Zähldaten

Zielgrössen Y_i sind Anzahl Ereignisse in einem bestimmten Zeitraum.

Beispiele:

- Anzahl Schäden, Versicherungsfälle in einem Jahr
- Anzahl Diagnosen pro PatientIn
- Anzahl Pflanzen in einem Gebiet

$Y_i \sim \mathcal{P}(\lambda_i)$ und λ_i ist abhängig von erklärenden Variablen x_1, x_2, x_3

Poissonregression

Gegeben sind n unabhängige poissonverteilte Zielgrössen Y_i mit Erwartungswert $E(Y_i) = \lambda_i$

und λ_i hängt von erklärenden Variablen $x_1, x_2, \ldots$ in der folgenden Form ab:

$$g(\lambda_i) = \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Solche Modelle heissen **log-linear**.

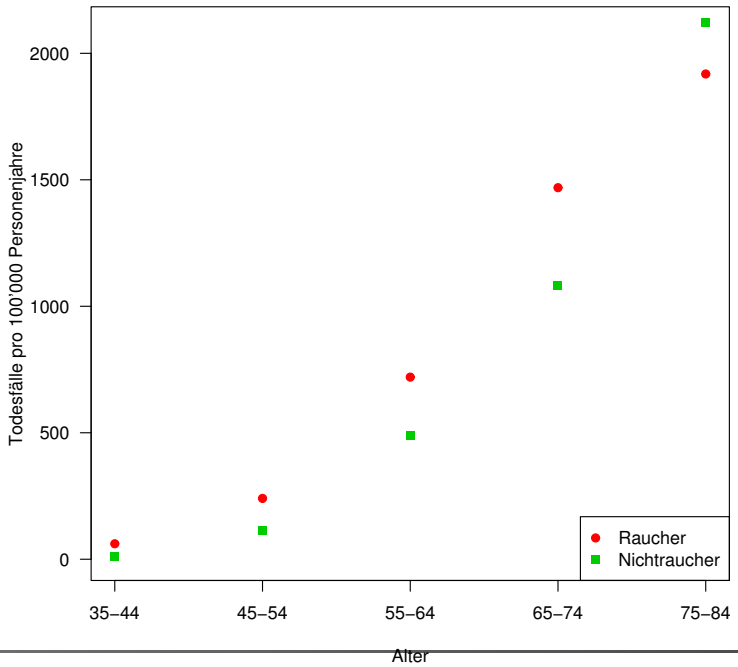
Todesfall durch Herzversagen

Daten:
Anzahl Todesfälle durch Herz-Kreislauferkrankung innerhalb von 10 Jahren (1951-1961) unter männlichen britischen Aerzten.

Ist Rauchen ein Risikofaktor?

Wenn ja, wie gross ist der Effekt und ist er abhängig vom Alter?

Todesfälle pro 100'000 Personenjahre



R Output

```
Call: glm(formula=deaths ~ offset(log(pers.years)) + smoker + age.n
          +age.nsq + smoker * age.n, family = poisson, data = doll
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -10.79176    0.45008 -23.978  < 2e-16 ***
smoker1         1.44097    0.37220   3.872 0.000108 ***
age.n           2.37648    0.20795  11.428  < 2e-16 ***
age.nsq        -0.19768    0.02737  -7.223 5.08e-13 ***
smoker1:age.n  -0.30755    0.09704  -3.169 0.001528 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703
```


Beobachtete und erwartete Anzahl Todesfälle

Alter	Raucher	Beob. Anz.	Erwartete Anz.
35–44	ja	32	29.6
45–54	ja	104	106.8
55–64	ja	206	208.2
65–74	ja	186	182.8
75–84	ja	102	102.6
35–44	nein	2	3.4
45–54	nein	12	11.5
55–64	nein	28	27.7
65–74	nein	28	30.2
75–84	nein	31	31.1

Generalized Linear Models

- Die Verteilung der Zielvariablen Y_i gehört einer einfachen Exponentialfamilie an.
- Die erklärenden Variablen gehen als Linearkombination $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$ ins Modell ein.
- Der Erwartungswert μ_i von Y_i ist durch die Linkfunktion g mit η verknüpft:
  $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots$

Einfache Exponentialfamilie

Die Wahrscheinlichkeitsfunktion $P(Y = y)$ oder die Dichte $f(y)$ von Y ist von der folgenden Form

$$exp\left\{\frac{1}{\phi}[\theta y + c(\theta)] + d(\phi, y)\right\}$$

θ ist der kanonische Parameter.
ϕ ist der Dispersionsparameter. Oft ist $\phi = \sigma^2$ oder 1.

Es gilt: $E(Y) = -c'(\theta)$ und $Var(Y) = -c''(\theta)\phi = V(\mu)\phi$.
$Var(\mu)$ heisst Varianzfunktion.

Beispiele Exponentialfamilie

Verteilung	$E(Y)$	$Var(Y)$	θ	ϕ
Normal	μ	σ^2	μ	σ^2
Binomial	np	$np(1-p)$	$\log(\frac{p}{1-p})$	1
Poisson	λ	λ	$\log\lambda$	1
Gamma	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\frac{\lambda}{\alpha}$	$\frac{1}{\alpha}$

Verteilung	$c(\theta)$	$d(\phi, y)$
Normal	$-\mu^2/2$	$-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$
Binomial	$-n\log(1+e^\theta)$	$\log\binom{n}{y}$
Poisson	$-e^\theta$	$-\log y!$
Gamma	$\log(\theta)$	$\frac{\log(y)-\log(\phi)}{\phi} - \log(y) - \log(\Gamma(\frac{1}{\phi})$

Goodness of Fit

Wie gut passt das Modell? Deviance = Lack of fit.

$\hat{L}_c =$ $L_c(\hat{\beta}) =$ maximale Likelihood des betrachteten Modells (c= current)

$\hat{L}_f =$ maximale Likelihood des vollen Modells (perfekter Fit: $y_i = \hat{\mu}_i$)

(skalierte) Devianz:

$$D^* = -2\log(\frac{\hat{L}_c}{\hat{L}_f}) = -2[\log\hat{L}_c - \log\hat{L}_f]$$

Devianz eines GLM

Log-Likelihood:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{1}{\phi}[\theta_i y_i + c(\theta_i)] + d(\phi, y_i) \right\}$$

(skalierte) Devianz:

$$D^* = \frac{2}{\phi} \sum_{i=1}^{n} \left\{ (\tilde{\theta}_i - \hat{\theta}_i) y_i + c(\tilde{\theta}_i) - c(\hat{\theta}_i) \right\}$$

wobei $\tilde{\theta}_i$ der MLE für θ im vollen Modell und $\hat{\theta}_i$ der MLE im betrachteten Modell ist.

Devianz (Fort.)

Bei p unbekannten β–Parametern ist $D^* \overset{as}{\sim} \chi^2_{n-p}$, d. h. wenn D^* ungefähr gleich $n - p$, dann ist das Modell gut.

Falls $D^* > \chi^2_{n-p,0.95}$ besteht ein signifikanter Lack of Fit.

Poissonverteilung:

$$D^* = 2 \sum_{i=1}^{n} y_i \log(\frac{y_i}{\hat{\mu}_i}) - 2 \sum_{i=1}^{n} (y_i - \hat{\mu}_i)$$

Vergleich von Modellen

Modell (1): q Parameter D_1^* mit $df = n - q$
Modell (2): p Parameter D_2^* mit $df = n - p$

$q < p$, Modell (1) ist im Model (2) enthalten.

Modellvergleich heisst $H_0 : \beta_{q+1} = \ldots = \beta_p = 0$

$$D_1^* - D_2^* = -2 \left[\log \hat{L}_{c_1} - \log \hat{L}_{c_2} \right] \sim \chi^2_{p-q}$$

Falls $D_1^* - D_2^* > \chi^2_{p-q,0.95}$ ist, genügt das kleinere Modell (1) nicht.

AIC Kriterium: $AIC = D^* + 2p$.

Residuen

Pearson Residuen:

$$\frac{y_i - \hat{y}_i}{\sqrt{\widehat{Var(Y_i)}}}$$

Devianz Residuen:

$$sign(y_i - \hat{y}_i)\sqrt{d_i}$$

wobe d_i die i-te Komponente der Devianz ist.

Anzahlen in Kontingenztafeln

- Beobachtung: Früherer Frass durch Blattläuse bewirkt chemische Veränderungen der Blätter, so dass diese weniger von Raupen angebohrt werden.
- Experiment: 2 Bäume, meherere Blätter hinsichtlich Blattlausfrass und Miniererbefall untersucht.
- Variablen: Anzahl Blätter, mit/ohne Blattläuse (Aphid), mit/ohne Löcher durch Raupen (Caterpillar), Baum 1 oder 2. Es gibt total 8=2*2*2 Beobachtungen.

Datensatz

```
> induced
   Tree   Aphid Caterpillar Count
1 Tree1  absent       holed    35
2 Tree1  absent         not  1750
3 Tree1 present       holed    23
4 Tree1 present         not  1146
5 Tree2  absent       holed   146
6 Tree2  absent         not  1642
7 Tree2 present       holed    30
8 Tree2 present         not   333
```

Volles Modell

```
Call: glm(formula = Count ~ Tree * Aphid * Caterpillar,
             family = poisson, data = induced)
Deviance Residuals:
[1]  0  0  0  0  0  0  0  0
Coefficients:              Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.555348   0.169031  21.034  < 2e-16 ***
Tree2                      1.428259   0.188204   7.589 3.23e-14 ***
Aphidp                    -0.419854   0.268421  -1.564  0.11778
Caterpillno                3.912023   0.170713  22.916  < 2e-16 ***
Tree2:Aphidp              -1.162555   0.335011  -3.470  0.00052 ***
Tree2:Caterpillno         -1.491959   0.191314  -7.798 6.27e-15 ***
Aphidp:Caterpillno        -0.003484   0.271097  -0.013  0.98975
Tree2:Aphidp:Caterpillno -0.009634   0.342474  -0.028  0.97756

    Null deviance:  6.5734e+03  on 7  degrees of freedom
Residual deviance: -4.2277e-13  on 0  degrees of freedom
AIC: 73.521
```


Ohne 3-Fach-Wechselwirkung

Modellvergleich:

```
> anova(id,id2,test="Chi")
Analysis of Deviance Table

Model 1: Count ~ Tree * Aphid * Caterpillar
Model 2: Count ~ Tree * Aphid * Caterpillar - Tree:Aphid:Caterpilla
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         0 -4.228e-13
2         1    0.00079 -1 -0.00079   0.97756
```

Ohne 2-Fach-Wechselwirkung A:C

Modellvergleich:

```
> anova(id2,id3,test="Chi")
Analysis of Deviance Table

Model 1: Count ~ Tree * Aphid * Caterpillar - Tree:Aphid:Caterpilla
Model 2: Count ~ Tree * Aphid * Caterpillar - Tree:Aphid:Caterpilla
    Aphid:Caterpillar
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         1    0.00079
2         2    0.00409 -1 -0.00329   0.95423
```


„Gutes“ Modell

```
> summary(id3)
Call:
glm(formula = Count ~ Tree * Aphid * Caterpillar - Tree:Aphid:Caterp
          Aphid:Caterpillar, family = poisson, data = induced)
Coefficients:      Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.55670    0.13215  26.915   <2e-16 ***
Tree2               1.42895    0.15244   9.374   <2e-16 ***
Aphidp             -0.42327    0.03763 -11.250   <2e-16 ***
Caterpillno         3.91064    0.13261  29.489   <2e-16 ***
Tree2:Aphidp       -1.17118    0.06877 -17.030   <2e-16 ***
Tree2:Caterpillno -1.49280    0.15419  -9.682   <2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6.5734e+03  on 7  degrees of freedom
Residual deviance: 4.0853e-03  on 2  degrees of freedom
AIC: 69.526
```