)+factor(suits)*factor(alko), family=poisson())
m2=glm(n~factor(mari)*factor(suits)+factor(alko)*factor(mari)
      +factor(suits)*factor(alko), family=poisson())
m3=glm(n~factor(mari)*factor(suits)*factor(alko), family=poisson())
```

udel	OR (alkohol, suits)
m0	1
m1	17,7
m2	7,8
m3, marihuaanata	7,6
m3, marihuaanaga	13,8

Parim mudel

```
> anova(m2, m3, test="Chisq")
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1      0.37399
2      0      0.00000 1 0.37399 0.5408
```