Generalisierte Lineare Modelle:

Eine Einführung mit 

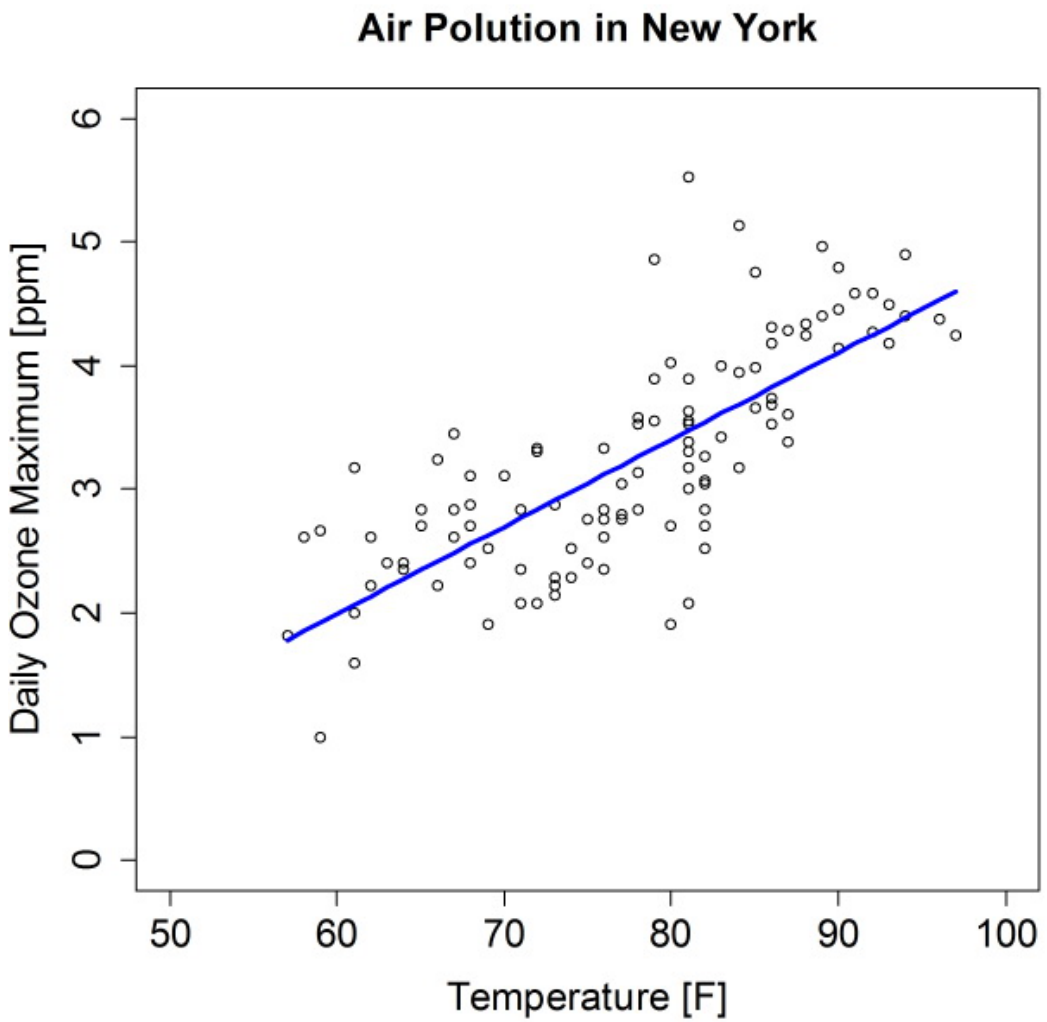
Herwig Friedl

Inhalt

- Lineares Modell (LM): Wiederholung
- Transformation auf Normalverteilung (Box-Cox)
- Lineare Exponentialfamilie (LEF)
- Generalisiertes Lineares Modell (GLM)
- Quasi-Likelihood Schätzung (Überdispersion)
- Logistische Regression (Binomiale Responses)
- Loglineare Modelle (Poisson Responses)
- GLMs mit zufälligen Effekten (EM Algorithmus)

Linear Regression: Modell

- Paare $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$; $\dim(\mathbf{x}_i) = p$; $\mathbf{x}_i$ fest
- Annahme: lineares Modell beschreibt Erwartungswert
- Modell: $\mathrm{E}(y_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, $\mathrm{var}(y_i) = \sigma^2$, $\mathrm{cov}(y_i, y_{i'}) = 0$



Linear Regression: Parameterschätzer

- LSE $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
- $y_i \overset{ind}{\sim} \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2) \Rightarrow \text{MLE } \hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$
- MLE $\hat{\sigma}^2 = \frac{1}{n} \text{SSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_i (y_i - \hat{\mu}_i)^2$, $\text{E}(\hat{\sigma}^2) = (1 - \frac{p}{n})\sigma^2$ (biased)
- $S^2 = \frac{1}{n-p} \text{SSE}(\hat{\boldsymbol{\beta}})$ (unbiased)
- $y_i \overset{ind}{\sim} \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2) \Rightarrow \text{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$
- ANOVA: $\text{SST} = \text{SSR}(\hat{\boldsymbol{\beta}}) + \text{SSE}(\hat{\boldsymbol{\beta}})$
- $\text{SST} = \sum_i (y_i - \bar{y})^2$, $\text{SSR}(\hat{\boldsymbol{\beta}}) = \sum_i (\hat{\mu}_i - \bar{y})^2$
- $R^2 = \dfrac{\text{SSR}(\hat{\boldsymbol{\beta}})}{\text{SST}} = 1 - \dfrac{\text{SSE}(\hat{\boldsymbol{\beta}})}{\text{SST}} \in (0, 1)$,

  $R^2_{adj} = 1 - \dfrac{\text{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\text{SST}/(n-1)} \notin (0, 1)$,

Linear Regression: Grenzen

Probleme:

- $y_i \not\sim \text{Normal}(\text{E}(y_i), \text{var}(y_i))$
- $\text{E}(y_i) \neq \mathbf{x}_i^\top \boldsymbol{\beta} \in \mathbb{R}$
- $\text{var}(y_i) \neq \sigma^2$ gleich (homoskedastisch) $\forall i = 1, \ldots, n$

Lösungsideen:

- transformiere y_i so dass $g(y_i) \overset{ind}{\sim} \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$
- verwende GLM mit $y_i \overset{ind}{\sim} \text{LEF}(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}), \phi V(\mu_i))$

Box-Cox Transformation

Definiere für **positive** Responses ($y > 0$)

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{falls } \lambda \neq 0, \\ \log y, & \text{falls } \lambda = 0, \end{cases}$$

Für $\lambda \to 0$ strebt $y(\lambda) \to \log y$, so dass $y(\lambda)$ eine stetige Funktion in λ ist.

Annahme: es existiert ein Wert λ für den

$$y_i(\lambda) \sim \text{Normal}(\mu_i(\lambda) = \mathbf{x}_i^\top \boldsymbol{\beta}(\lambda), \sigma^2(\lambda))$$

Berechne MLE bzgl. marginaler Verteilung der Responses y.

Box-Cox Transformation

Die Transformationsfamilie liefert dafür

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2(\lambda)}} \exp\left(-\dfrac{\left(\frac{(y^\lambda-1)}{\lambda}-\mu(\lambda)\right)^2}{2\sigma^2(\lambda)}\right) y^{\lambda-1}, & \lambda \neq 0, \\ \dfrac{1}{\sqrt{2\pi\sigma^2(\lambda)}} \exp\left(-\dfrac{(\log y-\mu(\lambda))^2}{2\sigma^2(\lambda)}\right) y^{-1}, & \lambda = 0. \end{cases}$$

- Falls $\lambda \neq 0$ und $\mu(\lambda) = \mathbf{x}^\top \boldsymbol{\beta}(\lambda)$, dann ist

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\lambda^2\sigma^2(\lambda)}} \exp\left(-\frac{\left(y^\lambda-1-\lambda\mathbf{x}^\top\boldsymbol{\beta}(\lambda)\right)^2}{2\lambda^2\sigma^2(\lambda)}\right) |\lambda| y^{\lambda-1}.$$

Box-Cox Transformation

Mit $\beta_0 = 1 + \lambda\beta_0(\lambda)$ und $\beta_j = \lambda\beta_j(\lambda)$, $j = 1, \ldots, p-1$, sowie $\sigma^2 = \lambda^2\sigma^2(\lambda)$ gilt

$$f(y|\lambda, \mu(\lambda), \sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\lambda^2\sigma^2(\lambda)}} \exp\left(-\frac{\left(y^\lambda - 1 - \lambda\mathbf{x}^\top\boldsymbol{\beta}(\lambda)\right)^2}{2\lambda^2\sigma^2(\lambda)}\right)|\lambda|y^{\lambda-1}$$

$$f(y|\lambda, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^\lambda - \mathbf{x}^\top\boldsymbol{\beta})^2}{2\sigma^2}\right)|\lambda|y^{\lambda-1}.$$

- Ist $\lambda = 0$, so sei $\beta_j = \beta_j(\lambda)$, $j = 0, \ldots, p-1$, und $\sigma^2 = \sigma^2(\lambda)$

$$f(y|0, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mathbf{x}^\top\boldsymbol{\beta})^2}{2\sigma^2}\right)y^{-1}.$$

Wäre λ bekannt, dann ist MLE sehr einfach zu berechnen!

Box-Cox Transformation

Relevanter Teil der Log-Likelihood Funktion der Stichprobe:

- $\lambda \neq 0$:

$$\ell(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i^{\lambda} - \mathbf{x}_i^{\top} \boldsymbol{\beta} \right)^2 + n \log |\lambda| + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

- $\lambda = 0$:

$$\ell(0, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\log y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta} \right)^2 - \sum_{i=1}^{n} \log y_i$$

Box-Cox Transformation: MLE's

Für festes λ lösen MLE's $\hat{\boldsymbol{\beta}}_\lambda$ und $\hat{\sigma}^2_\lambda$ (vgl. multiples LM) die Scoregleichung:

$$\frac{\partial \ell(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{y})}{\partial \boldsymbol{\beta}} = \begin{cases} \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{x}_i (y_i^\lambda - \mathbf{x}_i^\top \boldsymbol{\beta}) = \mathbf{0}, & \lambda \neq 0, \\ \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{x}_i (\log y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) = \mathbf{0}, & \lambda = 0\,, \end{cases}$$

$$\frac{\partial \ell(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{y})}{\partial \sigma^2} = \begin{cases} -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i^\lambda - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = 0, & \lambda \neq 0, \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (\log y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = 0\,, & \lambda = 0\,. \end{cases}$$

Box-Cox Transformation: MLE's

Wir erhalten

$$\hat{\boldsymbol{\beta}}_\lambda = \begin{cases} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}^\lambda, & \lambda \neq 0, \\ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \log \mathbf{y}, & \lambda = 0, \end{cases}$$

$$\hat{\sigma}^2_\lambda = \frac{1}{n} \mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda) = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} (y_i^\lambda - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda)^2, & \lambda \neq 0, \\ \frac{1}{n} \sum_{i=1}^{n} (\log y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda)^2, & \lambda = 0. \end{cases}$$

$\mathbf{y}^\lambda$ (bzw. $\log \mathbf{y}$) sind elementeweise gerechnet. $\mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)$ ist die Fehlerquadratsumme von y^λ (bzw. $\log y$) in $\hat{\boldsymbol{\beta}}_\lambda$ für λ fest.

Wegen dieser Parameterisierung ist $\mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)$ in $\lambda = 0$ unstetig.

Box-Cox Transformation: Profile-Likelihood Schätzer

Profile (Log-) Likelihoodfunktion $p\ell(\lambda|\mathbf{y}) = \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}^2_\lambda|\mathbf{y}) =$

$$= \begin{cases} -\frac{n}{2}\log \mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda) + n\log|\lambda| + (\lambda-1)\sum_{i=1}^n \log y_i, & \lambda \neq 0, \\ -\frac{n}{2}\log \mathrm{SSE}_0(\hat{\boldsymbol{\beta}}_0) - \sum_{i=1}^n \log y_i, & \lambda = 0. \end{cases}$$

Für $\lambda \neq 0$ folgt

$$p\ell(\lambda|\mathbf{y}) = -\frac{n}{2}\log \sum_{i=1}^n \frac{(y_i^\lambda - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda)^2}{\lambda^2} + (\lambda-1)\sum_{i=1}^n \log y_i$$

$$= -\frac{n}{2}\log \sum_{i=1}^n \left((y_i^\lambda - 1)/\lambda - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\lambda)\right)^2 + (\lambda-1)\sum_{i=1}^n \log y_i,$$

und somit $\lim_{\lambda\to 0} p\ell(\lambda|\mathbf{y}) = p\ell(0|\mathbf{y})$ (stetig).

Box-Cox Transformation: Profile-Likelihood Schätzer

Ignorieren wir auch $-\sum_{i=1}^{n} \log y_i$, so folgt

$$
\begin{aligned}
p\ell(\lambda|y) &= -\frac{n}{2}\log\sum_{i=1}^{n}\left((y_i^\lambda - 1)/\lambda - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}(\lambda)\right)^2 + \lambda\sum_{i=1}^{n}\log y_i \\
&= -\frac{n}{2}\log\sum_{i=1}^{n} r_i^2 + \log\left(\exp\left[n\lambda\frac{1}{n}\sum_{i=1}^{n}\log y_i\right]\right) \\
&= -\frac{n}{2}\log\sum_{i=1}^{n} r_i^2 + \log\left(c^{n\lambda}\right) = -\frac{n}{2}\log\sum_{i=1}^{n} r_i^2 + \frac{n}{2}\log(c^\lambda)^2 \\
&= -\frac{n}{2}\log\sum_{i=1}^{n}\left(\frac{r_i}{c^\lambda}\right)^2
\end{aligned}
$$

mit Residuen r_i zum Modell $\mathbf{x}_i^\top\boldsymbol{\beta}(\lambda)$ für Responses $(y_i^\lambda - 1)/\lambda$, und mit der Konstanten $c = \exp(\frac{1}{n}\sum_i \log y_i)$.

Box-Cox Transformation: Profile-Likelihood Schätzer

LRT: $H_0 : \lambda = \lambda_0$ gegen $H_1 : \lambda \neq \lambda_0$. LRT Statistik definiert als

$$\Lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} L(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{y})}{\sup_{\theta \in \Theta} L(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{y})},$$

Unter gewissen Regularitätsbedingung gilt

$$\begin{aligned} -2 \log(\Lambda(\mathbf{y})) &= -2 \left(\ell(\lambda_0, \hat{\boldsymbol{\beta}}_{\lambda_0}, \hat{\sigma}^2_{\lambda_0} | \mathbf{y}) - \ell(\hat{\lambda}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | \mathbf{y}) \right) \\ &= -2 \left(p\ell(\lambda_0 | \mathbf{y}) - p\ell(\hat{\lambda} | \mathbf{y}) \right) \xrightarrow{D} \chi^2_1. \end{aligned}$$

Wegen $-2(p\ell(\lambda_0|\mathbf{y}) - p\ell(\hat{\lambda}|\mathbf{y})) \sim \chi^2_1$ beinhaltet ein $(1-\alpha)$ Konfidenzintervall alle Werte λ_0, für die $p\ell(\lambda_0|\mathbf{y})$ maximal $\frac{1}{2}\chi^2_{1;1-\alpha}$ von $p\ell(\hat{\lambda}|\mathbf{y})$ entfernt ist ($\chi^2_{1;0.95} = 3.841$, $\chi^2_{1;0.99} = 6.635$).

Box-Cox Transformation: Eigenschaften

Log-Transformation ($\lambda = 0$): für $\log y_i \sim \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ gilt

$$\begin{aligned} \text{median}(\log y_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{E}(\log y_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{var}(\log y_i) &= \sigma^2. \end{aligned}$$

Die originalen y_i sind lognormalverteilt mit

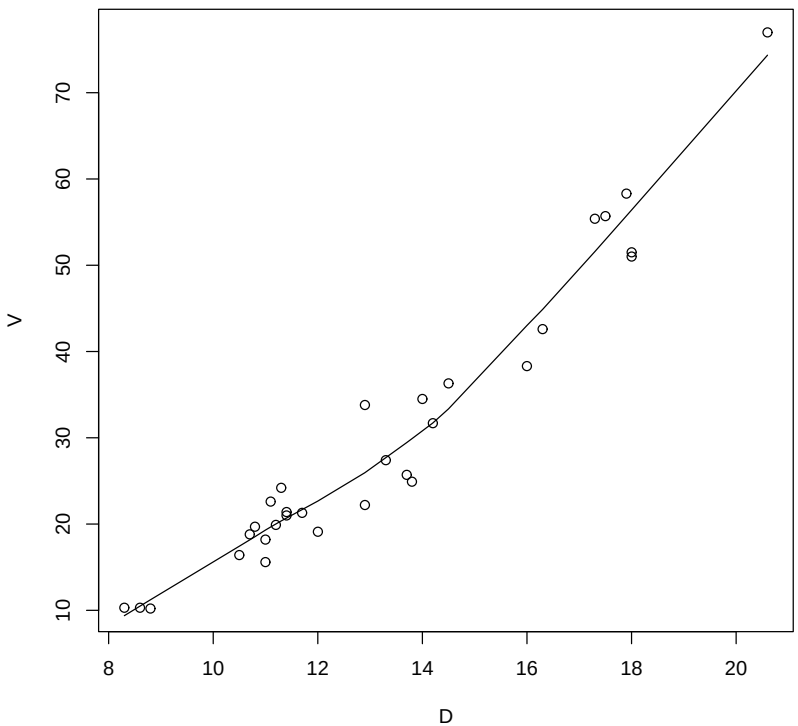
$$\begin{aligned} \text{median}(y_i) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \\ \text{E}(y_i) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \sigma^2/2) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\sigma^2/2), \\ \text{var}(y_i) &= \big(\exp(\sigma^2) - 1\big) \exp(2\mathbf{x}_i^\top \boldsymbol{\beta} + \sigma^2). \end{aligned}$$

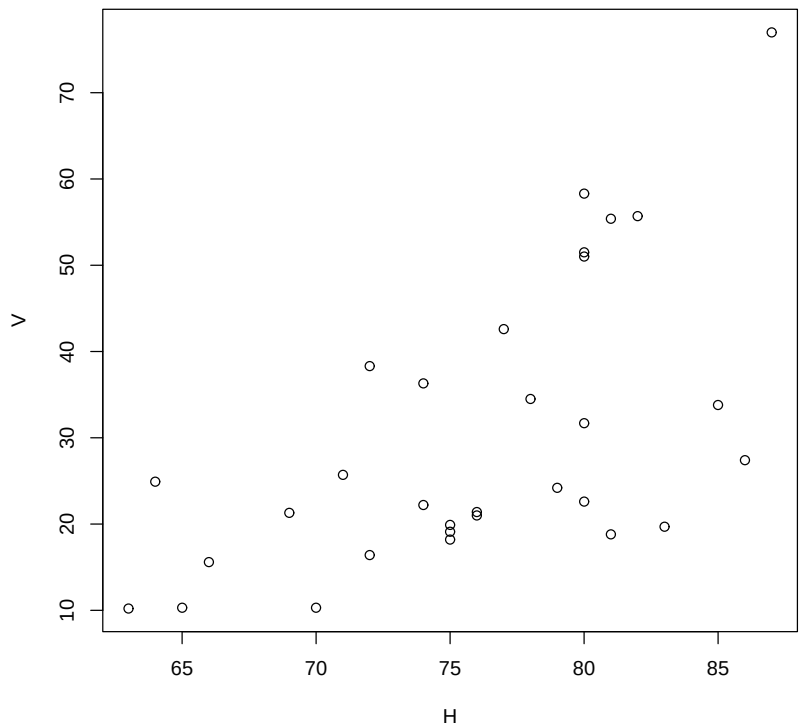
Additives Modell für Erwartungswert (und für Median) von $\log y_i$ ist multiplikativ für Median und Erwartungswert von y_i. $\text{E}(y_i)$ ist das $1 < \exp(\sigma^2/2)$-fache des Medians und die Varianz ist nicht mehr konstant für $i = 1, \ldots, n$.

Box-Cox Transformation: Beispiel

$n = 31$ Black Cherry Bäume. Zusammenhang zwischen Holzvolumen V in feet^3, Baumhöhe H in feet und Durchmesser D in inches (1 inch = 2.54 cm, 12 inches = 1 foot).

```
> trees <- read.table("trees.dat", header=TRUE); attach(trees)
> plot(D, V); lines(lowess(D, V)) # curvature (wrong scale?)
> plot(H, V)  # increasing variance?
```

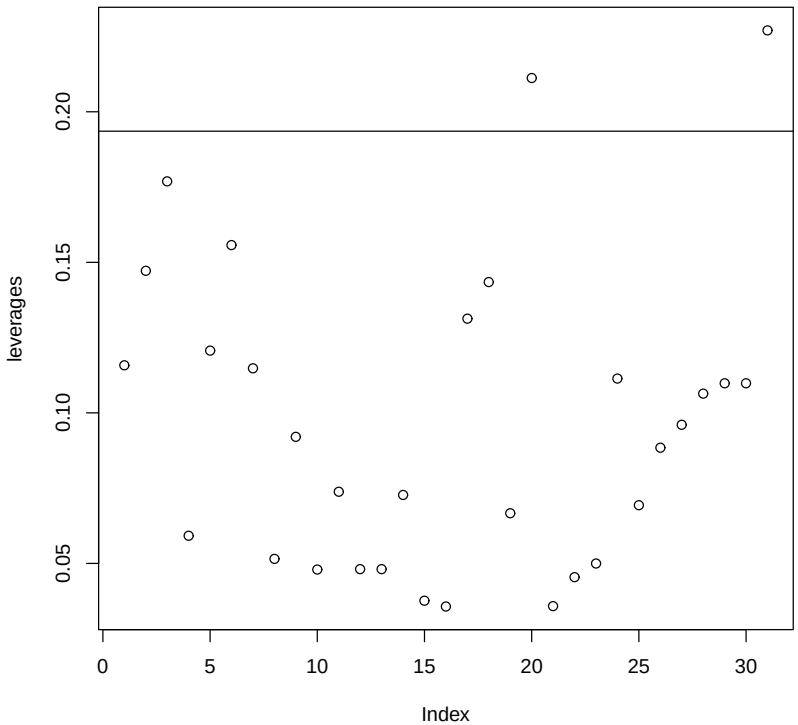


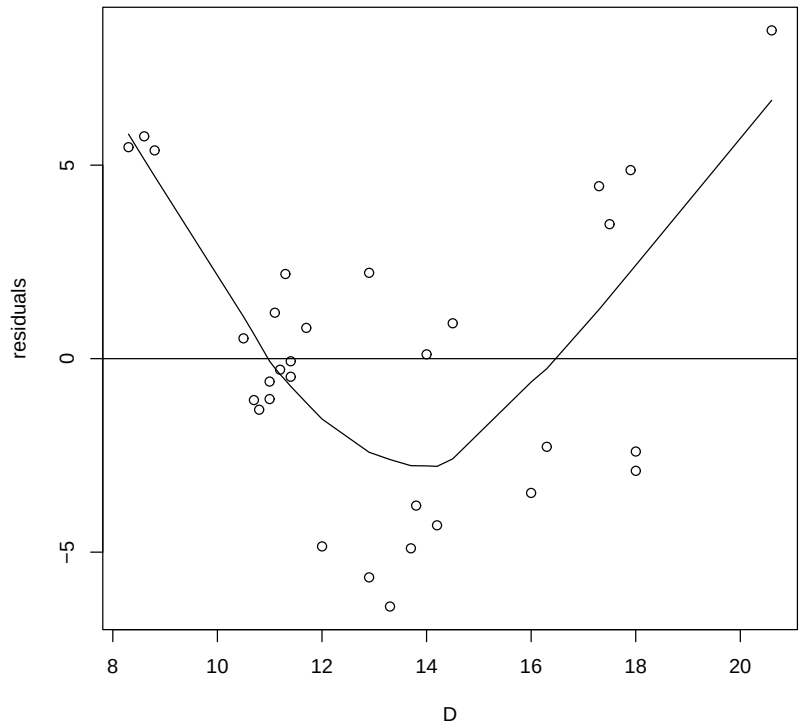


Box-Cox Transformation: Beispiel

```
> (mod <- lm(V ~ H + D)) # still fit a linear model for volume
Coefficients:
(Intercept)            H            D
   -57.9877       0.3393       4.7082

> plot(lm.influence(mod)$hat, ylab = "leverages")
> h.crit <- 2*mod$rank/length(V)
> abline(h.crit, 0) # 2 leverage points

> plot(D, residuals(mod), ylab="residuals"); abline(0, 0)
> lines(lowess(D, residuals(mod))) # sink in the middle
```

Box-Cox Transformation: Beispiel

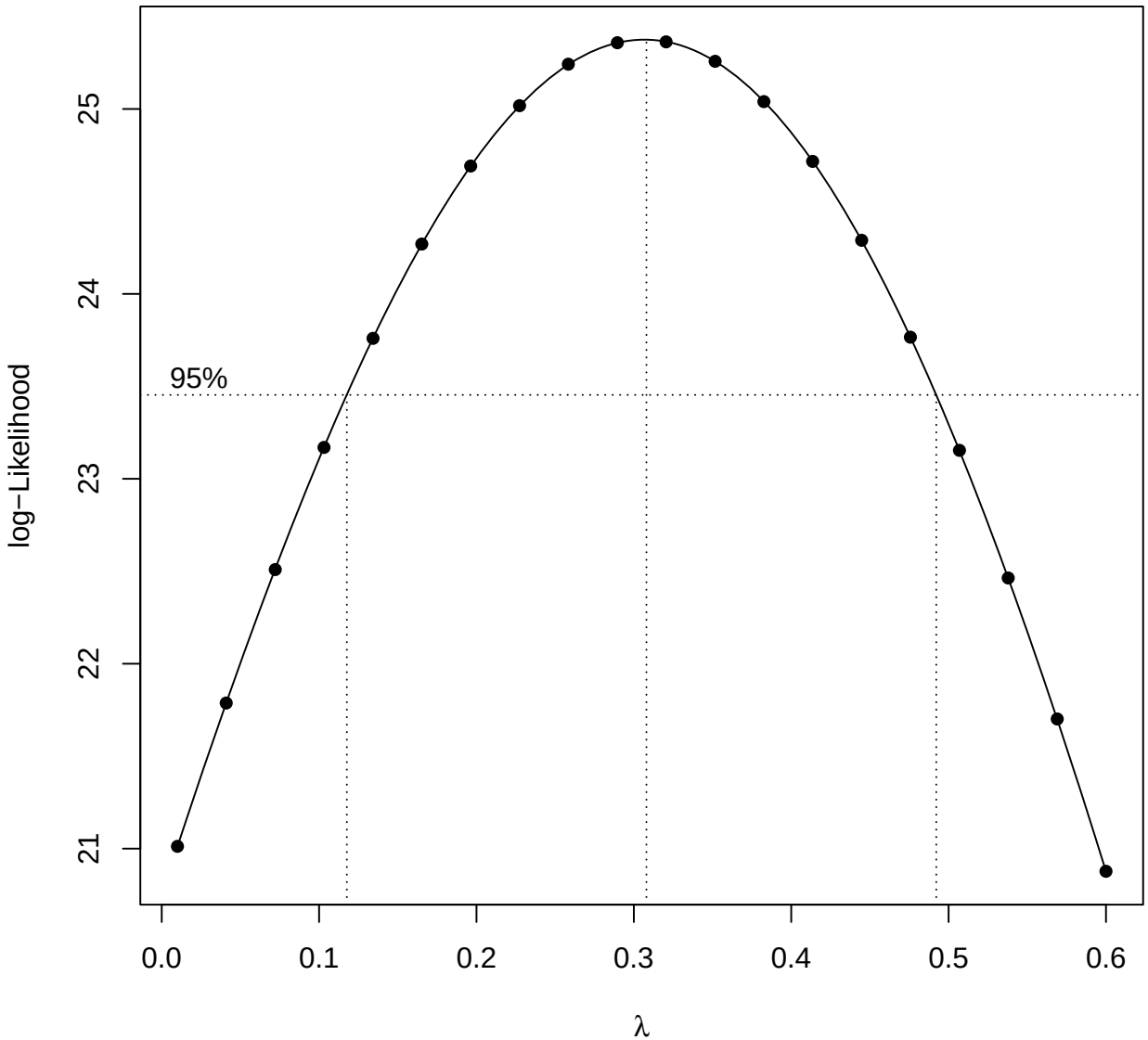




```
> library(MASS)

> bc<-boxcox(V~H+D,lambda=seq(0.0,0.6,length=100),plotit=FALSE))
> ml.index <- which(bc$y == max(bc$y))
> bc$x[ml.index]
[1] 0.3090909

> boxcox(V~H+D, lambda = seq(0.0, 0.6,len = 18)) # plot it now
```

```
> # directly calculate pl(lambda|y) - doesn't work if lambda=0!
> require(MASS)
> bc.trafo <- function(y, lambda) (y^lambda - 1)/lambda
> n <- length(V)
> lambda <- seq(0.01, 0.4, len=20) # avoide lambda=0
> res <- matrix(0, nrow = length(lambda), 2)
> C <- exp(mean(log(V))) # scaling constant
> for(i in seq_along(lambda)) {
+  r <- resid(lm(bc.trafo(V, lambda[i]) ~ H + D))
+  pl <- -(n/2) * log(sum((r/(C^lambda[i]))^2))
+  res[i, ] <- c(lambda[i], pl)
> }

> boxcox(V~H+D, lambda = lambda)  # compare with box cox
> points(res[,1], res[,2], pch=16) # add points to verify match
```

Box-Cox Transformation: Beispiel



Box-Cox Transformation: Beispiel

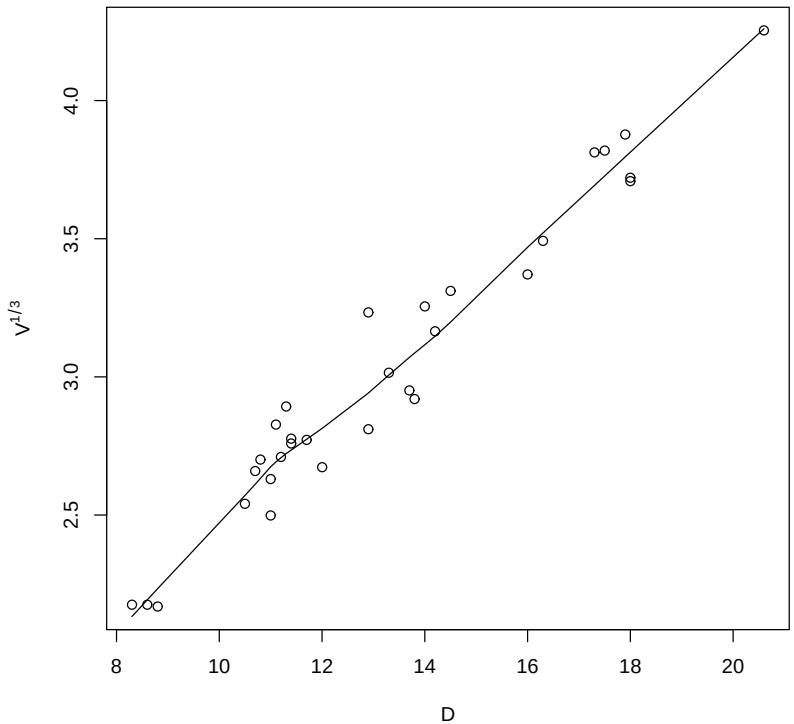
Volumenmessung verhält sich kubisch in Höhe und Durchmesser?

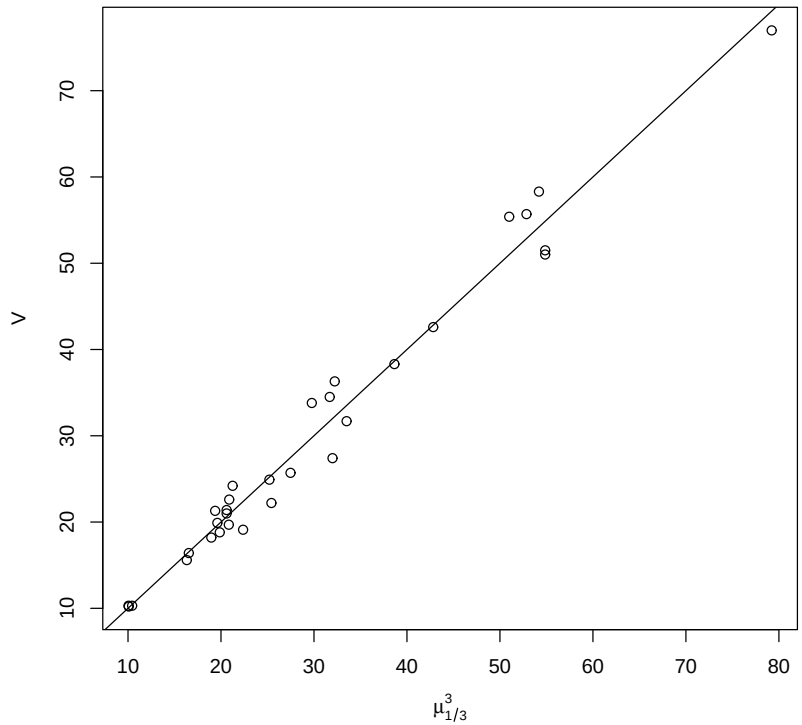
```
> plot(D, V^(1/3), ylab=expression(V^{1/3}))
> lines(lowess(D, V^(1/3))) # curvature almost removed

> (mod1 <- lm(V^(1/3) ~ H + D))
Coefficients:
(Intercept)            H            D
   -0.08539      0.01447      0.15152
```

Für $\lambda = 1/3$ fest ist $\widehat{\text{median}}(V) = \hat{\mu}_{1/3}^3$ mit $\text{E}(V^{1/3}) = \mu_{1/3}$.
$\hat{\text{E}}(V) = \hat{\mu}_{1/3}^3(1 + 3\hat{\sigma}_{1/3}^2/\hat{\mu}_{1/3}^2)$. Vergleiche Responses mit Medianen.

```
> mu <- fitted(mod1)
> plot(mu^3, V) # fitted median modell
```

Box-Cox Transformation: Beispiel



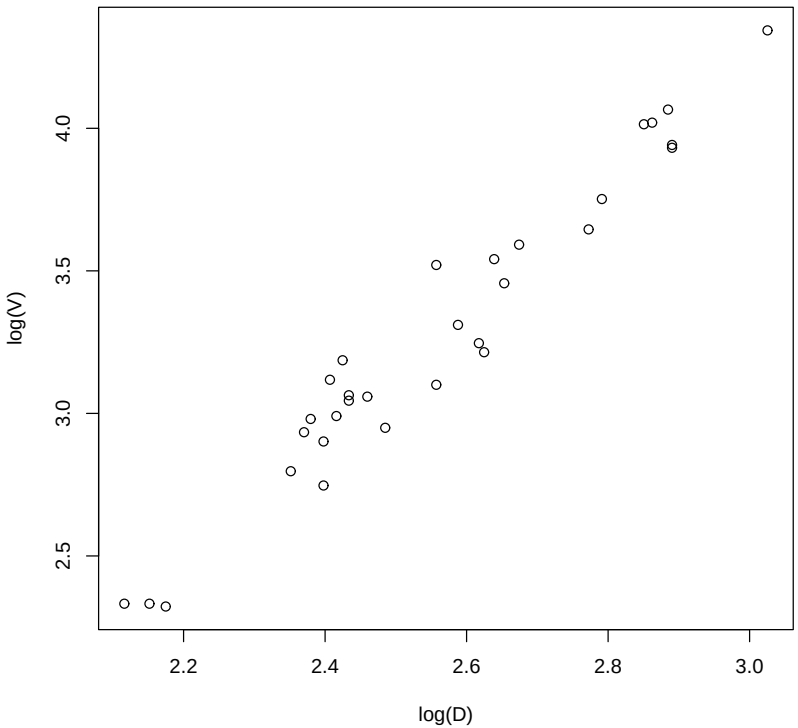


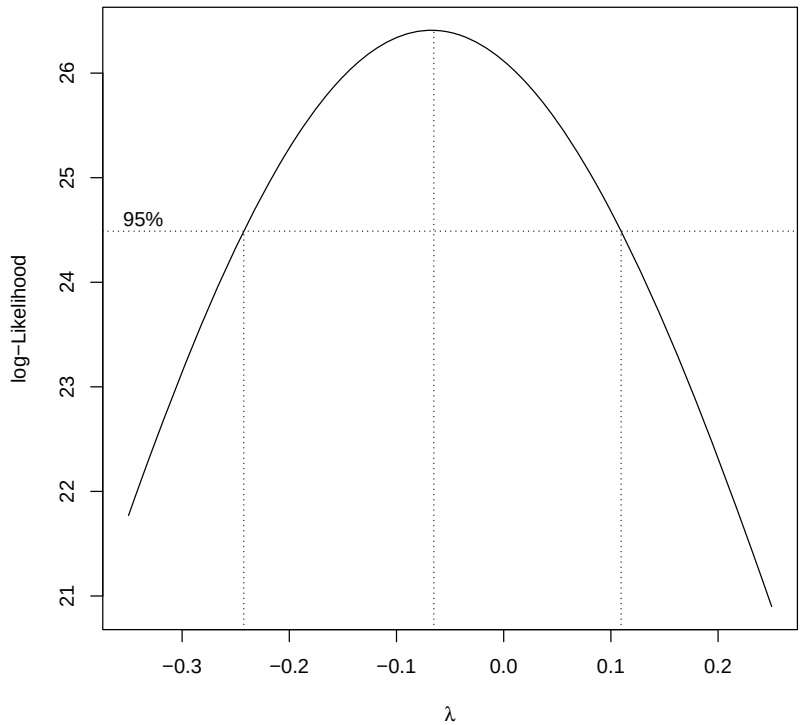
Box-Cox Transformation: Beispiel

Krümmung kann vielleicht durch log-Transformation aller Variablen reduziert werden (Regression auf $\log(D)$ und $\log(H)$). Soll man jetzt jedoch auf der $\log(V)$ Achse modellieren?

```
> plot(log(D), log(V)) # shows nice linear relationship

> lm(log(V) ~ log(H) + log(D)) # response log(V) or still V?
Coefficients:
(Intercept)       log(H)       log(D)
     -6.632        1.117        1.983

> boxcox(V~log(H)+log(D), lambda=seq(-0.35,0.25,length=100))
```

Box-Cox Transformation: Beispiel





Box-Cox Transformation: Beispiel

Welches der beiden Modelle ist *besser*? Vergleich mittels LRT.
Bette beide Modelle ein in die **Modellfamilie**

$$V^* \sim \text{Normal}(\beta_0 + \beta_1 H^* + \beta_2 D^*, \sigma^2)$$

$$V^* = (V^{\lambda_V} - 1)/\lambda_V$$
$$H^* = (H^{\lambda_H} - 1)/\lambda_H$$
$$D^* = (D^{\lambda_D} - 1)/\lambda_D$$

Vergleiche Profile-Likelihoodfunktion in $\lambda_V = 1/3$, $\lambda_H = \lambda_D = 1$ ($\text{E}(V^{1/3}) = \beta_0 + \beta_1 H + \beta_2 D$), mit jenem in $\lambda_V = \lambda_H = \lambda_D = 0$ ($\text{E}(\log(V)) = \beta_0 + \beta_1 \log(H) + \beta_2 \log(D)$).

Box-Cox Transformation: Beispiel

```
> bc1 <- boxcox(V ~ H + D, lambda = 1/3, plotit=FALSE)
> bc1$y
[1] 25.33313

> bc2 <- boxcox(V ~ log(H) + log(D), lambda = 0, plotit=FALSE)
> bc2$y
[1] 26.11592
```

LRT Statistik: $-2(25.333 - 26.116) = 1.566$ (nicht signifikant).

Box-Cox Transformation: Beispiel

Bemerkung: Schätzer zu $\log(H)$ nahe bei Eins ($\hat{\beta}_1 = 1.117$) und Schätzer zu $\log(D)$ nahe bei Zwei ($\hat{\beta}_2 = 1.983$).

Baum durch **Zylinder** oder **Kegel** beschreibbar. Volumen $\pi h d^2/4$ (Zylinder) oder $\pi h d^2/12$ (Kegel), also

$$\log(V) = c + 1\log(H) + 2\log(D)$$

mit $c = \log(\pi/4)$ (Zylinder) oder $c = \log(\pi/12)$ (Kegel).

Vorsicht: D von inches auf feet konvertieren $\Rightarrow D/12$ als Prädiktor.

Box-Cox Transformation: Beispiel

```
> lm(log(V) ~ log(H) + log(D/12))
Coefficients:
(Intercept)       log(H)    log(D/12)
     -1.705        1.117        1.983
```

Konvertierung beeinflusst nur Interceptwert!

Fixiere Slopes (β_1, β_2) auf (1, 2) und schätze nur Intercept β_0, d.h. betrachte das Modell

$$\mathrm{E}(\log(V)) = \beta_0 + 1\log(H) + 2\log(D/12)\,.$$

Den Term $1\log H + 2\log(D/12)$ nennt man **offset** (Prädiktor mit festem Parameter).

Box-Cox Transformation: Beispiel

```
> (mod3 <- lm(log(V) ~ 1 + offset(log(H) + 2*log(D/12))))
Coefficients:
(Intercept)
     -1.199

> log(pi/4)
 [1] -0.2415645
> log(pi/12)
 [1] -1.340177
```

Holzvolumen kann eher durch ein Kegelvolumen als durch das eines Zylinders beschrieben werden, hat jedoch ein etwas größeres Volumen als ein Kegel.

Lineare Exponentialfamilie (LEF): Definition

Definition: Eine Zufallsvariable y sei aus einer Verteilung mit Dichte- oder Wahrscheinlichkeitsfunktion

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

für bekannte Funktionen $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ mit $a(\phi) > 0$.

Ist ϕ fest, nennt man $f(y|\theta)$ **einparametrige, lineare Exponentialfamilie in kanonischer Form** (LEF) mit kanonischem Parameter θ.

Lineare Exponentialfamilie (LEF)

Bemerkung: Allgemeine Exponentialfamilie (AEF) definiert durch

$$f(y|\theta) = h(y)p(\theta)\exp\left\{\sum_{j=1}^{k} t_j(y)w_j(\theta)\right\},$$

mit reellen Funktionen $h(y) \geq 0$ und $t_1(y), \ldots, t_k(y)$ in y sowie $p(\theta) \geq 0$ und $w_1(\theta), \ldots, w_k(\theta)$ in θ.
Um darin LEF zu erkennen, schreiben wir AEF um zu

$$f(y|\theta) = \exp\left\{\sum_{j=1}^{k} t_j(y)w_j(\theta) + \log(p(\theta)) + \log(h(y))\right\}.$$

Setze $k = 1$ (einparametrig), $t(y) = y$ (linear), $w(\theta) = \theta$ (kanonische Parametrisierung), sowie $\log(p(\theta)) = -b(\theta)$ und $\log(h(y)) = c(y, \phi)$. Bis auf Skalierung ist dies eine LEF.

Lineare Exponentialfamilie: Momente

Bemerkung: Für die AEF ($\log(p(\theta)) = -b(\theta)$) gilt

$$\mathrm{E}\left(\sum_{j=1}^{k} \frac{\partial w_j(\theta)}{\partial \theta_l} t_j(y)\right) = -\frac{\partial}{\partial \theta_l} \log(p(\theta)),$$

$$\mathrm{var}\left(\sum_{j=1}^{k} \frac{\partial w_j(\theta)}{\partial \theta_l} t_j(y)\right) = -\frac{\partial^2}{\partial \theta_l^2} \log(p(\theta)) - \mathrm{E}\left(\sum_{j=1}^{k} \frac{\partial^2 w_j(\theta)}{\partial \theta_l^2} t_j(y)\right).$$

Ist $k = 1$, $t(y) = y$ und $w(\theta) = \theta$, liefert dies $\mathrm{E}(y) = b'(\theta)$ und $\mathrm{var}(y) = b''(\theta)$.

Lineare Exponentialfamilie: Score & Information

Bemerkung: Für Scorefunktion und Informationszahl gilt:

$$\mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = 0,$$

$$\mathrm{var}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = \mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2 = \mathrm{E}\left(-\frac{\partial^2 \log f(y|\theta)}{\partial \theta \partial \theta^\top}\right).$$

Hier:

$$\mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = \frac{1}{a(\phi)}\mathrm{E}(y - b'(\theta)) = 0,$$

also $\mathrm{E}(y) = b'(\theta)$, und

$$\mathrm{E}\left(\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2 = -\frac{1}{a(\phi)}b''(\theta) + \frac{1}{a^2(\phi)}\mathrm{var}(y) = 0$$

Also $\mathrm{var}(y) = a(\phi)b''(\theta)$.

Lineare Exponentialfamilie: Kumulanten

Sei $\mathrm{E}(y) = b'(\theta) = \mu$ und somit $\mathrm{var}(y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$.

Varianz von y ist Produkt zweier Funktionen: $V(\mu)$ und $a(\phi)$.
Varianzfunktion $V(\mu)$, **Dispersionsparameter** ϕ.

$b(\theta)$ nennt man **Kumulantenfunktion**.
Kumulantenerzeugenden Funktion $K(t) = \log M(t)$.
k-te Kumulante κ_k gegeben durch $K^{(k)}(t)|_{t=0}$.

Zusammenhang mit Momenten

$$\begin{aligned}
\kappa_1(y) &= \mathrm{E}(y) \\
\kappa_2(y) &= \mathrm{E}(y-\mu)^2 \\
\kappa_3(y) &= \mathrm{E}(y-\mu)^3 \\
\kappa_4(y) &= \mathrm{E}(y-\mu)^4 - 3\mathrm{var}^2(y)\,.
\end{aligned}$$

Lineare Exponentialfamilie: Kumulantenerzeugende

Für die LEF gilt

$$\begin{aligned} 1 &= \int_{\mathbb{R}} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy \\ &= \exp\left(-\frac{b(\theta)}{a(\phi)}\right) \int_{\mathbb{R}} \exp\left(\frac{y}{a(\phi)}\,\theta + c(y,\phi)\right) dy\,, \end{aligned}$$

also

$$\exp\left(\frac{b(\theta)}{a(\phi)}\right) = \int_{\mathbb{R}} \exp\left(\frac{y}{a(\phi)}\,\theta + c(y,\phi)\right) dy\,.$$

Lineare Exponentialfamilie: Kumulantenerzeugende

Wegen

$$\exp\left(\frac{b(\theta)}{a(\phi)}\right) = \int_{\mathbb{R}} \exp\left(\frac{y}{a(\phi)}\,\theta + c(y,\phi)\right) dy$$

folgt als Momentenerzeugende Funktion $M(t)$

$$\begin{aligned}
\mathrm{E}(e^{ty}) &= \exp\left(-\frac{b(\theta)}{a(\phi)}\right) \int_{\mathbb{R}} \exp\left(\frac{y}{a(\phi)}\Big(\theta + a(\phi)t\Big) + c(y,\phi)\right) dy \\
&= \exp\left(-\frac{b(\theta)}{a(\phi)}\right) \exp\left(\frac{b\Big(\theta + a(\phi)t\Big)}{a(\phi)}\right) \\
&= \exp\left(\frac{b\Big(\theta + a(\phi)t\Big) - b(\theta)}{a(\phi)}\right) .
\end{aligned}$$

Lineare Exponentialfamilie: Kumulantenerzeugende

Wegen

$$\mathrm{E}(e^{ty}) = \exp\left(\frac{b\Big(\theta + a(\phi)t\Big) - b(\theta)}{a(\phi)}\right)$$

folgt als Kumulantenerzeugende Funktion

$$K(t) = \log M(t) = \frac{b\Big(\theta + a(\phi)t\Big) - b(\theta)}{a(\phi)}\,.$$

Die k-te Kumulante von y ist somit

$$\kappa_k(y) = K^{(k)}(t)|_{t=0} = a(\phi)^{k-1} b^{(k)}\Big(\theta + a(\phi)t\Big)\Big|_{t=0} = a(\phi)^{k-1} b^{(k)}(\theta)\,.$$

Lineare Exponentialfamilie: MLE für Erwartung

Annahme: $y_1, \ldots, y_n$ (iid) Zufallsstichprobe aus LEF(θ).
MLE für μ ist Nullstelle der **Scorefunktion**

$$\sum_{i=1}^{n} \frac{\partial \log f(y_i|\theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{\partial \log f(y_i|\theta)}{\partial \theta} \frac{\partial \theta}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - b'(\theta)}{a(\phi)} \frac{\partial \theta}{\partial \mu} .$$

Mit $b'(\theta) = \mu$ und wegen (Ableitung der inversen Funktion)

$$\frac{\partial \mu}{\partial \theta} = \frac{\partial b'(\theta)}{\partial \theta} = b''(\theta) = V(\mu)$$

ist die Scorefunktion

$$\sum_{i=1}^{n} \frac{\partial \log f(y_i|\theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - \mu}{a(\phi) V(\mu)} = \sum_{i=1}^{n} \frac{y_i - \mu}{\text{var}(y_i)} .$$

Lineare Exponentialfamilie: MLE für Erwartung

MLE $\hat{\mu}$ löst

$$\sum_{i=1}^{n} \frac{\partial \log f(y_i|\theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - \mu}{a(\phi)V(\mu)} = 0\,.$$

Entspricht der Ableitung der Fehlerquadratsumme beim LM mit $\text{var}(y_i) = \sigma^2$ (Normalverteilung).

Bekannte Lösung für gesamte LEF:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

Lineare Exponentialfamilie: Dispersion

Von nun an Annahme: beobachtungsspezifische Funktionen $a_i(\cdot)$ hängen **nur von einem** globalen Dispersionsparameter ϕ ab! (sonst ist die Anzahl der Dispersionsparameter gleich n).

Beispiel: N Mittel $\bar{y}_k$ von Stichproben mit Umfängen $n_1, \ldots, n_N$. Nur diese Mittel $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ki}$ sind beobachtet.
Falls y_{ki} (iid) Zufallsstichprobe mit $\mathrm{E}(y_{ki}) = \mu$ und $\mathrm{var}(y_{ki}) = \sigma^2$, dann $\mathrm{E}(\bar{y}_k) = \mu$ und $\mathrm{var}(\bar{y}_k) = \sigma^2/n_k = a_k \cdot \phi$ mit $a_k = 1/n_k$ bekannt und unbekannter Dispersion $\phi = \sigma^2$.

Daher werden wir uns im Folgenden ausschließlich auf den Fall $a_i(\phi) = a_i \cdot \phi$ mit bekannten Gewichten a_i beschränken. Unter diesem Modell hängt der MLE $\hat{\mu}$ nicht mehr von ϕ ab.

Lineare Exponentialfamilie: Mitglieder

- **Normalverteilung** $y \sim \text{Normal}(\mu, \sigma^2)$:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right), \qquad y \in \mathbb{R}.$$

Setze $\theta = \mu$ und $\phi = \sigma^2$, so führt dies zur LEF mit

$$a = 1\,, \qquad b(\theta) = \theta^2/2\,, \qquad c(y,\phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)\,,$$

wofür gilt

$$\mathrm{E}(y) = b'(\theta) = \theta = \mu$$
$$\mathrm{var}(y) = \phi b''(\theta) = \phi \cdot 1 = \sigma^2$$
$$\kappa_k(y) = 0 \qquad \text{für } k > 2\,.$$

Lineare Exponentialfamilie: Mitglieder

- **Poissonverteilung** $y \sim \text{Poisson}(\mu)$:

$$f(y|\mu) = \frac{\mu^y}{y!} e^{-\mu} = \exp\left(y \log \mu - \mu - \log y!\right) , \qquad y = 0, 1, 2, \ldots .$$

Setze $\theta = \log \mu$ und $\phi = 1$, so führt dies zur LEF mit

$$a = 1 , \qquad b(\theta) = \exp(\theta) , \qquad c(y, \phi) = -\log y! ,$$

wofür gilt

$$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = \exp(\theta) = \mu \\ \mathrm{var}(y) &= b''(\theta) = \exp(\theta) = \mu \\ \kappa_k(y) &= \exp(\theta) = \mu \qquad \text{für } k > 2 . \end{aligned}$$

Die Dispersion ist bei der Poissonverteilung bekannt Eins und somit *wirklich* kein freier Parameter.

Lineare Exponentialfamilie: Mitglieder

- **Gammaverteilung** $y \sim \text{Gamma}(a, \lambda)$:

$$f(y|a,\lambda) = \exp(-\lambda y)\lambda^a y^{a-1}\frac{1}{\Gamma(a)}\,, \qquad a, \lambda, y > 0\,.$$

Dafür ist $\text{E}(y) = a/\lambda$ und $\text{var}(y) = a/\lambda^2$.
Reparametrisierung $\mu = \nu/\lambda$, $\nu = a$ gibt $\text{E}(y) = \mu$, $\text{var}(y) = \mu^2/\nu$ und Dichtefunktion für $\mu, \nu, y > 0$

$$\begin{aligned} f(y|\mu,\nu) &= \exp\left(-\frac{\nu}{\mu}y\right)\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\frac{1}{\Gamma(\nu)} \\ &= \exp\left(-\frac{\nu}{\mu}y + \nu\log\nu - \nu\log\mu + (\nu-1)\log y - \log\Gamma(\nu)\right) \\ &= \exp\left(\frac{y\left(-\frac{1}{\mu}\right) + \log\frac{1}{\mu}}{1/\nu} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right)\,. \end{aligned}$$

Lineare Exponentialfamilie: Mitglieder

Dichtefunktion für $\mu, \nu, y > 0$

$$f(y|\mu,\nu) = \exp\left(\frac{y\left(-\frac{1}{\mu}\right) + \log\frac{1}{\mu}}{1/\nu} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right).$$

Setze $\theta = -1/\mu$ und $\phi = 1/\nu$, so führt dies zur LEF mit

$$a = 1\,,\ b(\theta) = -\log(-\theta)\,,\ c(y,\phi) = \frac{1}{\phi}\log\frac{1}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma\left(\frac{1}{\phi}\right),$$

wofür gilt

$$\mathrm{E}(y) = b'(\theta) = -\frac{1}{\theta} = \mu$$

$$\mathrm{var}(y) = \phi b''(\theta) = \phi\frac{1}{\theta^2} = \frac{1}{\nu}\mu^2$$

$$\kappa_k(y) = (k-1)!\,\nu\left(\frac{\mu}{\nu}\right)^k \qquad \text{für } k > 2\,.$$

Varianz ist proportional zum Quadrat des Erwartungswertes.

Lineare Exponentialfamilie: Mitglieder

- **Inverse Gaussverteilung** $y \sim \text{InvGauss}(\mu, \sigma^2)$:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{1}{2\sigma^2 y}\left(\frac{y-\mu}{\mu}\right)^2\right)$$
$$= \exp\left(\frac{y\left(-\frac{1}{2\mu^2}\right)+\frac{1}{\mu}}{\sigma^2} - \frac{1}{2\sigma^2 y} - \frac{1}{2}\log\left(2\pi\sigma^2 y^3\right)\right), \quad y > 0.$$

Mit $\theta = -\frac{1}{2\mu^2}$, $(\mu = (-2\theta)^{-1/2})$ und $\phi = \sigma^2$ gibt dies LEF mit

$$a = 1, \quad b(\theta) = -(-2\theta)^{1/2}, \quad c(y,\phi) = -\frac{1}{2}\left(\frac{1}{\phi y} + \log\left(2\pi\phi y^3\right)\right)$$

$$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = (-2\theta)^{-1/2} = \mu, \\ \mathrm{var}(y) &= \phi b''(\theta) = \phi(-2\theta)^{-3/2} = \sigma^2\mu^3, \\ \kappa_3(y) &= 3\sigma^4\mu^5, \qquad \kappa_4(y) = 15\sigma^6\mu^7. \end{aligned}$$

Varianz wächst proportional zu μ^3.

Lineare Exponentialfamilie: Mitglieder

- **Standardisierte Binomialverteilung** $my \sim \text{Binomial}(m, \pi)$:

$$
\begin{aligned}
f(y|m,\pi) = \Pr(Y = y) = \Pr(mY = my) &= \binom{m}{my} \pi^{my} (1-\pi)^{m-my} \\
&= \exp\left(\log \binom{m}{my} + my \log \pi + m(1-y) \log(1-\pi) \right) \\
&= \exp\left(\frac{y \log \frac{\pi}{1-\pi} - \log \frac{1}{1-\pi}}{1/m} + \log \binom{m}{my} \right), \quad y = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1.
\end{aligned}
$$

Mit $\theta = \log \frac{\pi}{1-\pi}$, $(\pi = e^{\theta}/(1+e^{\theta}))$ und $\phi = 1$ ist dies LEF mit

$$
a = \frac{1}{m}, \quad b(\theta) = \log \frac{1}{1-\pi} = \log(1 + \exp(\theta)), \quad c(y, \phi) = \log \binom{\frac{1}{\phi}}{\frac{y}{\phi}}.
$$

Lineare Exponentialfamilie: Mitglieder

- **Standardisierte Binomialverteilung** $my \sim \text{Binomial}(m, \pi)$:

$$\begin{aligned}
\text{E}(y) &= b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \pi \,, \\
\text{var}(y) &= a \cdot \phi b''(\theta) = \frac{1}{m} 1 \frac{\exp(\theta)}{(1 + \exp(\theta))^2} = \frac{1}{m} \pi (1 - \pi) \,, \\
\kappa_3(y) &= \frac{1}{m^2} (1 - 2\pi) \pi (1 - \pi) \,, \\
\kappa_4(y) &= \frac{1}{m^3} (1 - 6\pi(1 - \pi)) \pi (1 - \pi) \,.
\end{aligned}$$

Relative Häufigkeit y. Absolute Häufigkeit my ist binomialverteilt. Bemerke, dass Dispersion bekannt Eins ist und m reziprok als Gewicht eingeht.

Lineare Exponentialfamilie: Quasi-Likelihood

Scorefunktion für μ bei LEF hängt nur von der Varianz ab.
Idee: verwende auch Varianz zu der gar kein LEF Mitglied existiert. Man spricht von **Quasi-Scorefunktion**.

Annahme: Dispersion sei $a \cdot \phi = \phi$, also Gewicht a sei Eins.

Definition: Für eine Zufallsvariable y mit $\mathrm{E}(y) = \mu$ und $\mathrm{var}(y) = \phi V(\mu)$ mit bekannter Varianzfunktion $V(\cdot)$ ist die **(Log-)Quasi-Likelihoodfunktion** $q(\mu|y)$ definiert über die Beziehung

$$\frac{\partial q(\mu|y)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)},$$

oder äquivalent dazu durch

$$q(\mu|y) = \int^{\mu} \frac{y - t}{\phi V(t)} dt + \text{Funktion in } y \text{ (und } \phi).$$

Lineare Exponentialfamilie: Quasi-Likelihood

Die Ableitung $\partial q/\partial \mu$ nennt man **Quasi-Scorefunktion**.
Verglichen mit der Scorefunktion hat sie folgende Eigenschaften gemeinsam

$$\mathrm{E}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = 0,$$

$$\mathrm{var}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = \frac{\mathrm{var}(y)}{\phi^2 V^2(\mu)} = \frac{1}{\phi V(\mu)} = -\mathrm{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right).$$

Lineare Exponentialfamilie: Quasi-Likelihood

Satz (Wedderburn, 1974) Für eine Beobachtung y mit $\mathrm{E}(y) = \mu$ und $\mathrm{var}(y) = \phi V(\mu)$ hat die Log-Likelihoodfunktion $\ell(\mu|y) = \log f(y|\mu)$ die Eigenschaft

$$\frac{\partial \ell(\mu|y)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)},$$

genau dann, wenn die Dichte- bzw. Wahrscheinlichkeitsfunktion von y in der Form

$$\exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

geschrieben werden kann, wobei θ eine Funktion von μ, und ϕ unabhängig von μ ist.

Lineare Exponentialfamilie: Quasi-Likelihood

Beweis:

$\Rightarrow$: Integration bezüglich μ liefert

$$\begin{aligned}
\ell(\mu|y) &= \int \frac{\partial \ell(\mu|y)}{\partial \mu} d\mu = \int \frac{y-\mu}{\phi V(\mu)} d\mu \\
&= \frac{y}{\phi} \underbrace{\int \frac{1}{V(\mu)} d\mu}_{\theta} - \frac{1}{\phi} \underbrace{\int \frac{\mu}{V(\mu)} d\mu}_{b(\theta)} \\
&= \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \, .
\end{aligned}$$

Lineare Exponentialfamilie: Quasi-Likelihood

Beweis:

$\Leftarrow$: Mit den Kumulanten der LEF gilt $\mathrm{E}(y) = \mu = b'(\theta)$ und $\mathrm{var}(y) = \phi V(\mu) = \phi b''(\theta)$. Daher ist

$$\frac{d\mu}{d\theta} = \frac{db'(\theta)}{d\theta} = b''(\theta) = V(\mu)\,.$$

Da aber $\ell(\mu|y) = (y\theta - b(\theta))/\phi + c(y,\phi)$ und θ eine Funktion von μ ist, folgt

$$\begin{aligned}\frac{\partial \ell(\mu|y)}{\partial \mu} &= \frac{y}{\phi}\frac{d\theta}{d\mu} - \frac{b'(\theta)}{\phi}\frac{d\theta}{d\mu} \\ &= \frac{y-\mu}{\phi V(\mu)}\,.\end{aligned}$$

Lineare Exponentialfamilie: QL Modelle

- $V(\mu) = 1$, $\phi = \sigma^2$, $y, \mu \in \mathbb{R}$, (vgl. mit $y \sim \text{Normal}(\mu, \sigma^2)$):

$$\theta = \int d\mu = \mu\,,$$

$$q(\mu|y) = \int^{\mu} \frac{y - t}{\sigma^2} dt + \text{Funktion in } y = -\frac{(y-\mu)^2}{2\sigma^2}\,.$$

- $V(\mu) = \mu$, $0 < \mu$, $0 \leq y$, (vgl. mit $y \sim \text{Poisson}(\mu)$):

$$\theta = \int \frac{1}{\mu} d\mu = \log \mu\,,$$

$$q(\mu|y) = \int^{\mu} \frac{y - t}{t} dt = y \log \mu - \mu\,.$$

Lineare Exponentialfamilie: QL Modelle

- $V(\mu) = \mu^2$, $0 < \mu$, $0 \leq y$, (vgl. mit $y \sim \text{Gamma}(\mu, 1)$):

$$\theta = \int \frac{1}{\mu^2} d\mu = -\frac{1}{\mu}\,,$$
$$q(\mu|y) = \int^{\mu} \frac{y-t}{t^2} dt = -\frac{y}{\mu} - \log \mu\,.$$

- $V(\mu) = \mu^3$, $0 < \mu$, $0 \leq y$, (vgl. mit $y \sim \text{InvGauss}(\mu, 1)$):

$$\theta = \int \frac{1}{\mu^3} d\mu = -\frac{1}{2\mu^2}\,,$$
$$q(\mu|y) = \int^{\mu} \frac{y-t}{t^2} dt = -\frac{y}{2\mu^2} + \frac{1}{\mu}\,.$$

Lineare Exponentialfamilie: QL Modelle

- $V(\mu) = \mu^k$, $0 < \mu$, $0 \leq y$, $k \geq 3$:

$$\theta = \int \frac{1}{\mu^k} d\mu = -\frac{1}{(k-1)\mu^{k-1}},$$

$$q(\mu|y) = \int^{\mu} \frac{y-t}{t^k} dt = \frac{1}{\mu^k}\left(\frac{\mu^2}{k-2} - \frac{y\mu}{k-1}\right).$$

- $V(\mu) = \mu(1-\mu)$, $0 < \mu < 1$, $0 \leq y \leq 1$, (vgl. mit $my \sim \text{Binomial}(m, \mu)$):

$$\theta = \int \frac{1}{\mu(1-\mu)} d\mu = \log \frac{\mu}{1-\mu},$$

$$q(\mu|y) = \int^{\mu} \frac{y-t}{t(1-t)} dt = y \log \frac{\mu}{1-\mu} + \log(1-\mu).$$

Lineare Exponentialfamilie: QL Modelle

- $V(\mu) = \mu^2(1-\mu)^2$, $0 < \mu < 1$, $0 \leq y \leq 1$:

$$\theta = \int \frac{1}{\mu^2(1-\mu)^2} d\mu = 2 \log \frac{\mu}{1-\mu} - \frac{1}{\mu} + \frac{1}{1-\mu},$$

$$q(\mu|y) = \int^{\mu} \frac{y-t}{t^2(1-t)^2} dt = (2y-1) \log \frac{\mu}{1-\mu} - \frac{y}{\mu} - \frac{1-y}{1-\mu}.$$

- $V(\mu) = \mu + \mu^2/k$, $0 < \mu$, $0 \leq y$, $0 < k$, (vgl. mit $y \sim \text{NegBinomial}(k, \mu)$):

$$\theta = \int \frac{1}{\mu + \mu^2/k} d\mu = \log \frac{\mu}{k+\mu},$$

$$q(\mu|y) = \int^{\mu} \frac{y-t}{t+t^2/k} dt = y \log \frac{\mu}{k+\mu} + k \log \frac{1}{k+\mu}.$$

Lineare Exponentialfamilie: Quasi Dichte

Durch Spezifikation der Erwartungswert/Varianz-Beziehung ist auch Dichtefunktion spezifizierbar.
Aus der (Log)-QL Funktion folgt mit

$$\omega(\mu) = \int_{\mathbb{R}} \exp(q(\mu|y))dy$$

als **Quasi-Dichte** (vgl. Nelder & Lee, 1992)

$$f_q(y|\mu) = \frac{\exp(q(\mu|y))}{\omega(\mu)} .$$

Normalisierungsfunktion $\omega(\mu) \neq 1$, wenn Varianz $\phi V(\mu)$ zu keinem LEF Mitglied gehört. Andererseits ist $\omega(\mu) = 1$, $\forall \mu$, falls zur Varianz eine LEF existiert.

Lineare Exponentialfamilie: Quasi Dichte

Zur Quasi-Dichte $f_q(y|\mu)$ korrespondiert Log-Likelihoodfunktion

$$\ell_q(\mu|y) = \log(f_q(y|\mu)) = q(\mu|y) - \log(\omega(\mu))$$

und Scorefunktion

$$\frac{\partial \ell_q(\mu|y)}{\partial \mu} = \frac{\partial q(\mu|y)}{\partial \mu} - \frac{\partial \log(\omega(\mu))}{\partial \mu}.$$

Dieser Score unterscheidet sich vom Quasi-Score um

$$\begin{aligned}
\frac{\partial \log(\omega(\mu))}{\partial \mu} &= \frac{1}{\omega(\mu)} \frac{\partial \omega(\mu)}{\partial \mu} = \frac{1}{\omega(\mu)} \int \frac{\partial \exp(q(\mu|y))}{\partial \mu} dy \\
&= \frac{1}{\omega(\mu)} \int \frac{\partial q(\mu|y)}{\partial \mu} \exp(q(\mu|y)) dy \\
&= \int \frac{y-\mu}{\phi V(\mu)} f_q(y|\mu) dy = \mathrm{E}_q\left(\frac{y-\mu}{\phi V(\mu)}\right) = \frac{\mu_q - \mu}{\phi V(\mu)}.
\end{aligned}$$

Lineare Exponentialfamilie: Quasi Dichte

Unterschied liegt in

$$\frac{\partial \log(\omega(\mu))}{\partial \mu} = \frac{\mu_q - \mu}{\phi V(\mu)}.$$

Hierbei bezeichnet

$$\mu_q = \int y f_q(y|\mu) dy$$

den Quasi-Mean von y. Falls $\mu_q - \mu$ verglichen mit $y - \mu$ klein ist, ist der Maximum-Quasi-Likelihood Schätzer nahe dem Maximum-Likelihood Schätzer bezüglich der Quasi-Verteilung.

Generalisiertes Lineares Modell (GLM)

Klasse der Generalisierten Linearen Modelle (GLM):
Parametrisierung der Form

stochastische Komponente: $y_i \overset{ind}{\sim} \text{LEF}(\theta_i),\ \text{E}(y_i) = \mu_i = \mu(\theta_i)$

systematische Komponente: $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$

Linkfunktion: $g(\mu_i) = \eta_i$.

Responsevektor $\mathbf{y} = (y_1, \ldots, y_n)^\top$ aus unabhängigen y_i mit $\text{E}(y_i) = \mu_i$ und $\text{var}(y_i) = a_i \phi V(\mu_i)$. Dispersion ist Produkt $a_i \phi$.

$\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{i,p-1})^\top$ sei $p \times 1$ Vektor bekannter Prädiktoren zusammengefasst zur $n \times p$ Designmatrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ sei $p \times 1$ Vektor unbekannter Parameter,

$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ sei $n \times 1$ Vektor mit den Linearen Prädiktoren,

$g(\cdot)$ sei eine bekannte (monoton und 2x stetig differenzierbare) Linkfunktion.

Generalisiertes Lineares Modell (GLM)

Wesentlichen Unterschiede zum herkömmlichen LM sind:

- keine allgemeine Additivität bzgl. nicht-beobachtbarer Fehlerterme ϵ_i wie beim LM,
- Varianzstruktur kann auch vom Erwartungswert abhängen,
- Funktion des Erwartungswertes wird linear modelliert (nicht zu verwechseln mit einer Transformation der Response).

Interessiert in der Konstruktion eines Schätzers von $\boldsymbol{\beta}$ und an einem Maß für die Güte der Anpassung. Beides ist für MLE sehr einfach und stellt nur Generalisierung der Resultate beim LM dar.

Generalisiertes Lineares Modell: MLE

Seien $y_1, \ldots, y_n$ unabhängige Responses aus derselben LEF mit Parameter (θ_i, ϕ), so ist die Log-Likelihood der Stichprobe

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a_i \phi} + c(y_i, \phi) \right) .$$

Unter der Annahme $\mu = \mu(\boldsymbol{\beta})$ folgt die Scorefunktion

$$\frac{\partial \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} , \qquad j = 0, 1, \ldots, p-1 .$$

Mit der Definition des linearen Prädiktors $\eta = \mathbf{x}^\top \boldsymbol{\beta}$ gilt beim GLM

$$\frac{\partial \mu}{\partial \boldsymbol{\beta}} = \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \boldsymbol{\beta}} = \frac{\partial \mu}{\partial g(\mu)} \mathbf{x} = \frac{\mathbf{x}}{g'(\mu)}$$

und deshalb

$$\frac{\partial \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} , \qquad j = 0, 1, \ldots, p-1 .$$

Generalisiertes Lineares Modell: MLE

Da $b'(\theta_i) = \mu_i$ und wegen $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ gilt

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}))\,.$$

Den Link $g(\cdot) = b'^{-1}(\cdot)$, also $g(\mu_i) = \theta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, nennt man **kanonisch**. Damit wird θ direkt durch η modelliert und es folgt

$$g'(\mu) = \frac{\partial g(\mu)}{\partial \mu} = \frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)}\,.$$

Scorefunktion vereinfacht sich für **kanonische Links** zu

$$\frac{\partial \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i \phi} x_{ij}\,, \qquad j = 0, 1, \ldots, p\,.$$

Falls $a_i = 1$ gilt bei Modellen mit Intercept ($x_{i0} = 1,\ \forall i$)

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i\,.$$

Generalisiertes Lineares Modell: MLE

Sind alle $a_i = 1$ und die Dispersion ϕ bekannt, so folgt bei kanonischen Links

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}(\boldsymbol{\beta})) &= \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))}{\phi}\right) \prod_{i=1}^{n} \exp(c(y_i, \phi)) \\ &= \exp\left(\frac{1}{\phi}\sum_{i=1}^{n} y_i \mathbf{x}_i^\top \boldsymbol{\beta} - b(\mathbf{x}_i^\top \boldsymbol{\beta})\right) \prod_{i=1}^{n} \exp(c(y_i, \phi)) \\ &= g(T(\mathbf{y})|\boldsymbol{\beta}) h(\mathbf{y}) \,, \end{aligned}$$

und $T(\mathbf{y}) = \mathbf{X}^\top \mathbf{y}$ ist somit **suffiziente Statistik** für $\boldsymbol{\beta}$. Bemerke, dass $\dim(T(\mathbf{y})) = \dim(\boldsymbol{\beta}) = p$ gilt, und dass eine suffiziente Statistik ausschließlich bei einer kanonischen Linkfunktion existiert.

Generalisiertes Lineares Modell: MLE

MLE $\hat{\boldsymbol{\beta}}$ ist Nullstelle der Scorefunktion. Gleichungssystem nur numerisch (iterativ) lösbar. Einzige Ausnahme ist das LM, in dem μ linear in $\boldsymbol{\beta}$ ist. Für alle anderen Situationen ist $\mu = g^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$.

Newton-Raphson Methode liefert Iteration $(t = 0, 1, \dots)$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right)^{-1} \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} \frac{\partial \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta}} \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}.$$

Ableitungen auf der rechten Seite sind in $\boldsymbol{\beta}^{(t)}$ betrachtet. In Matrixnotation folgt für den Scorevektor

$$\frac{\partial \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \mathbf{X}^\top \mathbf{D} \mathbf{W} (\mathbf{y} - \boldsymbol{\mu}),$$

mit $\mathbf{D} = \operatorname{diag}(d_i)$ und $\mathbf{W} = \operatorname{diag}(w_i)$, wobei

$$d_i = g'(\mu_i),$$
$$w_i = \left(a_i V(\mu_i) g'^2(\mu_i) \right)^{-1}.$$

Generalisiertes Lineares Modell: MLE

Als negative Hessematrix der Log-Likelihoodfunktion resultiert

$$-\frac{\partial^2 \ell(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{1}{\phi}\mathbf{X}^\top \left(\frac{\partial \mathbf{DW}}{\partial \boldsymbol{\eta}^\top}(\mathbf{y}-\boldsymbol{\mu}) - \mathbf{DW}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^\top} \right) \mathbf{X}$$
$$= \frac{1}{\phi}\mathbf{X}^\top \left(\mathbf{W} - \frac{\partial \mathbf{DW}}{\partial \boldsymbol{\eta}^\top}(\mathbf{y}-\boldsymbol{\mu}) \right) \mathbf{X} ,$$

wegen $\partial \boldsymbol{\mu}/\partial \boldsymbol{\eta}^\top = \mathbf{D}^{-1}$. Weiters ist $d_i w_i = \left(a_i V(\mu_i) g'(\mu_i)\right)^{-1}$

$$\frac{\partial d_i w_i}{\partial \eta_i} = -\frac{a_i V'(\mu_i)\frac{\partial \mu_i}{\partial \eta_i} g'(\mu_i) + a_i V(\mu_i) g''(\mu_i)\frac{\partial \mu_i}{\partial \eta_i}}{(a_i V(\mu_i) g'(\mu_i))^2}$$
$$= -\frac{V'(\mu_i) g'(\mu_i) + V(\mu_i) g''(\mu_i)}{a_i V^2(\mu_i) g'^3(\mu_i)} .$$

Generalisiertes Lineares Modell: MLE

Zusammenfassen von

$$w_i^* = w_i - \frac{\partial d_i w_i}{\partial \eta_i}(y_i - \mu_i)$$

zu $\mathbf{W}^*$ (mit $\mathrm{E}(\mathbf{W}^*) = \mathbf{W}$) liefert Newton-Raphson Vorschrift

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\mathbf{X}^\top \mathbf{W}^{*(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}^{(t)} \mathbf{W}^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}), \quad t = 0, 1, \ldots .$$

Bemerke, dass das Produkt Scorevektor mal inverse Hessematrix hier **unabhängig vom Dispersionsparameter** ϕ ist.

Generalisiertes Lineares Modell: MLE

Mittels **Pseudobeobachtungen** (adjusted dependent variates)

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{*-1}\mathbf{D}\mathbf{W}(\mathbf{y} - \boldsymbol{\mu})$$

die Newton-Raphson Vorschrift umschreiben in eine **Iterative (Re)Weighted Least Squares** Prozedur (IWLS/IRLS)

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^\top \mathbf{W}^{*(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{*(t)} \mathbf{z}^{(t)} .$$

Generalisiertes Lineares Modell: MLE

Für **kanonische Links** gilt $g'(\mu) = 1/V(\mu)$ und $g''(\mu) = -V'(\mu)/V^2(\mu)$ und es verschwinden die Ableitungen

$$\frac{\partial d_i w_i}{\partial \eta_i} = -\frac{V'(\mu_i)/V(\mu_i) - V'(\mu_i)/V(\mu_i)}{a_i/V(\mu_i)} = 0$$

$\Rightarrow \mathbf{W}^* = \mathbf{W}$, und Pseudobeobachtungen vereinfachen sich zu

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

mit $\mathbf{V} = \text{diag}(V(\mu_i))$. Dies liefert als Iterationsvorschrift

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}.$$

Generalisiertes Lineares Modell: MLE

Erinnerung an LM: LS Schätzer definiert als

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

(keine iterative Lösung notwendig da $\mathbf{W} = \mathbf{I}$ und $\mathbf{z} = \mathbf{y}$).

Generell wird gerne mit erwarteter Information gearbeiten. Wegen $\mathrm{E}(\mathbf{X}^\top \mathbf{W}^* \mathbf{X}) = \mathbf{X}^\top \mathbf{W} \mathbf{X}$ folgt für diese **Fisher Scoring Technik** wieder

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}$$

mit den Pseudobeobachtungen

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}(\mathbf{y} - \boldsymbol{\mu}) \,.$$

Dafür ist $\mathrm{E}(\mathbf{z}) = \mathbf{X}\boldsymbol{\beta}$ und wegen $\mathrm{var}(\mathbf{y}) = \phi(\mathbf{D}\mathbf{W}\mathbf{D})^{-1}$ folgt $\mathrm{var}(\mathbf{z}) = \mathbf{D}\,\mathrm{var}(\mathbf{y})\mathbf{D} = \phi \mathbf{W}^{-1}$.

Generalisiertes Lineares Modell: MLE Asymptotik

Entwickle Scorefunktion um wahren Parameter $\boldsymbol{\beta}$, d.h.

$$\mathbf{0} = \left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\mu})}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} \approx \left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\mu})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} + \left.\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\mu})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) .$$

Ersetze Information durch Erwartungswert $-\mathbf{X}^\top\mathbf{W}\mathbf{X}$, also

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}\mathbf{W}(\mathbf{y} - \boldsymbol{\mu}) .$$

Wegen $\text{var}(\mathbf{y}) = \phi(\mathbf{D}\mathbf{W}\mathbf{D})^{-1}$ ergibt sich

$$\begin{aligned} \text{E}(\hat{\boldsymbol{\beta}}) &\approx \boldsymbol{\beta} \\ \text{var}(\hat{\boldsymbol{\beta}}) &\approx (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}\mathbf{W}\,\text{var}(\mathbf{y})\mathbf{W}\mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} = \phi(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} , \end{aligned}$$

mit $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta})$. Fahrmeir & Kaufmann (1985) zeigten sogar

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{D} \text{Normal}_p(\mathbf{0}, n\phi(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}) .$$

Generalisiertes Lineares Modell: Pearson Statistik

Schätzer für Varianzmatrix $\text{var}(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X})^{-1}$

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}$$

hängt möglicherweise noch von ϕ ab.
LM: σ^2 durch biaskorrigierte mittlere Fehlerquadratsumme

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

schätzen. Unter Annahme einer LEF gilt $\phi = \text{var}(y_i)/a_i V(\mu_i)$, für alle $i = 1, \ldots, n$. Falls $\boldsymbol{\beta}$ bekannt, würden μ_i bekannt sein und

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{a_i V(\mu_i)}$$

wäre erwartungstreu für ϕ.

Generalisiertes Lineares Modell: Pearson Statistik

Da aber $\boldsymbol{\beta}$ unbekannt, verwendet man biaskorrigierte Größe

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)} = \frac{1}{n-p} X^2 \,,$$

was auch als mittlere (generalisierte) **Pearson Statistik** X^2 bezeichnet wird. Die einzelnen (nicht quadrierten) Summanden von X^2, d.h.

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i V(\hat{\mu}_i)}} \,,$$

nennt man **Pearson Residuen**. Diese entsprechen den gewöhnlichen (nicht standardisierten) Residuen $y_i - \hat{\mu}_i$ beim LM.

Generalisiertes Lineares Modell: Scorefunktion

Für $y_i \overset{ind}{\sim} \text{LEF}(\theta_i(\boldsymbol{\beta}))$, $i = 1, \ldots, n$, betrachten wir die Momente der Scorefunktion (bzgl. $\boldsymbol{\beta}$). Sei $\ell_i = \log f(y_i|\theta_i(\boldsymbol{\beta}))$ und $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, dann hält für $0 \leq j, k \leq p-1$

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j},$$

$$\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} = -\left(\frac{1}{a_i \phi V(\mu_i)} + (y_i - \mu_i) \frac{a_i \phi V'(\mu_i)}{(a_i \phi V(\mu_i))^2} \right) \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} + (y_i - \mu_i) \frac{1}{a_i \phi V(\mu_i)} \frac{\partial^2 \mu_i}{\partial \beta_j \partial \beta_k},$$

$$\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k} = \left(\frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \right)^2 \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k}.$$

Generalisiertes Lineares Modell: Scorefunktion

Deren Erwartungswerte genügen den bekannten Eigenschaften

$$\mathrm{E}\left(\frac{\partial \ell_i}{\partial \beta_j}\right) = 0\,,$$

$$\mathrm{E}\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = -\frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k}\,,$$

$$\mathrm{E}\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k}\right) = \frac{\mathrm{var}(y_i)}{(a_i \phi V(\mu_i))^2} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} = \frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k}$$

$$= -\mathrm{E}\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right)\,.$$

Hält dies auch für die Momente der Quasi-Scorefunktion?

Generalisiertes Lineares Modell: Scorefunktion

Annahmen beim QL Modell: $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{var}(y_i) = a_i \phi V(\mu_i)$, sowie $g(\mu_i) = x_i^\top \beta$.

Als Quasi-Scorefunktion resultiert

$$\frac{\partial q(\mu_i(\boldsymbol{\beta})|y_i)}{\partial \beta_j} = \frac{\partial q(\mu_i(\boldsymbol{\beta})|y_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} .$$

Entspricht genau dem i-ten Summanden in der Scorefunktion.

Man spricht auch bei QL Ansätzen von **kanonischem** Link, falls $g'(\mu) = 1/V(\mu)$ gilt.

MQL Schätzer entspricht dem MLE, falls zur angenommenen Varianz ein Mitglied aus der LEF existiert.

Generalisiertes Lineares Modell: Scorefunktion

Sei $q_i = q(\mu_i(\boldsymbol{\beta})|y_i)$, $i = 1, \ldots, n$, so gilt

$$\mathrm{E}\left(\frac{\partial q_i}{\partial \beta_j}\right) = \mathrm{E}\left(\frac{\partial q_i}{\partial \mu_i}\right)\frac{\partial \mu_i}{\partial \beta_j} = 0\,,$$

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 q_i}{\partial \beta_j \partial \beta_k}\right) &= \mathrm{E}\left(\frac{\partial^2 q_i}{\partial \mu_i^2}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k} + \frac{\partial q_i}{\partial \mu_i}\frac{\partial^2 \mu_i}{\partial \beta_j \partial \beta_k}\right)\\ &= -\frac{1}{a_i \phi V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\,,\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(\frac{\partial q_i}{\partial \beta_j}\frac{\partial q_i}{\partial \beta_k}\right) &= \mathrm{E}\left(\left(\frac{\partial q_i}{\partial \mu_i}\right)^2\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\right) = \frac{1}{a_i \phi V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\\ &= -\mathrm{E}\left(\frac{\partial^2 q_i}{\partial \beta_j \partial \beta_k}\right)\,.\end{aligned}$$

All die Resultate von zuvor halten auch für QL Modelle.

Generalisiertes Lineares Modell: Deviance

Bewerte Güte der **Modellanpassung**?
LRT zu $H_0 : \mu = g^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$ gegen $H_1 : \mu \neq g^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$.
H_0: das betrachtete Modell mit p Parametern ist korrekt.
LRT-Statistik

$$\Lambda(\mathbf{y}) = \frac{\sup\limits_{\boldsymbol{\mu}=g^{-1}(\boldsymbol{\eta})} L(\boldsymbol{\mu}|\mathbf{y})}{\sup\limits_{\boldsymbol{\mu}} L(\boldsymbol{\mu}|\mathbf{y})} .$$

MLE $\hat{\boldsymbol{\beta}}$ maximiert die Likelihoodfunktion unter H_0.
Das uneingeschränktes Maximum erzielt man für $\hat{\mu}_i = y_i$ für alle $i = 1, \ldots, n$. Man spricht hierbei vom **vollen (saturierten) Modell** mit n frei wählbaren Parameter. Dies liefert

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\boldsymbol{\mu}}|\mathbf{y})}{L(\mathbf{y}|\mathbf{y})} .$$

Wir bilden $-2 \log \Lambda(\mathbf{y})$ und erhalten die **skalierte Deviance**

$$\frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\Big(\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y})\Big) .$$

Generalisiertes Lineares Modell: Deviance

Diese Schreibweise ist für LEF berechtigt, denn es gilt

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y}) &= \sum_{i=1}^{n}\left(\frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{a_i\phi} + c(y_i,\phi)\right) \\ &\quad - \sum_{i=1}^{n}\left(\frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{a_i\phi} + c(y_i,\phi)\right) \\ &= \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i\left(\hat{\theta}_i - \tilde{\theta}_i\right) - \left(b(\hat{\theta}_i) - b(\tilde{\theta}_i)\right)}{a_i}\,,\end{aligned}$$

wobei $\tilde{\theta}$ jenen Wert von $\theta(\mu)$ bezeichnet, falls $\mu = y$.
Die unskalierte Deviance auch schreibbar als

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i\,.$$

$\Rightarrow$ alternative Definition von Residuen. **Deviance Residuen**:

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}\,.$$

Generalisiertes Lineares Modell: Deviance

MLE $\hat{\boldsymbol{\mu}}$ maximiert $\ell(\boldsymbol{\mu}|\mathbf{y})$ und minimiert die Deviance (da $\ell(\mathbf{y}|\mathbf{y})$ unabhängig von Parameter). Deviance ist eine Verallgemeinerung der Fehlerquadratsumme beim LM.

Beispiel: Seien $y_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2)$ und σ^2 fest. So ist

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

$$\ell(\mathbf{y}|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2)\,.$$

Die skalierte Deviance ist somit

$$\frac{1}{\phi}D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2}\,\text{SSE}(\hat{\boldsymbol{\beta}}) \sim \chi^2_{n-p}\,.$$

Für andere Mitglieder der LEF wird dieses Ergebnis auch gerne verwendet, jedoch ist dabei große Vorsicht geboten.

Generalisiertes Lineares Modell: Deviance

Beispiel: $m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)$ mit $y_i = 0, 1/m_i, \ldots, 1$, dann

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i} - m_i \log \frac{1}{1-\hat{\mu}_i} + \log \binom{m_i}{m_i y_i} \right\}$$

$$\ell(\mathbf{y}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{y_i}{1-y_i} - m_i \log \frac{1}{1-y_i} + \log \binom{m_i}{m_i y_i} \right\} .$$

Wegen $\phi = 1$ und $a_i = 1/m_i$ resultiert als (skalierte) Deviance

$$\frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2 \sum_{i=1}^{n} \left\{ m_i y_i \left(\log \frac{\hat{\mu}_i}{y_i} + \log \frac{1-y_i}{1-\hat{\mu}_i} \right) - m_i \log \frac{1-y_i}{1-\hat{\mu}_i} \right\}$$

$$= 2 \sum_{i=1}^{n} m_i \left\{ (1-y_i) \log \frac{1-y_i}{1-\hat{\mu}_i} + y_i \log \frac{y_i}{\hat{\mu}_i} \right\} .$$

Bemerke, dass für $y_i = 0$ oder $y_i = 1$ unabhängig von $\hat{\mu}_i$ (da $x \log x = 0$ für $x = 0$) der dazugehörende Teil in der Deviance-Komponente verschwindet.

Generalisiertes Lineares Modell: Deviance

Entsprechend kann man auch eine **Quasi-Deviance** definieren als

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\phi\Big(q(\hat{\boldsymbol{\mu}}|\mathbf{y}) - q(\mathbf{y}|\mathbf{y})\Big) = -2\sum_{i=1}^{n}\int_{y_i}^{\hat{\mu}_i}\frac{y_i - t}{V(t)}dt\,.$$

Diese Größe ist positiv außer an der Stelle $\mathbf{y} = \hat{\boldsymbol{\mu}}$. Natürlich erzeugt auch hier der MQLE die minimale Quasi-Deviance.

Generalisiertes Lineares Modell: Deviance

Beispiel: Für Prozentwerte y_i ist $V(\mu_i) = \mu_i(1-\mu_i)$ manchmal zu groß. Alternative Varianzfunktion $V(\mu_i) = \mu_i^2(1-\mu_i)^2$.

Als QL Komponente folgt für $0 < \hat{\mu}_i < 1$ und $0 \leq y_i \leq 1$

$$\begin{aligned} q(\hat{\mu}_i|y_i) &= \int^{\hat{\mu}_i} \frac{y_i - t}{t^2(1-t)^2} dt \\ &= (2y_i - 1) \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i} - \frac{y_i}{\hat{\mu}_i} - \frac{1-y_i}{1-\hat{\mu}_i} \,. \end{aligned}$$

Generalisiertes Lineares Modell: Deviance

Die entsprechende Quasi-Deviance ist daher

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2 \sum_{i=1}^{n} \left\{ (2y_i - 1) \left(\log \frac{1 - y_i}{1 - \hat{\mu}_i} - \log \frac{y_i}{\hat{\mu}_i} \right) + (2\hat{\mu}_i - 1) \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i (1 - \hat{\mu}_i)} \right\}.$$

Bemerke: in $y_i = 0$ und $y_i = 1$ ist Quasi-Deviance nicht definiert, der MQLE $\hat{\boldsymbol{\beta}}$ jedoch existiert sehr wohl.

MQLE $\hat{\boldsymbol{\beta}}$ ist definiert als die Lösung von $\partial q(\mu(\boldsymbol{\beta})|\mathbf{y})/\partial \boldsymbol{\beta} = \mathbf{0}$ (für beide kritischen Beobachtungen unproblematisch).

Generalisiertes Lineares Modell: MQLE

Annahme: $y_1, \ldots, y_n$ seien unabhängige Responses mit $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{var}(y_i) = \phi V(\mu_i)$.

Postulieren wir noch das strukturelle Modell $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, so ist der MQLE $\hat{\boldsymbol{\beta}}$ definiert als Nullstelle der Quasi-Scorefunktion

$$\begin{aligned} U(\boldsymbol{\beta}|\mathbf{y}) &= \frac{\partial q(\boldsymbol{\mu}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \,. \end{aligned}$$

Generalisiertes Lineares Modell: MQLE

Mit $\mathbf{V} = \text{diag}(V(\mu_i))$ und $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}^\top$ folgt als Quasi-Score

$$U(\boldsymbol{\beta}|\mathbf{y}) = \frac{1}{\phi}\mathbf{D}^\top\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\,.$$

Bereits gezeigt, dass

$$\begin{aligned}
\text{E}(U(\boldsymbol{\beta}|\mathbf{y})) &= \mathbf{0} \\
\text{var}(U(\boldsymbol{\beta}|\mathbf{y})) &= -\text{E}\left(\frac{\partial}{\partial\boldsymbol{\beta}^\top}U(\boldsymbol{\beta}|\mathbf{y})\right) = \frac{1}{\phi}\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}\,.
\end{aligned}$$

Für QL Modelle spielt $\frac{1}{\phi}\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}$ dieselbe Rolle wie die Fisher-Information bei Likelihood Modellen. Im Speziellen ist die asymptotische Varianz/Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ gleich

$$\text{var}(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D})^{-1}\,.$$

Generalisiertes Lineares Modell: MQLE

MQLE $\hat{\boldsymbol{\beta}}$ mittels Newton-Raphson Methode mit Fisher-Scoring. Initialisierung mit $\boldsymbol{\beta}^{(0)}$ ergibt erste Update

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + (\mathbf{D}^{(0)\top}\mathbf{V}^{(0)-1}\mathbf{D}^{(0)})^{-1}\mathbf{D}^{(0)\top}\mathbf{V}^{(0)-1}(\mathbf{y} - \boldsymbol{\mu}^{(0)}) .$$

MQLE $\hat{\boldsymbol{\beta}}$ resultiert bei Konvergenz. Bemerke, dass auch diese Iterationsvorschrift unabhängig vom Dispersionsparameter ϕ ist.

MQLE verhält sich genau so wie MLE. Eine ML Schätzung von ϕ ist jedoch unpraktikabel. Daher verwenden wir wieder die mittlere Pearson Statistik

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} .$$

Generalisiertes Lineares Modell: Tests

Konzept der **nested models** (ineinander geschachtelte Untermodelle). Entspricht dem Testen von Hypothesen der Form

$$H_0 : \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1}$$
$$H_1 : \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1} + \beta_q x_q + \cdots + \beta_{p-1} x_{p-1} \,,$$

($q < p$) oder äquivalent dazu dem Test von

$$H_0 : \beta_q = \cdots = \beta_{p-1} = 0$$
$$H_1 : \beta_0, \ldots, \beta_{p-1} \text{ beliebig.}$$

Sei M das unter H_1 spezifizierte Modell, und M_0 jenes unter H_0 (somit $M_0 \subset M$). Bezeichnen $\hat{\mu}_0$ und $\hat{\mu}$ die geschätzten Erwartungen, dann erhalten wir als korrespondierende Deviancen

$$D(M_0) = D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) = -2\phi\Big(\ell(\hat{\boldsymbol{\mu}}_0|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y})\Big)$$
$$D(M) = D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\phi\Big(\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y})\Big) \,.$$

Generalisiertes Lineares Modell: Tests

Differenz dieser beiden Deviancen

$$D(M_0|M) = D(M_0) - D(M) = D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})$$
$$= -2\phi\Big(\ell(\hat{\boldsymbol{\mu}}_0|\mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}|\mathbf{y})\Big)$$

beschreibt den Unterschied in Güte beider Modelle.

Die Differenz der skalierten Deviancen $D(M_0|M)/\phi$ entspricht der **Likelihood-Quotienten Teststatistik** um H_0 zu testen.

Generalisiertes Lineares Modell: Tests

Deviance des eingeschränkten Modells zerlegen in

$$D(M_0) = D(M_0|M) + D(M)\,.$$

Term $D(M_0|M)$ beschreibt Zuwachs in der Diskrepanz zwischen Daten und Modellanpassung wenn M_0 anstatt M verwendet wird. Für **normalverteilte Responses** entspricht dies der Zerlegung

$$\mathrm{SSE}(M_0) = \Big(\mathrm{SSE}(M_0) - \mathrm{SSE}(M)\Big) + \mathrm{SSE}(M)\,.$$

Die beiden Terme rechts sind stochastisch unabhängig mit $\mathrm{SSE}(M)/\sigma^2 \sim \chi^2_{n-p}$ und $(\mathrm{SSE}(M_0) - \mathrm{SSE}(M))/\sigma^2 \sim \chi^2_{p-q}$. Um H_0 zu testen verwenden wir beim LM die Statistik

$$\frac{\Big(\mathrm{SSE}(M_0) - \mathrm{SSE}(M)\Big)/(p-q)}{\mathrm{SSE}(M)/(n-p)} \sim F_{p-q,n-p}\,,$$

Generalisiertes Lineares Modell: Tests

Falls ϕ **bekannt**: betrachte bei GLM die Deviancereduktion

$$\frac{D(M_0) - D(M)}{\phi} = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi},$$

die (Regularitätsbedingungen!) asymptotisch χ^2_{p-q}-verteilt ist.

Falls ϕ **unbekannt**: verwende

$$\frac{\Big(D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})\Big)/(p-q)}{D(\mathbf{y}, \hat{\boldsymbol{\mu}})/(n-p)} \sim F_{p-q,n-p}\,.$$

Hierbei ist ϕ unter H_1 geschätzt. Statt $D(\mathbf{y}, \hat{\boldsymbol{\mu}})/(n-p)$ kann auch $X^2/(n-p)$ zur Dispersionsschätzung verwendet werden.

Generalisiertes Lineares Modell: Tests

Wald Test: Hypothese der Form $H_0 : \beta_j = 0$, $j = 1, \ldots, p-1$.

Teststatistik basiert auf MLE $\hat{\boldsymbol{\beta}}$, d.h.

$$\left(\frac{\hat{\beta}_j}{\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}} \right)^2 \overset{H_0}{\sim} \chi^2_1 \,.$$

Generalisiertes Lineares Modell: Tests

Verallgemeinern auf **allgemeine lineare Hypothese**

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi} \qquad \text{gegen} \qquad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi} \, .$$

$\mathbf{C}$ feste $s \times p$ Matrix ($\text{rank}(\mathbf{C}) = s \leq p$), $\boldsymbol{\xi}$ fester $s \times 1$ Vektor.
Erinnerung: $\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \mathbf{F}^{-1}(\boldsymbol{\beta}))$ mit Fisher Information

$$\mathbf{F}(\boldsymbol{\beta}) = -\text{E}\left(\frac{\partial^2 \ell(\boldsymbol{\mu}|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right) = \frac{1}{\phi}(\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X}) \, .$$

Unter H_0 ist $\text{E}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$ mit $\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}^\top$.
Betrachte daher

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})^\top \left(\mathbf{C}\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}^\top\right)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}) \overset{H_0}{\sim} \chi^2_s \, .$$

Statistik misst Distanz zwischen $\mathbf{C}\hat{\boldsymbol{\beta}}$ und $\mathbf{C}\boldsymbol{\beta}$, wobei mit der inversen (asymptotischen) Kovarianzmatrix $\mathbf{C}\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}^\top$ gewichtet wird.

Generalisiertes Lineares Modell: Tests

Likelihood-Quotienten Test: betrachte dazu

$$-2\Big(\ell(\tilde{\boldsymbol{\beta}}|\mathbf{y}) - \ell(\hat{\boldsymbol{\beta}}|\mathbf{y})\Big) \overset{H_0}{\sim} \chi^2_s\,,$$

die Abweichung zwischen unrestringierten Maximum $\ell(\hat{\boldsymbol{\beta}}|y)$ und dem unter H_0 restringierten Maximum $\ell(\tilde{\boldsymbol{\beta}}|\mathbf{y})$, wobei $\tilde{\boldsymbol{\beta}}$ der MLE unter der Restriktion $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$ ist.

Beide Modelle (unter H_0 und unter H_1) sind anzupassen, um LRT Statistik berechnen zu können.

Generalisiertes Lineares Modell: Tests

Score Test: Scorefunktion $\mathbf{s}(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{X}^\top \mathbf{D}\mathbf{W}(\mathbf{y} - \boldsymbol{\mu})$ für das unrestringierte Modell ist $\mathbf{0}$, falls im unrestringierten MLE $\hat{\boldsymbol{\beta}}$ ausgewertet, also $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$.

Ersetzen wir $\hat{\boldsymbol{\beta}}$ durch (unter H_0) restringierten MLE $\tilde{\boldsymbol{\beta}}$, so wird $\mathbf{s}(\tilde{\boldsymbol{\beta}})$ sich signifikant von $\mathbf{0}$ unterscheiden, wenn H_0 nicht zutrifft.

Wegen $\mathrm{E}(\mathbf{s}(\boldsymbol{\beta})) = \mathbf{0}$ und $\mathrm{var}(\mathbf{s}(\boldsymbol{\beta})) = \mathrm{E}(\mathbf{s}(\boldsymbol{\beta})\mathbf{s}(\boldsymbol{\beta})^\top) = \mathbf{F}(\boldsymbol{\beta})$ liegt es nahe, als Score Teststatistik

$$\mathbf{s}(\tilde{\boldsymbol{\beta}})^\top \mathbf{F}^{-1}(\tilde{\boldsymbol{\beta}})\mathbf{s}(\tilde{\boldsymbol{\beta}}) \overset{H_0}{\sim} \chi^2_s$$

zu verwenden.

Generalisiertes Lineares Modell: Konstante Varianz

Beispiel Black Cherry Trees: V wurde Box-Cox transformiert und angenommen, dass $V^{1/3}$ oder $\log V$ normalverteilt.

Alternative Annahme: V normalverteilt mit $\mathrm{E}(V) = \mu$ und $g(\mu) = \eta$ mit $g(\mu) \neq \mathrm{id}(\mu)$.

Habe V konstante Varianz und sei $g(\mu) = \mu^{1/3} = \eta$ $(\mu = \eta^3)$ mit linearem Prädiktor $\eta = \beta_0 + \beta_1 H + \beta_2 D$.

```
> attach(trees)
> (pm <- glm(V ~ H + D, family = gaussian(link=power(1/3))))

Coefficients:
(Intercept)            H            D
   -0.05132      0.01429      0.15033

Degrees of Freedom: 30 Total (i.e. Null);  28 Residual
Null Deviance:      8106
Residual Deviance: 184.2        AIC: 151.2
```

Generalisiertes Lineares Modell: Konstante Varianz

`Null` Modell: Modell nur mit Intercept (iid) mit df $= n - 1 = 30$ und Deviance 8106.

Mit beiden Prädiktoren im Modell Deviance $= 184.2$ bei df $= n - p = 28$.

Unter AIC versteht man das Akaike Informationskriterium

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) + 2k\,,$$

Hier ist die Parameteranzahl $k = 4$ (β_0, β_1, β_2, und ϕ).

Generalisiertes Lineares Modell: Konstante Varianz

```
> AIC(pm)
[1] 151.2102

> logLik(pm) # maximized log-likelihood function
'log Lik.' -71.60508 (df=4)
> sum(log(dnorm(V, pm$fit, sqrt(summary(pm)$dispersion*28/31))))
[1] -71.60508

> -2*logLik(pm) + 2*4
'log Lik.' 151.2102 (df=4)

> sum(residuals(pm)^2) # compare with Residual Deviance
[1] 184.1577
> sum((V-mean(V))^2) # Null Deviance
[1] 8106.084
```

Schätzer sind fast identisch jenen aus Beispiel in Abschnitt 1.3.

Generalisiertes Lineares Modell: Konstante Varianz

Statt Annahme der Lognormal-Verteilung für V können wir Log-Link verwenden, also $g(\mu) = \log \mu = \eta$ (bzw. $\mu = \exp(\eta)$), mit $\eta = \beta_0 + \beta_1 \log H + \beta_2 \log D$.

```
> glm(V ~ log(H) + log(D), family = gaussian(link=log))

Coefficients:
(Intercept)       log(H)       log(D)
     -6.537        1.088        1.997

Degrees of Freedom: 30 Total (i.e. Null);  28 Residual
Null Deviance:      8106
Residual Deviance: 179.7        AIC: 150.4
```

Schätzer entsprechen etwa jenen unter Lognormal-Verteilung in Abschnitt 1.3. Deviance ist etwas geringer als zuvor.

Generalisiertes Lineares Modell: Quasi-Likelihood

Beispiel Bioassay: Verfahren zur Kontrolle von Insekten.
Suche zuerst passende Varianzstruktur und verwende diese dann als Modellannahme für Anzahlen.
Standardannahme bei Anzahlen ist Poisson mit $\mathrm{E}(y) = \mathrm{var}(y)$.

Im Folgenden führen manche Aspekte zur Annahme einer Varianz quadratisch in der Erwartung, $\mathrm{var}(y) = \phi\mu^2$. Für die **Variationskoeffizienten** ergibt sich dann

$$\frac{\sqrt{\mathrm{var}(y_i)}}{\mathrm{E}(y_i)} = \frac{\sqrt{\phi\mu_i^2}}{\mu_i} = \sqrt{\phi}\,,$$

also sind diese **konstant** für alle Beobachtungen.

Generalisiertes Lineares Modell: Quasi-Likelihood

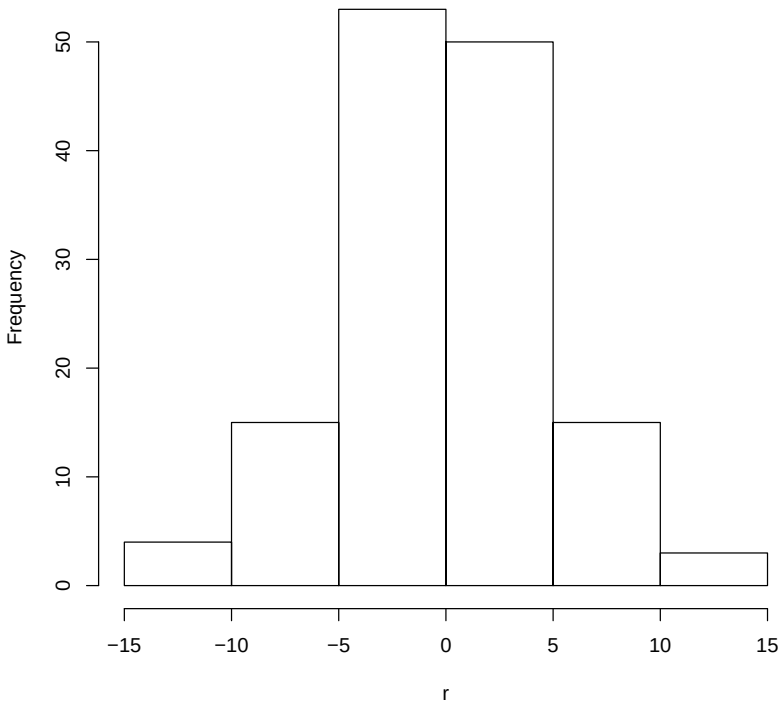
Feldversuch liefert $n = 140$ Insektenanzahlen (`counts`) zweier unabhängiger Wiederholungen (`plots`) in jedem der 10 Blöcke (`block`) für jedes von sieben Verfahren (`treatment`).

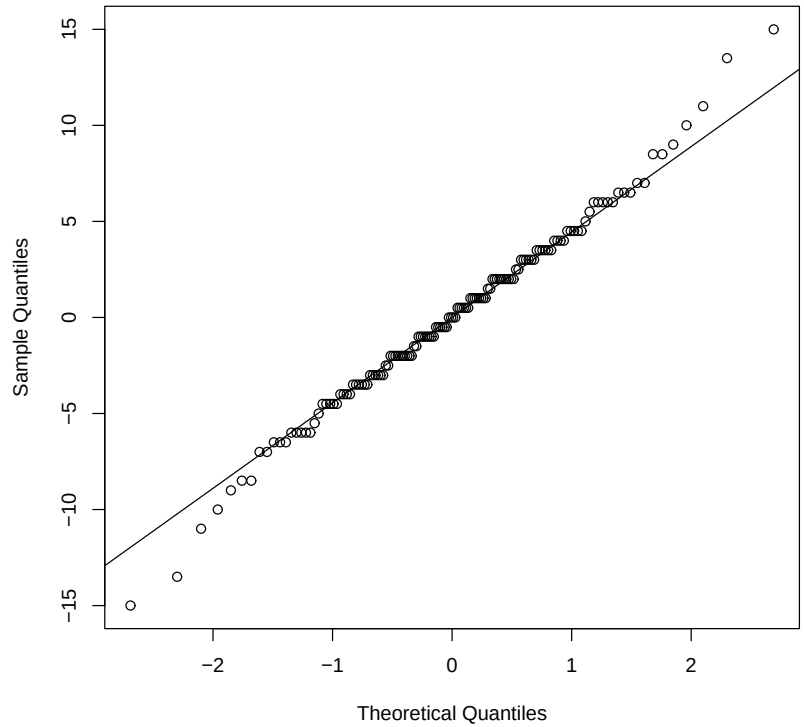
```
> insects <- read.table("insects.dat")
> (block     <- factor(rep(1:10, len=140)))
  [1] 1  2  3  4  5  6  7  8  9  10 1  2  3 ...
Levels: 1 2 3 4 5 6 7 8 9 10
> (plot      <- factor(rep(rep(1:2, each=10), times=7)))
  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 ...
Levels: 1 2
> (treatment <- factor(rep(1:7, each=20)))
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 ...
Levels: 1 2 3 4 5 6 7
> insects<-data.frame(insects,data.frame(treatment,plot,block))
> colnames(insects)<-c("counts","treatment","plot","block")
> attach(insects)
```

Generalisiertes Lineares Modell: Quasi-Likelihood

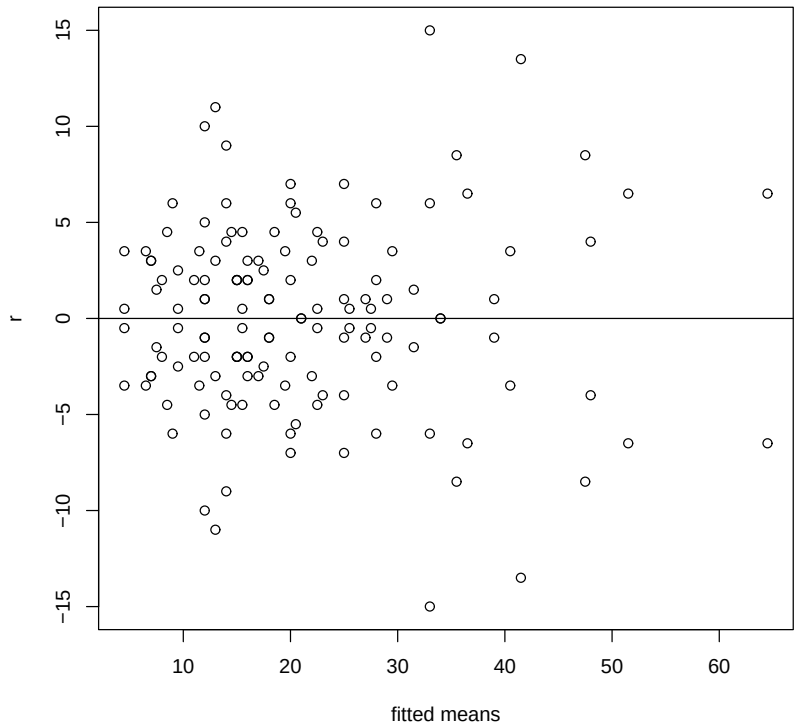
```
> i.lm <- lm(counts ~ treatment*block)      # calc 70 cell means
> r <- residuals(i.lm); f <- fitted(i.lm)   # model diagnostics
> hist(r, main="")
> qqnorm(r, main=""); qqline(r)
> plot(f, r, xlab="fitted means"); abline(0,0)

> (cell.mean <- tapply(counts, list(treatment, block), mean))
     1    2    3    4  5    6    7    8    9   10
1 25.0 39.0 33.0  9.5 16 20.0 28.0 22.0 20.0 29.0
2  7.5 18.5 14.0  8.5  9 16.0 20.5 23.0 18.0 14.5
:

> cell.sd   <- tapply(counts, list(treatment, block), sd)

> plot(cell.mean, cell.sd); abline(mean(cell.sd), 0)
> abline(lsfit(as.vector(cell.mean), as.vector(cell.sd)))
```

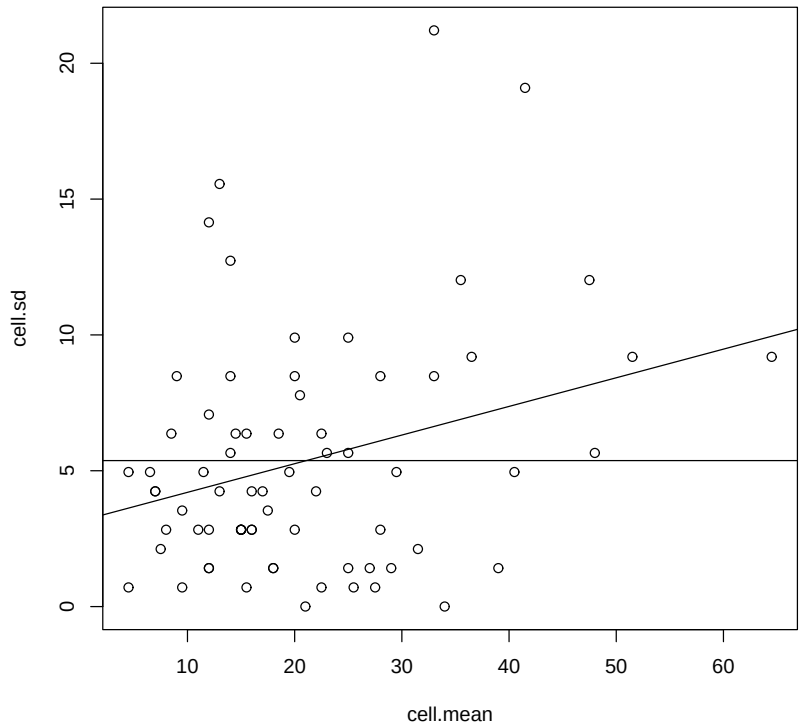
Generalisiertes Lineares Modell: Quasi-Likelihood





Generalisiertes Lineares Modell: Quasi-Likelihood





Generalisiertes Lineares Modell: Quasi-Likelihood

Welche Verteilung liegt den Counts zugrunde?

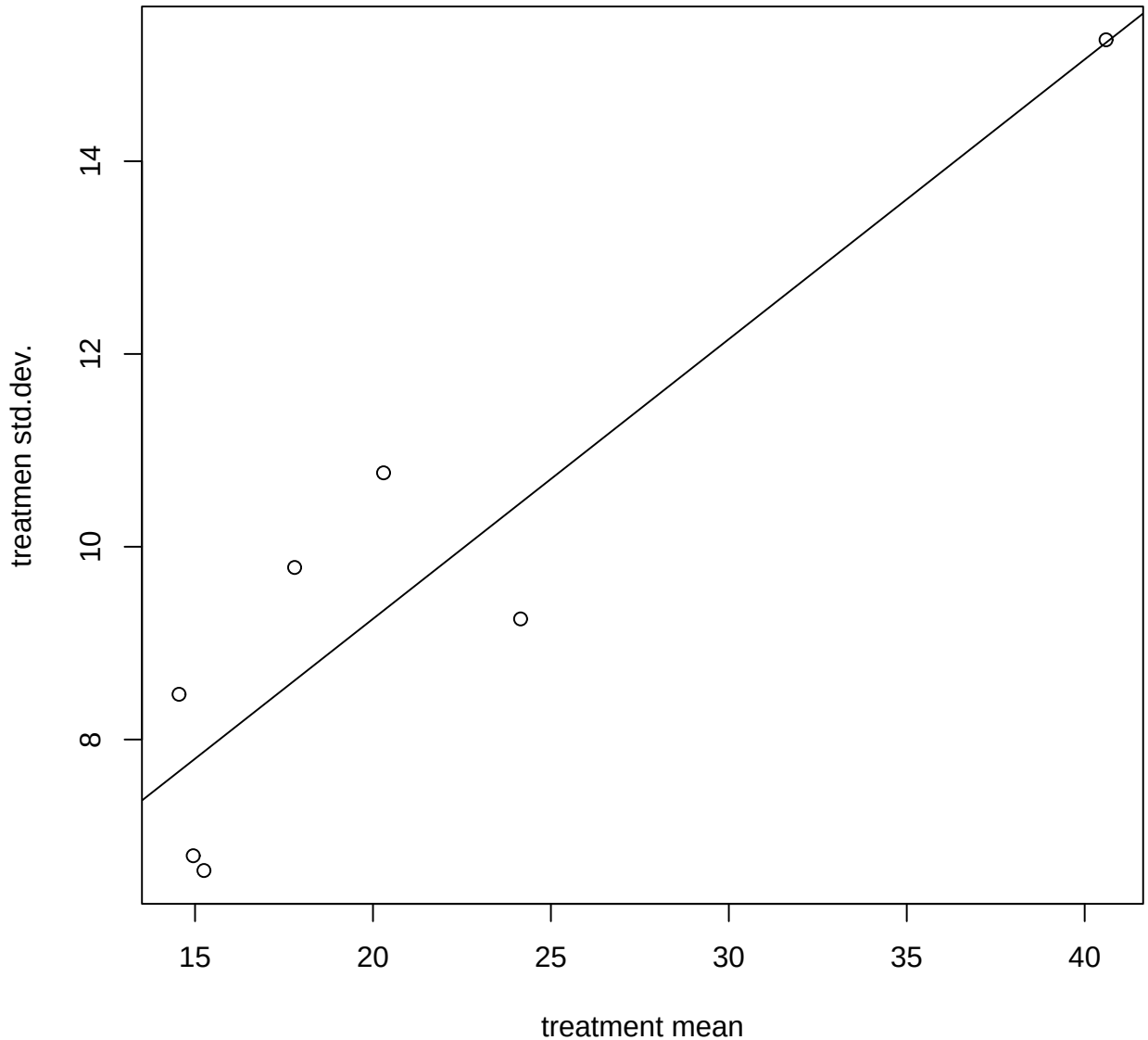
- Betrachte in jeder der 70 Zellen das Mittel beider `counts` und deren Standardabweichung.
- Prüfe dazu die Residuen r eines LM für die `counts`.
- Histogramm der Residuen sieht symmetrisch aus, während QQ-Plot etwas längere Schwänze in der empirischen Verteilung zeigt als bei Normalverteilung der Fall ist.
- Scatterplot Residuen gegen angepassten Werte zeigt leicht zunehmende Dispersion. Da 2 Beobachtungen/Zelle ist Plot symmetrisch um Null.
- Einfacher ist dies im Scatterplot der empirischen Momente zu erkennen.

Generalisiertes Lineares Modell: Quasi-Likelihood

Linearer Zuwachs der Standardabweichungen für wachsende Mittelwerte ist zu sehen, wenn man `treatment`-spezifische Analyse macht.
Standardabweichungen scheinen proportional den Mittelwerten zu sein, was ein quadratisches Modell für die Varianz der `counts` nahe legt, d.h. $\text{var}(y) = \phi\mu^2$ (so wie bei Gammaverteilung).

```
> trt.mean <- tapply(counts, treatment, mean)
> trt.sd   <- tapply(counts, treatment, sd)
> plot(trt.mean, trt.sd); abline(lsfit(trt.mean, trt.sd))
```

Generalisiertes Lineares Modell: Quasi-Likelihood



Generalisiertes Lineares Modell: Quasi-Likelihood

- In Gamma(μ, ϕ)-Verteilung auch Dispersion ϕ, durch die mittlere Pearsonstatistik geschätzt wird.
- Kanonischer Link ist die reziproke Funktion, Log-Links hat aber Vorteile (Interpretation der Parameter).
- Stufen der Prädiktoren sollen derart kodiert werden, dass Abweichungen von Basisstufe (`treatment=1`, `block=1`) beschrieben werden.
- Beginne mit Modell, das jede Zelle durch Parameterkombination eindeutig identifiziert. Interaktion `treatment * block` erlaubt dies und generiert 70 Parameterschätzungen.

Generalisiertes Lineares Modell: Quasi-Likelihood

```
> summary(i.glmmax<-glm(counts ~ treatment * block,
+                       family = Gamma(link=log)))
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.21888    0.29702   10.84   <2e-16 ***
treatment2         -1.20397    0.42004   -2.87   0.0055 **
:
treatment7          0.48243    0.42004    1.15   0.2547
block2              0.44469    0.42004    1.06   0.2934
:
block10             0.14842    0.42004    0.35   0.7249
treatment2:block2   0.45818    0.59403    0.77   0.4431
          :
treatment7:block10  0.09186    0.59403    0.15   0.8776
---
(Dispersion parameter for Gamma family taken to be 0.1764)
    Null deviance: 59.825  on 139  degrees of freedom
Residual deviance: 16.115  on  70  degrees of freedom
AIC: 1033.1
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Dispersion wird nun aus diesem Modell geschätzt und für weitere Modelle festgehalten.

```
> s.glmmax <- summary(i.glmmax)
> (phi <- s.glmmax$dispersion) # mean Pearson statistic
[1] 0.1764366

> df.max <- s.glmmax$df[2]                  # 70
> (dev.max <- s.glmmax$deviance/phi)        # scaled Deviance
[1] 91.33322
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Untersuche, ob auch Submodell ausreicht. Führe dazu *Analysis of Deviance* durch (entspricht ANOVA beim LM).

LRT-Statistik für Modellvergleich: die (durch dieselbe Dispersion) skalierte Deviance-Differenz. Halte dazu Wert $\hat{\phi}$ fest und berechne damit:

```
> s.glmmain <- summary(glm(counts ~ treatment + block,
+                          family = Gamma(link=log)))
> df.main <- s.glmmain$df[2]                # 124
> dev.main <- s.glmmain$deviance/phi        # 143.5044
> 1-pchisq(dev.main-dev.max, df.main-df.max)
[1] 0.545228
```

Generalisiertes Lineares Modell: Quasi-Likelihood

```
> s.glmt <- summary(glm(counts ~ treatment,
+                       family=Gamma(link=log)))
> df.t <- s.glmt$df[2]                        #  133
> dev.t <- s.glmt$deviance/phi                #  235.8278
> 1-pchisq(dev.t-dev.main, df.t-df.main)
[1] 5.551115e-16

> s.glmb <- summary(glm(counts ~ block,
                        family = Gamma(link=log)))
> df.b <- s.glmb$df[2]                        # 130
> dev.b <- s.glmb$deviance/phi                # 248.4376
> 1-pchisq(dev.b-dev.main, df.b-df.main)
[1] 0
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Damit ist es möglich, die *Analysis of Deviance* zu rechnen (hier unter Annahme, $\phi = \hat{\phi}$ bekannt).

	Modell	unskalierte Deviance	skalierte Deviance	df	p-Wert
1)	Haupteffekte und Interaktionen	16.11	91.33	70	
2)	beide Haupteffekte	25.32	143.50	124	
(2-1)	Test auf Interaktionen		52.17	54	0.545
3)	nur Behandlungseffekt	41.61	235.83	133	
(3-2)	Test auf Behandlungseffekt		92.32	9	5.55e-16
4)	nur Blockeffekt	43.83	248.44	130	
(4-2)	Test auf Blockeffekt		104.93	6	0

Interaktion `treatment:block` ist zu vernachlässigen, beide Haupteffekte `block` und `treatment` sind hoch signifikant.

Generalisiertes Lineares Modell: Quasi-Likelihood

Viel einfacher: verwende anova. Liefert Zerlegung des Modells.

```
> anova(glm(counts ~ treatment * block,
+           family = Gamma(link=log)), test="Chisq")

Analysis of Deviance Table
Model: Gamma, link: log
Response: counts

Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            139     59.825
treatment         6  18.2167       133     41.609 < 2.2e-16 ***
block             9  16.2892       124     25.319  5.56e-16 ***
treatment:block  54   9.2049        70     16.115    0.5452
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Möchte man Dispersion frei schätzen, sollte Option `test="F"` verwendet werden.

```
> anova(glm(counts ~ treatment * block,
+           family = Gamma(link=log)), test="F")

Analysis of Deviance Table
Model: Gamma, link: log
Response: counts

Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                            139     59.825
treatment         6  18.2167       133     41.609 17.208 4.1e-12
block             9  16.2892       124     25.319 10.258 6.6e-10
treatment:block  54   9.2049        70     16.115  0.966   0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Bleibe bei Modell mit beiden Haupteffekten (ohne Interaktion).
Aus `summary` Objekt `s.glmmain` erhält man

```
> round(s.glmmain$coefficients, digits=4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9767     0.1365 21.8016   0.0000
treatment2   -0.4746     0.1277 -3.7156   0.0003
treatment3   -0.2297     0.1277 -1.7983   0.0746
treatment4   -0.5285     0.1277 -4.1383   0.0001
treatment5   -0.3439     0.1277 -2.6929   0.0081
treatment6   -0.4729     0.1277 -3.7024   0.0003
treatment7    0.5139     0.1277  4.0239   0.0001
block2        0.3315     0.1527  2.1713   0.0318
block3        0.3381     0.1527  2.2148   0.0286
:
block9        0.4536     0.1527  2.9717   0.0036
block10       0.4293     0.1527  2.8123   0.0057
```

Generalisiertes Lineares Modell: Quasi-Likelihood

Schätzer der `treatment` Erwartungen und (punktweise) 95% Konfidenzintervalle für jeden `block` berechnen.
Vorhersage für alle Stufen von `treatment` in `block 1`.

```
> (new.i <- data.frame(treatment=levels(treatment),
+                      block=factor(rep(1,7))))
  treatment block
1         1     1
:
7         7     1
> (i.pred<-predict(i.glmmain,new=new.i,type="response",se.fit=T)
$fit
    1     2     3     4     5     6     7
19.62 12.21 15.60 11.57 13.91 12.23 32.81
$se.fit
    1     2     3     4     5     6     7
2.679 1.667 2.130 1.579 1.900 1.670 4.480
$residual.scale
[1] 0.4039
```

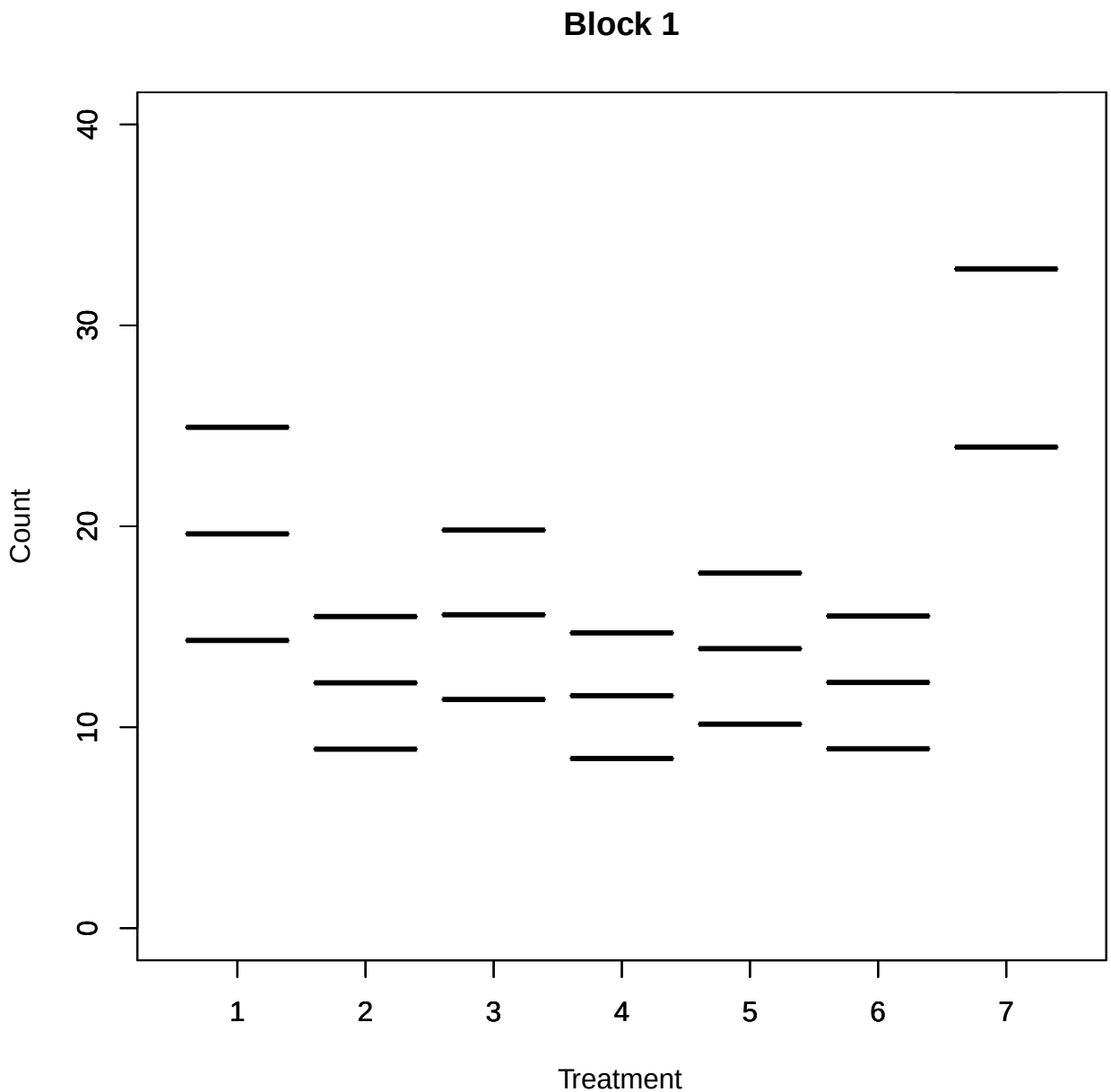
Generalisiertes Lineares Modell: Quasi-Likelihood

```
> fit   <- i.pred$fit
> upper <- fit + qt(0.975, df.main)*i.pred$se.fit
> lower <- fit - qt(0.975, df.main)*i.pred$se.fit

> plot(new.i$treatment, upper, style="box", xlab="Treatment",
+      ylab="Count", ylim=c(0.0,40.0), main="Block 1")
> par(new=TRUE)
> plot(new.i$treatment, fit,   style="box", ylim=c(0.0,40.0))
> plot(new.i$treatment, lower, style="box", ylim=c(0.0,40.0))
```

Ist `se.fit = TRUE` in `predict` gesetzt, wird auch Standardfehler der geschätzten Erwartungen gerechnet. Unter `residual.scale` wird hier die Wurzel von $\hat{\phi}$ verstanden, d.h.

```
> sqrt(s.glmmain$dispersion)
[1] 0.4039
```

Generalisiertes Lineares Modell: Quasi-Likelihood



Logistische Regression

Seien $m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)$, $y_i = 0, \frac{1}{m_i}, \ldots, 1$, wofür gilt

$$\mathrm{E}(y_i) = b'(\theta_i) = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \mu_i$$

$$\mathrm{var}(y_i) = a_i b''(\theta_i) = \frac{1}{m_i} \frac{\exp(\theta_i)}{(1 + \exp(\theta_i))^2} = \frac{1}{m_i} \mu_i (1 - \mu_i)$$

mit **bekannter Dispersion** $\phi = 1$.
Kanonischer Link $g(\mu) = b'^{-1}(\mu) = \theta$ ist der **Logitlink**

$$\text{logit}(\mu) = \log \frac{\mu}{1 - \mu} = \log \frac{m\mu}{m - m\mu} = \theta = \eta \quad \Rightarrow \; \mu = \frac{\exp(\eta)}{1 + \exp(\eta)} \, .$$

Logistische Regression

Name *Logit* bezieht sich auf Verteilung einer logistisch verteilten Zufallsvariablen mit Dichte

$$f(y|\mu,\tau) = \frac{\exp((y-\mu)/\tau)}{\tau\Big(1+\exp((y-\mu)/\tau)\Big)^2}\,, \qquad \mu \in \mathbb{R},\ \tau > 0\,,$$

wofür $\mathrm{E}(y) = \mu$ und $\mathrm{var}(y) = \tau^2\pi^2/3$ gilt.
Bezeichnen X_1 und X_2 zwei unabhängige Exponential(1)-verteilte Zufallsvariablen, so ist die Dichte von $Y = \log(X_1/X_2)$ gleich

$$f(y) = \frac{\exp(y)}{\Big(1+\exp(y)\Big)^2}\,, \qquad \text{sowie} \qquad F(y) = \frac{\exp(y)}{1+\exp(y)}$$

und Verteilungsfunktion $F(y)$ entspricht dem inversen Logitlink.

Logistische Regression: Links

Wegen $\mu = g^{-1}(\eta)$ und $0 < \mu < 1$ kann statt inversem Logitlink beliebig andere Verteilungsfunktion verwendet werden.

Bei $g^{-1}(\eta) = \Phi(\eta)$ spricht man von **Probitmodell**. Logit- und Probitlink sind symmetrische Links.

Extremwertverteilungen:
Maximum–Extremwertverteilung

$$F_{max}(y) = \exp(-\exp(-y)), \qquad y \in \mathbb{R}$$

mit $\mathrm{E}(y) = \gamma$ (Euler Konstante $\gamma = 0.577216$) und $\mathrm{var}(y) = \pi^2/6$. Verwenden wir diese Verteilungsfunktion als inverse Linkfunktion, so resultiert das **log-log Modell** mit $g(\mu) = -\log(-\log(\mu))$.

Logistische Regression: Links

Minimum–Extremwertverteilung
(falls $-y$ Maximum-Extremwertverteilung genügt)

$$F_{min}(y) = 1 - F_{max}(-y) = 1 - \exp(-\exp(y)), \qquad y \in \mathbb{R}$$

mit $\mathrm{E}(y) = -\gamma$ und $\mathrm{var}(y) = \pi^2/6$. Als Linkfunktion erhalten wir das **komplementäre log-log-Modell** $g(\mu) = \log(-\log(1-\mu))$.

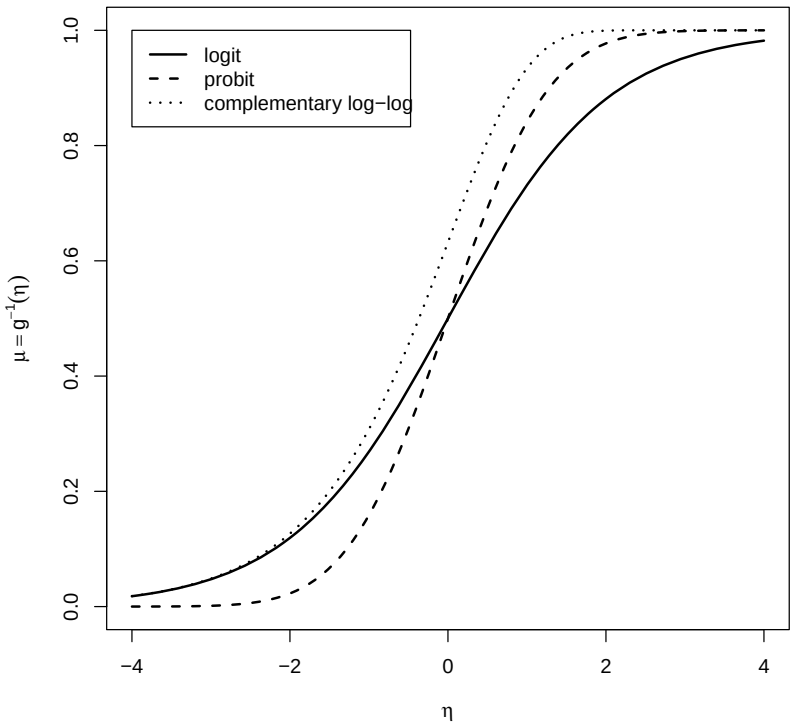
Beide Extremwertverteilungen führen zu asymmetrischen Linkfunktionen.

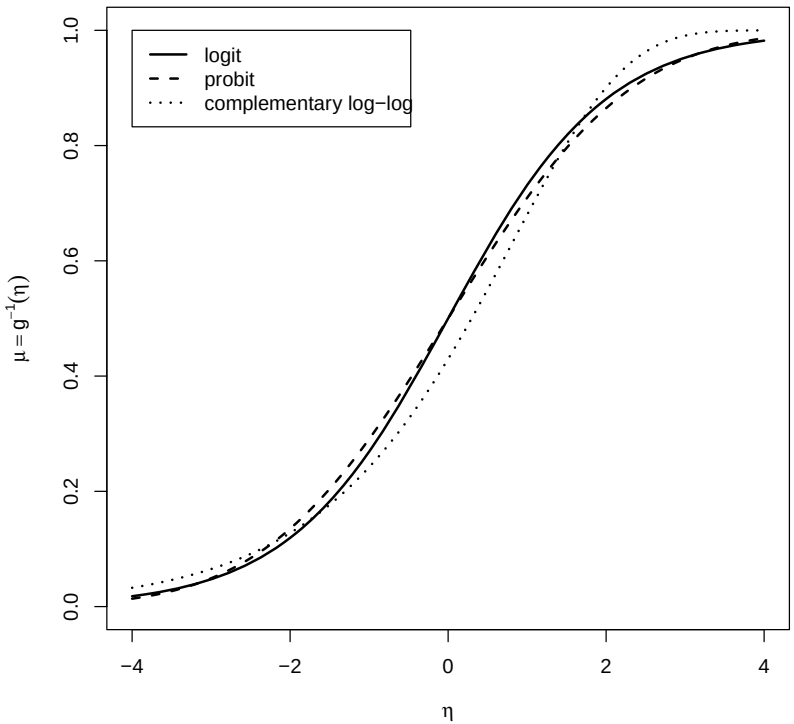
Logistische Regression: Links

R erlaubt für `family=binomial` als Spezifikation der Linkfunktion `logit`, `probit`, `cauchit`, sowie `log` und `cloglog`.

```
> euler <- 0.577216
> mu.logit  <-function(eta) 1/(1 + exp(-eta))
> mu.probit <-function(eta) pnorm(eta, 0, pi/sqrt(3))
> mu.cloglog<-function(eta) 1-exp(-exp(-euler+eta/sqrt(2)))
> plot(mu.logit, (-4): 4, xlim = c(-4, 4), ylim = c(0,1),
+      xlab = expression(eta),
´+      ylab = expression(mu == g^-1 * (eta)), lwd=2)
> curve(mu.probit,  (-4):4, add = TRUE, lty = 2, lwd=2)
> curve(mu.cloglog, (-4):4, add = TRUE, lty = 3, lwd=2)
> legend(-4, 1, c("logit", "probit", "complementary log-log"),
+        lty = 1:3, lwd=2)
```

Logistische Regression: Links





Logistische Regression: Deviance

Log-Likelihoodfunktion und (skalierte) Deviance

$$\ell(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{\mu_i}{1-\mu_i} - m_i \log \frac{1}{1-\mu_i} + \log \binom{m_i}{m_i y_i} \right\},$$

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} m_i \left\{ (1-y_i) \log \frac{1-y_i}{1-\hat{\mu}_i} + y_i \log \frac{y_i}{\hat{\mu}_i} \right\}.$$

Logistische Regression: Deviance

Für binäre Daten $y_i \in \{0, 1\}$ ($m_i = 1$ für alle i) erhalten wir

$$\ell(\mu_i|y_i) = \begin{cases} \log(1-\mu_i) & \text{für } \; y_i = 0\,, \\ \log \mu_i & \text{für } \; y_i = 1 \end{cases}$$

und

$$d(y_i, \hat{\mu}_i) = \begin{cases} -2\log(1-\hat{\mu}_i) & \text{für } \; y_i = 0\,, \\ -2\log \hat{\mu}_i & \text{für } \; y_i = 1\,. \end{cases}$$

Deviance-Inkrement $d(y_i, \hat{\mu}_i)$ beschreibt Anteil einer binären Beobachtung an maximierter Log-Likelihoodfunktion der Stichprobe

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = \sum_{i=1}^{n} \ell(\hat{\mu}_i|y_i) = -\frac{1}{2}\sum_{i=1}^{n} d(y_i, \hat{\mu}_i)\,.$$

Logistische Regression: Toleranzverteilungen

Bioassay: experimentelle Untersuchung basierend auf binäre Responses, z.B. Erprobung der Wirkung diverser Konzentrationen in Tierversuchen.
Anzahl Tiere, die darauf ansprechen, wird als binomiale Response betrachtet.
Beispiel: Insektizid auf Gruppen (**batches**) von Insekten mit bekannten Anzahlen angewendet. Wird einer Gruppe niedrige Dosis verabreicht, wird wahrscheinlich kein Insekt daran sterben. Wird einer anderen Gruppe hohe Dosis verabreicht, werden viele daran sterben.
Ob Insekt bei gegebenen Dosis stirbt oder nicht, hängt von **Toleranz** des Tieres ab. Insekten mit geringer Toleranz werden bei einer Dosis eher sterben als andere mit großer Toleranz.

Logistische Regression: Toleranzverteilungen

Annahme: es gibt eine Verteilung der Toleranz der Insekten.

Insekten mit Toleranz $< d_i$ werden an Dosis d_i sterben. Sei U Zufallsvariable, die mit Toleranz assoziiert ist mit Dichte $f(u)$.

Wahrscheinlichkeit, dass ein Tier bei Dosis d_i stirbt ist

$$p_i = \Pr(U \leq d_i) = \int_{-\infty}^{d_i} f(u)\, du\,.$$

Ist $U \sim \textbf{Normal}(\mu, \sigma^2)$, folgt dafür

$$p_i = \Phi\left(\frac{d_i - \mu}{\sigma}\right)\,.$$

Mit $\beta_0 = -\mu/\sigma$ und $\beta_1 = 1/\sigma$ ergibt dies

$$p_i = \Phi\left(\beta_0 + \beta_1 d_i\right) \quad \text{bzw.} \quad \text{probit}(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 d_i\,,$$

d.h. ein **Probitmodell** für Sterbewahrscheinlichkeit p_i abhängig von verabreichten Dosis d_i.

Logistische Regression: Toleranzverteilungen

Für U **logistisch** verteilt entspricht dies

$$p_i = \Pr(U \leq d_i) = \int_{-\infty}^{d_i} \frac{\exp((u-\mu)/\tau)}{\tau\Big(1+\exp((u-\mu)/\tau)\Big)^2}\, du$$

$$= \frac{\exp((d_i-\mu)/\tau)}{1+\exp((d_i-\mu)/\tau)}\,.$$

Mit $\beta_0 = -\mu/\tau$ und $\beta_1 = 1/\tau$ folgt

$$p_i = \frac{\exp(\beta_0+\beta_1 d_i)}{1+\exp(\beta_0+\beta_1 d_i)} \quad \text{bzw.} \quad \operatorname{logit}(p_i) = \beta_0+\beta_1 d_i\,.$$

Logistisch verteilte Toleranz liefert **logistisches Modell** für p_i.

Logistische Regression: Toleranzverteilungen

Beispiel: Wirkung von Gift auf *Tobacco Budworm*, der Tabakpflanzen schädigt. Gruppen zu je 20 Motten beiderlei Geschlechts (`sex`) wurden verschieden hohe Dosen eines Giftes ausgesetzt und die Anzahl der verstorbenen Tiere gezählt.

	Dosis in μg					
Geschlecht	1	2	4	8	16	32
männlich	1	4	9	13	18	20
weiblich	0	2	6	10	12	16

Logistische Regression: Toleranzverteilungen

Dosen sind 2-er Potenzen. Verwende $\log_2(\text{Dosis})$ als Prädiktor.

```
> (ldose <- rep(0:5, 2))
 [1] 0 1 2 3 4 5 0 1 2 3 4 5

> (sex <- factor(rep(c("M", "F"), c(6, 6))))
 [1] M M M M M M F F F F F F
Levels: F M

> (dead <- c(1,4,9,13,18,20,0,2,6,10,12,16))
 [1]  1  4  9 13 18 20  0  2  6 10 12 16
```

Logistische Regression: Toleranzverteilungen

- Spezifikation binomialer Responses in R mittels Matrix SF (success/failure), deren **erste** (zweite) Spalte Anzahl **Erfolge** (Misserfolge) enthält.
- Modell schätzt dann die **Erfolgswahrscheinlichkeit** (hier die Sterbewahrscheinlichkeit) bei einer Dosis.

```
> (SF <- cbind(dead, alive = 20-dead))
      dead alive
 [1,]    1    19
 [2,]    4    16
   :
[12,]   16     4
```

Logistische Regression: Toleranzverteilungen

```
> summary(budworm.lg <- glm(SF ~ sex*ldose, family = binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104
```

Logistische Regression: Toleranzverteilungen

Alternative Spezifikation durch numerischen Vektor mit Eintragungen s_i/a_i, wobei a_i die Anzahl an Versuchen und s_i die Anzahl der Erfolge bezeichnet. Die a_i mit `weights` spezifizieren.

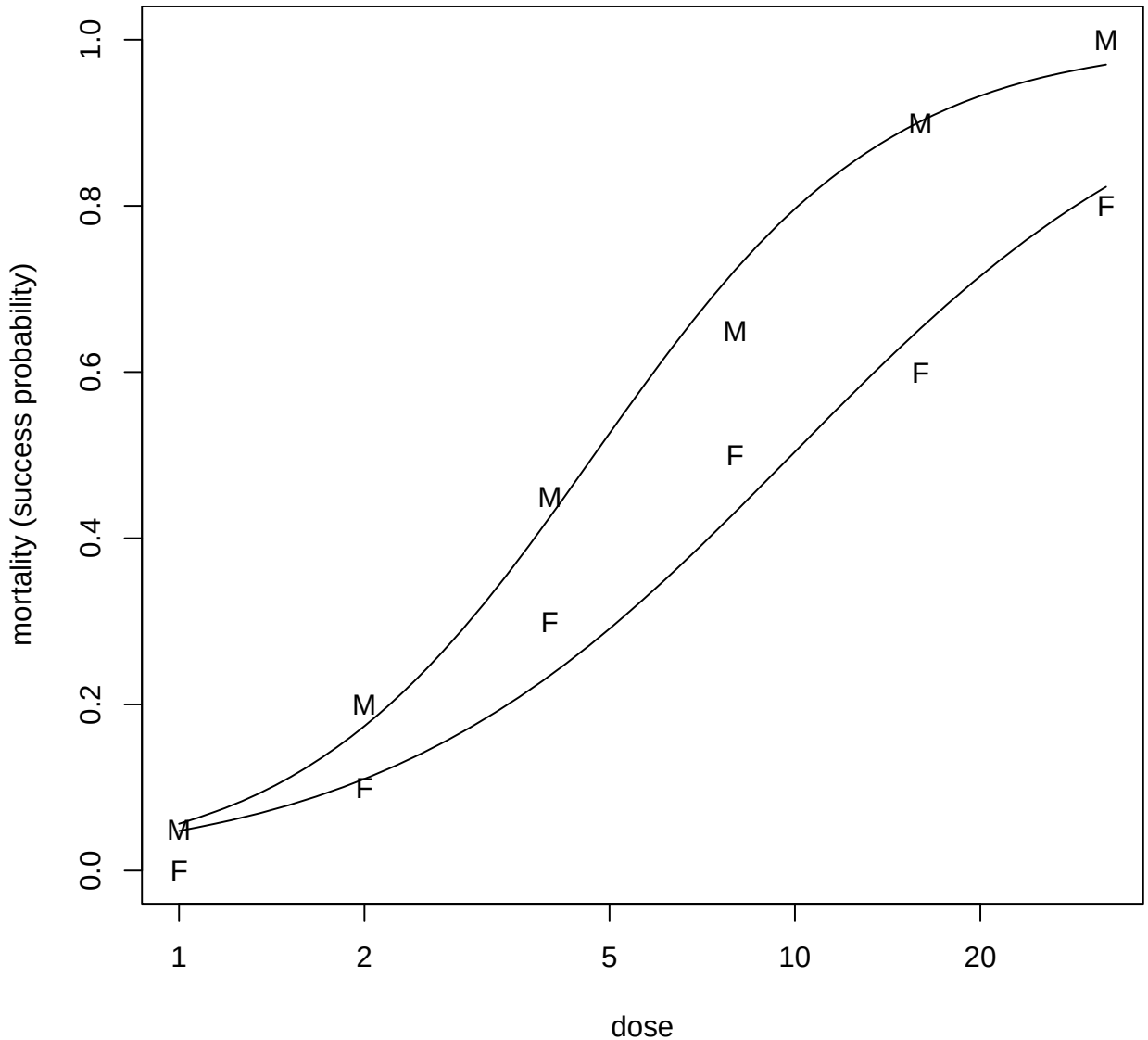
```
> summary(glm(dead/20 ~ sex*ldose, family = binomial,
+             weights=rep(20,12)))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistische Regression: Toleranzverteilungen

Ergebnis weist auf signifikante Steigung von `ldose` hin.
Erste Stufe (Basisstufe) von `sex` sind weibliche Tiere (`"F"` vor `"M"`), wird durch Intercept beschrieben.
Parameter `sexM:ldose` repräsentiert (nicht signifikant) größeren Anstieg bei männlichen Tiere, während `sexM` den (nicht signifikant) Unterschied in den Intercepts beschreibt.

```
> plot(c(1,32), c(0,1), type="n", xlab="dose", log="x")
> text(2^ldose, dead/20, as.character(sex))
> ld <- seq(0, 5, 0.1), l <- length(ld)
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+   sex=factor(rep("M",l,levels=levels(sex))),type="response"))
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+   sex=factor(rep("F",l,levels=levels(sex))),type="response"))
```

Logistische Regression: Toleranzverteilungen



Logistische Regression: Toleranzverteilungen

Parameter `sexM` scheint irrelevant. Er beschreibt den Unterschied bei Dosis 1μg ($\log_2(\text{Dosis}) = 0$). Ist man jedoch am Unterschied bei 8μg ($\log_2(\text{Dosis}) = 3$) interessiert, erhält man dafür

```
> summary(budworm.lg8 <- update(budworm.lg, .~sex*I(ldose-3)))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.2754     0.2305  -1.195  0.23215
sexM                1.2337     0.3770   3.273  0.00107 **
I(ldose - 3)        0.9060     0.1671   5.422 5.89e-08 ***
sexM:I(ldose - 3)   0.3529     0.2700   1.307  0.19117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistische Regression: Toleranzverteilungen

```
> anova(budworm.lg, test = "Chisq")

          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         11    124.876
sex        1    6.077        10    118.799   0.0137 *
ldose      1  112.042         9      6.757   <2e-16 ***
sex:ldose  1    1.763         8      4.994   0.1842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Signifikanter Geschlechtsunterschied bei Dosis 8μg.
Modell passt ausgezeichnet (Deviance 5 bei df $= 8$).
Analysis of Deviance bestätigt es.
Auf Interaktion können wir verzichten.

Logistische Regression: Toleranzverteilungen

Hinzunahme eines quadratischen Terms ist auch nicht notwendig.

```
> anova(update(budworm.lg, .~.+ sex*I(ldose^2)), test="Chisq")

                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                            11    124.876
sex              1    6.077        10    118.799   0.0137 *
ldose            1  112.042         9      6.757   <2e-16 ***
I(ldose^2)       1    0.907         8      5.851   0.3410
sex:ldose        1    1.240         7      4.611   0.2655
sex:I(ldose^2)   1    1.439         6      3.172   0.2303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analyse empfiehlt Modell mit zwei parallelen Geraden bzgl. Prädiktor- (logit)-Achse, eine für jedes Geschlecht.

Logistische Regression: Toleranzverteilungen

Schätzung der Dosis für gewünschte Ausfallswahrscheinlichkeit: reparametrisiere zuerst Modell, so dass jedes Geschlecht eigenen Intercept hat.

```
> summary(budworm.lg0<-glm(SF~sex+ldose-1, family=binomial))

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
sexF   -3.4732     0.4685  -7.413 1.23e-13 ***
sexM   -2.3724     0.3855  -6.154 7.56e-10 ***
ldose   1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 126.2269  on 12  degrees of freedom
Residual deviance:   6.7571  on  9  degrees of freedom
AIC: 42.867
```

Logistische Regression: Toleranzverteilungen

ξ_p ist Wert von $\log_2(\text{Dosis})$ mit Ausfallswahrscheinlichkeit p.
$2^{\xi_{0.5}}$ beschreibt 50% Ausfallsdosis (**50% lethal dose** oder **LD50**).
Dafür gilt bei Linkfunktion $g(p) = \beta_0 + \beta_1 \xi_p$

$$\xi_p = \frac{g(p) - \beta_0}{\beta_1} .$$

Dosis ξ_p hängt von $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ ab, also $\xi_p = \xi_p(\boldsymbol{\beta})$. Ersetzte $\boldsymbol{\beta}$ durch $\hat{\boldsymbol{\beta}}$ liefert Schätzer $\hat{\xi}_p = \xi_p(\hat{\boldsymbol{\beta}})$ wofür approximativ gilt

$$\hat{\xi}_p = \xi_p + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} .$$

Wegen $\mathrm{E}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\beta}$, folgt $\mathrm{E}(\hat{\xi}_p) \approx \xi_p$.

Logistische Regression: Toleranzverteilungen

Mit der Delta-Methode resultiert

$$\mathrm{var}(\hat{\xi}_p) = \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \mathrm{var}(\hat{\boldsymbol{\beta}}) \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} ,$$

wobei

$$\frac{\partial \xi_p}{\partial \beta_0} = -\frac{1}{\beta_1} , \qquad \frac{\partial \xi_p}{\partial \beta_1} = -\frac{g(p) - \beta_0}{\beta_1^2} = -\frac{\xi_p}{\beta_1} .$$

Funktion `dose.p` aus `MASS` und liefert für **weibliche** Tiere:

```
> require(MASS)
> dose.p(budworm.lg0, cf = c(1,3), p = (1:3)/4) # females
            Dose     SE
p = 0.25: 2.231 0.2499
p = 0.50: 3.264 0.2298
p = 0.75: 4.296 0.2747
```

Logistische Regression: Toleranzverteilungen

Für **männliche** Tiere resultiert:

```
> dose.p(budworm.lg0, cf = c(2,3), p = (1:3)/4) # males
            Dose     SE
p = 0.25: 1.197 0.2635
p = 0.50: 2.229 0.2260
p = 0.75: 3.262 0.2550
```

Eine geschätzte Dosis von $\log_2(\text{Dosis}) = 3.26$, also $\text{Dosis} = 9.60$, ist notwendig damit 50% der weiblichen Schädlinge ausfallen, aber nur eine $\text{Dosis} = 4.69$ für 50% der männlichen.

Logistische Regression: Toleranzverteilungen

Alternatives Probitmodell: liefert sehr ähnliche Resultate. So ergibt sich beispielsweise für **weibliche** Motten

```
> dose.p(update(budworm.lg0, family=binomial(link=probit)),
+        cf=c(1,3), p=(1:3)/4)
            Dose     SE
p = 0.25: 2.191 0.2384
p = 0.50: 3.258 0.2241
p = 0.75: 4.324 0.2669
```

Logistische Regression: Parameterinterpretation

Erwartungswert binärer Daten hängt vom zweistufigen Faktor x ab. Zellwahrscheinlichkeiten:

	$x = 1$	$x = 0$
$y = 1$	π_1	π_0
$y = 0$	$1 - \pi_1$	$1 - \pi_0$

Für $x = 1$ bezeichnet $\pi_1/(1 - \pi_1)$ die **Quote, Chance (odds)** für das Eintreten von $y = 1$ zu $y = 0$.
Deren Log-Transformation

$$\log \frac{\pi_1}{1 - \pi_1} = \text{logit}(\pi_1)$$

nennt man **log-odds** oder **Logit**.

Logistische Regression: Parameterinterpretation

Den Quotienten der Quote unter $x = 1$ zur Quote unter $x = 0$ bezeichnet man als **odds-ratio**

$$\psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)},$$

und dessen Log-Transformation ergibt das **log-odds ratio** oder die **Logit-Differenz**

$$\log \psi = \log \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \text{logit}(\pi_1) - \text{logit}(\pi_0).$$

Logistische Regression: Parameterinterpretation

Sei $\mu(x) = \Pr(y = 1|x)$ und $1 - \mu(x) = \Pr(y = 0|x)$, $x \in \{0, 1\}$.
Das Modell

$$\log \frac{\mu(x)}{1 - \mu(x)} = \beta_0 + \beta_1 x$$

liefert folgende Wahrscheinlichkeiten:

	$x = 1$	$x = 0$
$y = 1$	$\mu(1) = \dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\mu(0) = \dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
$y = 0$	$1 - \mu(1) = \dfrac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \mu(0) = \dfrac{1}{1 + \exp(\beta_0)}$

Als log-odds ratio folgt

$$\log \psi = \log \frac{\mu(1)/(1 - \mu(1))}{\mu(0)/(1 - \mu(0))} = \log \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \beta_1 \,.$$

Logistische Regression: Parameterinterpretation

Ist x ein allgemeiner Prädiktor und hält ein entsprechendes Modell, dann folgt dafür als Quote

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = \frac{\mu(x)}{1-\mu(x)} = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0)\exp(\beta_1)^x .$$

Interpretation: wächst der Prädiktor x um eine Einheit, so erhöht sich die Eintrittsquote von $y = 1$ multiplikativ um den Term $\exp(\beta_1)$.

Logistische Regression: Parameterinterpretation

Beispiel: Injektionsbehandlung bei 27 Krebspatienten sollte zum Abklingen des Karzinoms führen.
Wichtigste erklärende Variable LI (Labelling Index) beschreibt Zellteilungsaktivität nach erfolgter Behandlung.
Für $n = 14$ unterschiedliche LI Werte beschreibt $m_i y_i$ die Anzahl erfolgreicher Rückbildungen bei m_i Patienten:

$\mathtt{LI}_i$	m_i	$m_i y_i$	$\mathtt{LI}_i$	m_i	$m_i y_i$	$\mathtt{LI}_i$	m_i	$m_i y_i$
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

Logistische Regression: Parameterinterpretation

Annahme: m_i Patienten in der LI_i Gruppe sind homogen, d.h.

$$m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)\,, \qquad \text{mit} \qquad \log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 \text{LI}_i\,.$$

```
> li <- c(seq(8, 28, 2), 32, 34, 38)
> total <-c(2, 2, 3, 3, 3, 1, 3, 2, 1, 1, 1, 1, 1, 3)
> back  <-c(0, 0, 0, 0, 0, 1, 2, 1, 0, 1, 1, 0, 1, 2)
> SF <- cbind(back, nonback = total - back)
> summary(carcinoma <- glm(SF ~ li, family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
AIC: 24.29
```

Logistische Regression: Parameterinterpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
```

Interpretation:

- Nimmt LI um eine Einheit zu, wird die Rückbildungsquote mit $\exp(0.145) = 1.156$ multipliziert (Zunahme um 15.6%).
- Wahrscheinlichkeit für eine Rückbildung ist $1/2$, falls $\hat{\eta} = 0$, also für $\text{LI} = -\hat{\beta}_0/\hat{\beta}_1 = 26.07$.
- Für das Mittel aller 27 LI-Werte, $\sum_i \text{LI}_i m_i / \sum_i m_i = 20.07$, ist linearer Prädiktor $\hat{\beta}_0 + \hat{\beta}_1 20.07 = -0.8691$ (entspricht 29.54%). Beobachtet sind 9 Erfolge aus 27 Patienten, also 33.33%.

Logistische Regression: Parameterinterpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
```

Interpretation:

- Logistische Regressionskurve: $\partial\mu(x)/\partial x = \beta_1\mu(x)(1-\mu(x))$. Stärkster Anstieg in $\mu(x) = 1/2$, also bei $\texttt{LI} = 26.07$, und beträgt dort $\hat{\beta}_1/4 = 0.0362$.

- Frage: hängt Rückbildung signifikant vom LI-Wert ab? Der p-Wert beim Wald-Test von 1.46% bestätigt dies.

Logistische Regression: Parameterinterpretation

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
AIC: 24.29
```

Interpretation:

• Für das iid Zufallsstichprobenmodell resultiert als (NULL) Deviance 23.96 bei $df = 13$. Deviancedifferenz 8.30 bei Verlust von $df = 1$ entspricht $\chi^2_{1;1-\alpha}$-Quantil mit $\alpha = 0.004$ (noch deutlicher als Wald-Test).

Signifikante (positive) Assoziation zwischen `LI` und Rückbildung.

Logistische Regression: Parameterinterpretation

Einfacher mit:

```
> anova(carcinoma, test="Chisq")

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    13      23.96
li    1    8.299        12      15.66  0.00397 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistische Regression: Parameterinterpretation

Modell mit einzelne Patienten als Bernoullivariablen: liefert dieselben Parameterschätzer, aber andere Werte für Deviance und Freiheitsgrade.

```
> index <- rep.int(li, times=total)
> B<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(carcinomaB <- glm(B ~ index, family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
index         0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom
AIC: 30.07
```

Logistische Regression: Parameterinterpretation

Jedoch ist die Deviancedifferenz dieselbe wie zuvor:

```
> anova(carcinomaB, test="Chisq")

      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     26      34.37
index  1    8.299        25      26.07  0.00397 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bemerke, dass wieder Wahrscheinlichkeit für $y = 1$ (Rückbildung) modelliert wird.

Da bei Bernoullivariablen alle $m_i = 1$ sind, ist deren explizite Angabe durch `weights` nicht notwendig.

Logistische Regression: Logitmodell

Spezialfall der logistischen Regression nur mit Faktoren aber ohne Variablen im linearen Prädiktor!

Beispiel: Untersuchungen von 313 Frauen nach Operation des Zervixkarzinoms ergaben 123 Rezidive.
Von welchen Risikogrößen hängt der Rezidiveintritt ab?
Anzahl befallener Lymphknotenstationen (LK) und Befall der Zervix-Grenzzone (GZ) stellen potentielle Risikogrößen dar.
Rezidivfälle (y_{ij}/m_{ij}) als 3×4 Kontingenztafel:

	befallene LK-Stationen			
	0	1	2	≥ 3
GZ nicht befallen	21/124	7/21	9/16	13/13
GZ befallen	18/ 58	6/12	5/ 7	5/ 5
über GZ befallen	4/ 14	16/19	9/12	10/12

Logistische Regression: Logitmodell

Verhalten sich Risikogrößen **linear**, kann LK z.B. mit $0, 1, 2, 3$ und GZ mit $0, 1, 2$ kodiert werden und wir betrachten das Modell

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{LK}_i + \beta_2 \text{GZ}_i\,, \qquad i = 1, \ldots, n = 12\,.$$

```
> rez   <- c( 21, 7, 9,13,18, 6,5,5, 4,16, 9,10)
> total <- c(124,21,16,13,58,12,7,5,14,19,12,12)
> LK <- rep(0:3, 3);  GZ <- rep(0:2, each=4)
> SF <- cbind(nonrez=total-rez, rez)
> summary(rez.glm <- glm(SF ~ LK + GZ, family = binomial))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.551      0.198    7.82  5.4e-15 ***
LK            -1.069      0.154   -6.95  3.6e-12 ***
GZ            -0.586      0.178   -3.29    0.001 **
---
    Null deviance: 103.493  on 11  degrees of freedom
Residual deviance:  13.332  on  9  degrees of freedom
AIC: 51.21
```

Logistische Regression: Logitmodell

Ist Linearität in den Prädiktoren nicht gewährleistet, verwendet man jeweils besser Faktoren $\mathtt{L}$ und $\mathtt{G}$.
Seien $m_{ij}y_{ij} \sim \text{Binomial}(m_{ij}, \mu_{ij})$, wobei μ_{ij} die Wahrscheinlichkeit rezidivfrei zu sein bei $\mathtt{LK}$-Status i $(i = 1, 2, 3, 4)$ und $\mathtt{GZ}$-Befall j $(j = 1, 2, 3)$ beschreibt.
Unterscheide folgende Modelle

$$
\begin{aligned}
\text{logit}(\mu_{ij}) &= (\mathtt{L} * \mathtt{G})_{ij} \,, \\
\text{logit}(\mu_{ij}) &= \mathtt{L}_i + \mathtt{G}_j \,, \\
\text{logit}(\mu_{ij}) &= \mathtt{L}_i \,, \\
\text{logit}(\mu_{ij}) &= \mathtt{G}_j \,, \\
\text{logit}(\mu_{ij}) &= 1 \,.
\end{aligned}
$$

Das Modell $\mathtt{L} * \mathtt{G}$ beinhaltet die gesamte Information der Daten und ist daher voll (saturiert), d.h. es erlaubt 12 Parameter bei 12 Beobachtungen.

Logistische Regression: Logitmodell

```
> L <- factor(LK); G <- factor(GZ)
> rez.1   <- glm(SF ~ 1,     family=binomial)
> rez.L   <- glm(SF ~ L,     family=binomial)
> rez.G   <- glm(SF ~ G,     family=binomial)
> rez.LG  <- glm(SF ~ L + G, family=binomial)
> rez.sat <- glm(SF ~ L * G, family=binomial)

> anova(rez.1, rez.L, rez.LG, rez.sat, test="Chisq")
Model 1: SF ~ 1
Model 2: SF ~ L
Model 3: SF ~ L + G
Model 4: SF ~ L * G
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        11    103.493
2         8     21.373  3   82.120 < 2.2e-16 ***
3         6     10.798  2   10.575  0.005053 **
4         0      0.000  6   10.798  0.094825 .
```

Logistische Regression: Logitmodell

Interaktion zwischen L und G scheint nicht relevant zu sein!
Minimal notwendige Modell ist $\texttt{L} + \texttt{G}$. Dafür resultieren MLE

```
> rez.LG$coefficients
(Intercept)          L1          L2          L3          G1          G2
      1.604      -1.287      -1.779      -3.798      -0.729      -1.074
```

Vergleich zu vorhin:
der Slope -1.069 zu LK ist etwa vergleichbar mit -1.287,
$2 * -1.069 = -2.138$ überschätzt die L2 Stufe und
$3 * -1.069 = -3.207$ unterschätzt die L3 Stufe.

Diese Erkenntnis motiviert, nie künstlich linearisierte Prädiktoren sondern Faktoren mit freien Stufen in das Modell aufzunehmen.

Logistische Regression: Logitmodell

Geschätztes Modell für den linearen Prädiktor mit L und G ist

$$\hat{\eta} = 1.604 - 1.287(\mathtt{L} = 1) - 1.779(\mathtt{L} = 2) - 3.798(\mathtt{L} = 3)$$
$$-0.729(\mathtt{G} = 1) - 1.074(\mathtt{G} = 2)\,.$$

Parameter sind alle negativ und nehmen zu höheren Faktorstufen monoton ab. D.h., Wahrscheinlichkeit rezidivfrei zu sein nimmt ab je mehr Lymphknoten befallen sind und je intensiver der Grenzzonenbefall ausgeprägt ist.

Geschätzte Wahrscheinlichkeiten für Rezidivfreiheit:

```
> p <- predict(rez.LG, expand.grid(L=levels(L), G=levels(G)),
+              type="response")
> matrix(p, ncol=4, byrow = TRUE)
      [,1]  [,2]  [,3]   [,4]
[1,] 0.833 0.579 0.457 0.1003
[2,] 0.706 0.399 0.288 0.0511
[3,] 0.630 0.319 0.223 0.0367
```

Logistische Regression: Logitmodell

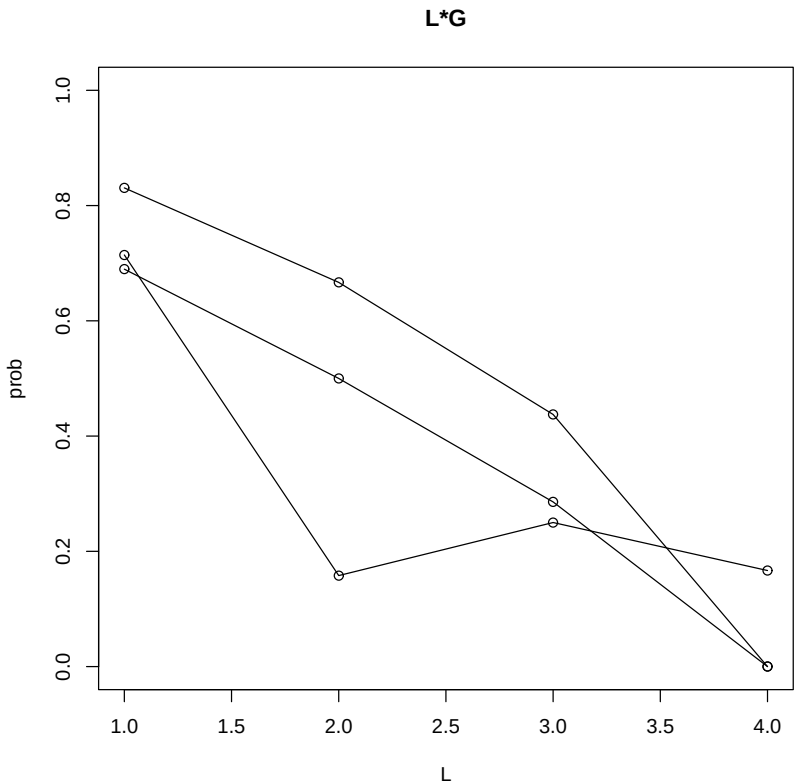
Volles Modell reproduziert relative Häufigkeiten in den Daten:

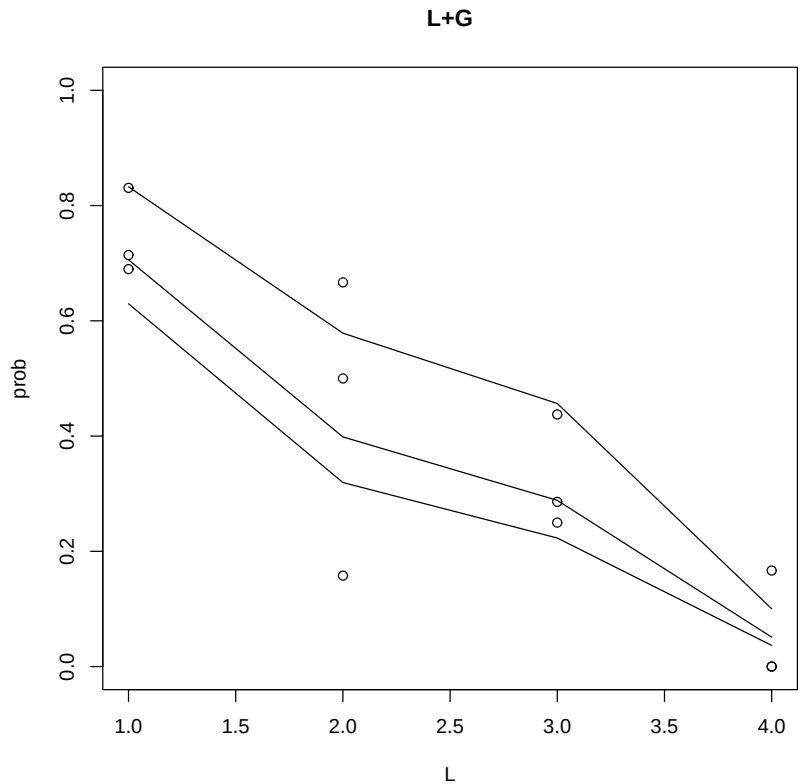
```
> p <- predict(rez.sat, expand.grid(L=levels(L), G=levels(G)),
+              type="response")
> matrix(p, ncol=4, byrow = TRUE)
      [,1]  [,2]  [,3]     [,4]
[1,] 0.831 0.667 0.438 9.752e-12
[2,] 0.690 0.500 0.286 2.208e-11
[3,] 0.714 0.158 0.250 0.167
```

Grafische Gegenüberstellung sämtlicher geschätzten Modelle:

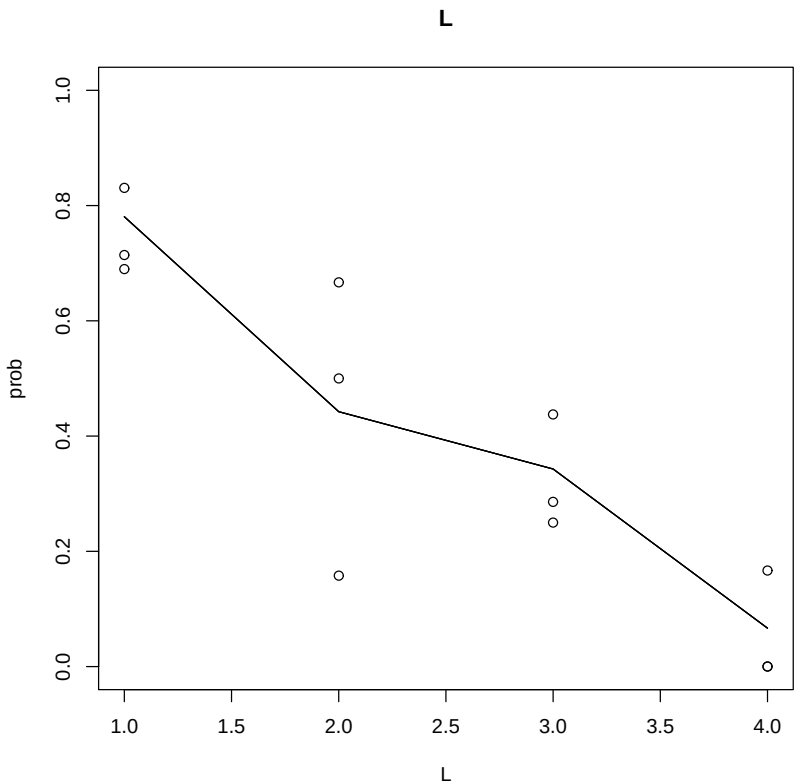
```
> plot(as.numeric(L), rez.LG$y, type="p", xlab="L", ylab="prob",
      ylim=c(0,1), main="L+G")
> lines(as.numeric(L)[G==0], fitted(rez.LG)[G==0], type="l")
> lines(as.numeric(L)[G==1], fitted(rez.LG)[G==1], type="l")
> lines(as.numeric(L)[G==2], fitted(rez.LG)[G==2], type="l")
```

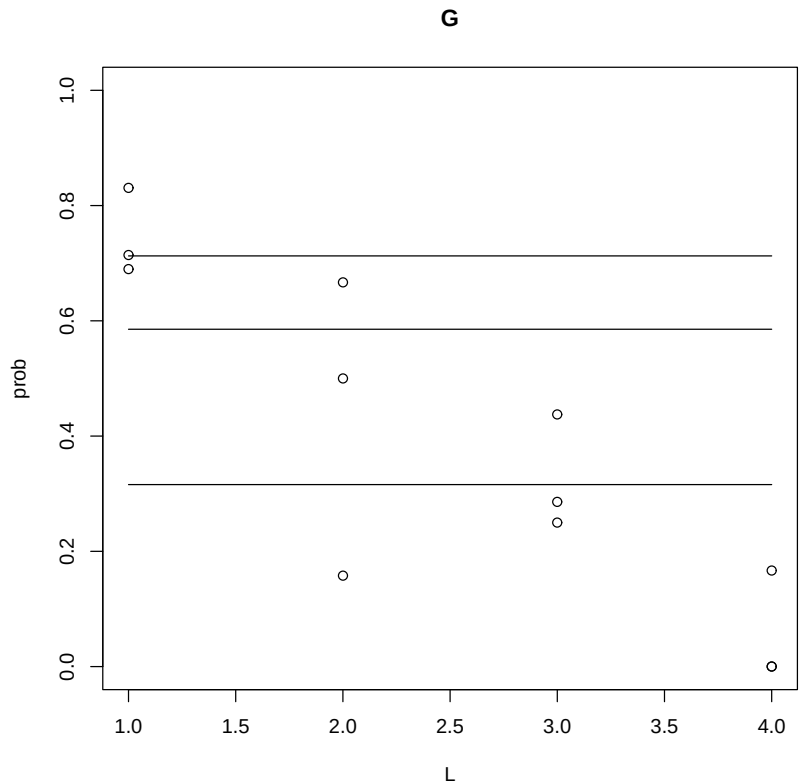
Logistische Regression: Logitmodell



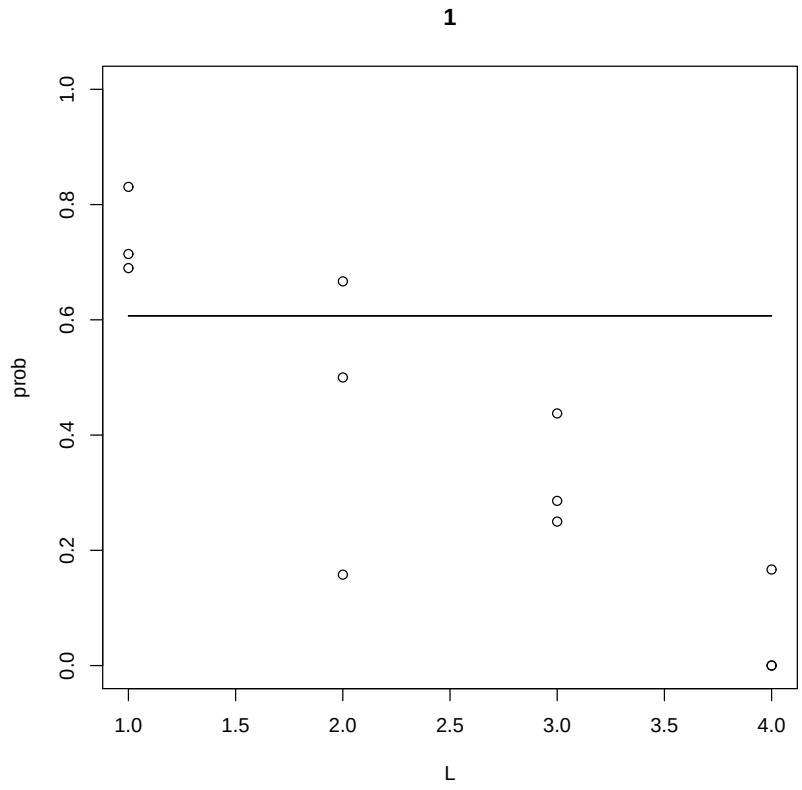


Logistische Regression: Logitmodell





Logistische Regression: Logitmodell



Logistische Regression: Überdispersion

Überdispersion: Varianz der Responses übersteigt nominale Varianz unter dem Modell. Seltener ist **Unterdispersion**.

Wir modellieren absolute Häufigkeiten und vergleichen daher Responsevarianz mit binomialer Varianz.

Wie erkennt man Überdispersion?
Falls Modell korrekt, ist Deviance asymptotisch χ^2_{n-p}-verteilt.
$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) > n - p = \mathrm{E}(\chi^2_{n-p})$ kann Hinweis auf Überdispersion sein.

Logistische Regression: Überdispersion

Andererseits kann $D(\mathbf{y}, \hat{\boldsymbol{\mu}}) > n - p$ auch alternativ begründet sein:

- fehlende Prädiktoren und/oder Interaktionsterme,
- Vernachlässigung nichtlinearer Effekte,
- falsche Linkfunktion,
- extreme Ausreißer,
- binäre Daten oder kleine Werte von m_i.

Schließe zuerst sämtliche Gründe mittels explorativer Datenanalyse und diagnostischer Verfahren aus.

Gründe für Überdispersion: Überdispersion kann verursacht werden durch Variation unter den Erfolgswahrscheinlichkeiten oder durch Korrelation unter den binären Responsevariablen.

Logistische Regression: Überdispersion

Zuerst eine recht allgemeine Aussage über die mögliche Varianzstruktur einer auf $[0, 1]$ definierten Zufallsvariablen.

Lemma

Es existiert genau dann eine auf $[0, 1]$ *verteilte Zufallsvariable* P *mit vorgegebenen Momenten* $E(P) = \pi$ *und* $var(P) = \sigma^2$, *wenn* $0 \leq \pi \leq 1$ *und* $0 \leq \sigma^2 \leq \pi(1 - \pi)$ *gilt.*

Beweis: Generell gelte $0 \leq \sigma^2$. Da $0 \leq P \leq 1$, folgt weiters $0 \leq \pi \leq 1$. Nun gilt $P^2 \leq P$ auf $[0, 1]$, woraus $\mathrm{E}(P^2) \leq \mathrm{E}(P)$ hervorgeht. Deswegen folgt

$$\sigma^2 = \mathrm{var}(P) = \mathrm{E}(P^2) - \mathrm{E}^2(P) \leq \pi - \pi^2 = \pi(1 - \pi)\,. \quad \square$$

Daher hat unter allen auf $[0, 1]$ definierten Zufallsvariablen die **Bernoulli-Variable maximale Varianz**.

Logistische Regression: Überdispersion

Lemma gilt auch für stand. binomialverteilte ZV y/m mit $\mathrm{E}(y/m) = \pi$. Dafür folgt damit $0 \leq \mathrm{var}(y/m) \leq \pi(1-\pi)$.

Für korrespondierende binomialverteilte ZV y mit $\mathrm{E}(y) = m\pi$ gilt für alle ganzzahligen $0 \leq y \leq m$ die Abschätzung $\mathrm{E}(y^2) \geq \mathrm{E}(y)$, also $\mathrm{var}(y) \geq \mathrm{E}(y) - \mathrm{E}^2(y) = \mathrm{E}(y)(1 - \mathrm{E}(y))$. Zusammen folgt

$$\max(0, m\pi(1-m\pi)) \leq \mathrm{var}(y) \leq m^2\pi(1-\pi)\,,$$

was für $m = 1$ (Bernoulli-Variable) zur Gleichung wird. Mit $\mathrm{E}(y) = m\pi = \mu$ resultiert dafür

$$\max(0, \mu(1-\mu)) \leq \mathrm{var}(y) \leq \mu(m-\mu)\,.$$

Logistische Regression: Überdispersion

Verwenden wir die für Dispersionsmodelle (relativ zum Binomialmodell) gebräuchliche Annahme $\text{var}(y) = \phi m\pi(1-\pi)$, so folgt für den Dispersionsparameter ϕ die Restriktion

$$\max\left(0, \frac{1-m\pi}{1-\pi}\right) \leq \phi \leq m\,.$$

Für unabhängige Stichprobenelemente $y_1, \ldots, y_n$ mit Varianzen $\text{var}(y_i) = \phi m_i \pi_i(1-\pi_i)$ folgt als Restriktion

$$\max_{i=1,\ldots,n}\left(0, \frac{1-m_i\pi_i}{1-\pi_i}\right) \leq \phi \leq \min_{i=1,\ldots,n}(m_i)\,.$$

Dispersion ϕ muss somit kleiner gleich allen Umfängen m_i sein.

Logistische Regression: Überdispersion

Die folgende Überlegung liefert eine plausible Motivation für die Notwendigkeit eines Dispersionsparameters im Verteilungsmodell.

Annahme: in der i-ten Experimentierumgebung werden m_i binäre $y_{i1}, \ldots, y_{im_i}$ beobachtet, alle mit gleichem $\mathrm{E}(y_{ij}) = \pi_i$, also auch mit $\mathrm{var}(y_{ij}) = \pi_i(1 - \pi_i)$.

Manchmal ist es naheliegend, in der Gruppe für alle Paare (y_{ij}, y_{ik}) denselben Korrelationskoeffizienten ρ anzunehmen mit

$$\rho = \frac{\mathrm{E}(y_{ij}y_{ik}) - \mathrm{E}(y_{ij})\mathrm{E}(y_{ik})}{\sqrt{\mathrm{var}(y_{ij})\,\mathrm{var}(y_{ik})}} = \frac{\Pr(y_{ij} = 1, y_{ik} = 1) - \pi_i^2}{\pi_i(1 - \pi_i)}\,, \quad j \neq k\,.$$

Logistische Regression: Überdispersion

Für die Summe $y_i = \sum_{j=1}^{m_i} y_{ij}$ folgt damit

$$\begin{aligned}
\mathrm{E}(y_i) &= m_i \pi_i = \mu_i \\
\mathrm{var}(y_i) &= \sum_{j=1}^{m_i} \mathrm{var}(y_{ij}) + \sum_{j \neq k}^{m_i} \mathrm{cov}(y_{ij}, y_{ik}) \\
&= m_i \pi_i (1 - \pi_i) + m_i (m_i - 1) \pi_i (1 - \pi_i) \rho \\
&= m_i \pi_i (1 - \pi_i) \big(1 + (m_i - 1)\rho\big) = \phi_i m_i \pi_i (1 - \pi_i) \,.
\end{aligned}$$

Varianz der Summe korrelierter Bernoullis unterscheidet sich von Varianz der Binomialverteilung um Faktor $\phi_i = \big(1 + (m_i - 1)\rho\big)$.

Positive Werte von ρ ergeben Überdispersion.
Wegen $\phi_i \geq 0$ ergibt sich $\rho \geq -1/(m_i - 1)$.
Ist $m_i = 1$, dann gibt es keine Dispersion, da hierfür $\phi_i = 1$ folgt.

Logistische Regression: Überdispersion

Summe homogener Bernoulli-Variablen ist binomialverteilt. Wie ist Summe heterogener Bernoulli-Variablen verteilt?

Lemma

Sei $y = \sum_{j=1}^{m} y_j$ *mit* $y_j \overset{ind}{\sim}$ *Binomial*$(1, \pi_j)$ *und* $\pi_1, \ldots, \pi_m$ *fest. Für die ersten beiden Momente von* y *gilt mit* $\overline{\pi} = \frac{1}{m}\sum_j \pi_j$ *und* $s_\pi^2 = \frac{1}{m-1}\sum_j (\pi_j - \overline{\pi})^2$

$$\kappa_1 = m\overline{\pi} \quad \textit{bzw.} \qquad \kappa_2 = m\overline{\pi}(1-\overline{\pi}) - (m-1)s_\pi^2 \,.$$

Beweis: Aus der Eigenschaft der Additivität der Kumulanten folgt $\kappa_1 = \sum_j \pi_j = m\overline{\pi}$ und $\kappa_2 = \sum_j \pi_j(1-\pi_j) = m\overline{\pi} - \sum_j \pi_j^2$. Da $\sum_j (\pi_j - \overline{\pi})^2 = \sum_j \pi_j^2 - m\overline{\pi}^2$ folgt $\sum_j \pi_j^2 = (m-1)s_\pi^2 + m\overline{\pi}^2$ und damit κ_2. □

Logistische Regression: Überdispersion

Die Varianz von y ist verglichen mit binomialer Varianz sogar um $(m-1)s_\pi^2$ kleiner.

Intuitiv hätte man mit größerer Varianz gerechnet.

In der Praxis ist meist nur bekannt, dass Variabilität in den π_j's besteht, aber die exakten Werte $p_1, \ldots, p_m$ dieser Parameter sind unbekannt.

Logistische Regression: Überdispersion

Lemma

Seien P_j stetige, auf $[0,1]$ iid Zufallsvariablen mit Dichte f_P und $E(P_j) = \pi$ und seien $y_j|(P_j = p_j) \overset{ind}{\sim} Binomial(1, p_j)$. Als Verteilung von y_j folgt unabhängig von f_P, $y_j \overset{ind}{\sim} Binomial(1, \pi)$ also $y = \sum_{j=1}^{m} y_j \sim Binomial(m, \pi)$.

Beweis: Wegen $y_j|(P_j = p_j) \overset{ind}{\sim} \text{Binomial}(1, p_j)$, folgt

$$\Pr(y_j = 1) = \int \Pr(y_j = 1|P_j = p) f_P(p) dp = \int p f_P(p) dp = \mathrm{E}(P_j) = \pi \,.$$

Analog ist $\Pr(y_j = 0) = 1 - \pi$ und somit $y_j \overset{ind}{\sim} \text{Binomial}(1, \pi)$. □

Logistische Regression: Überdispersion

Beta-Binomiale Varianz

Trotz Einbeziehung von Variabilität in den P_j's erhält man wegen identischer Erwartungswerte wieder die binomiale Varianz.
Nehmen wir Beta-Verteilung für latenten Variablen an, so folgt:

Lemma

Seien $P_1, \ldots, P_n$ unabhängig (nicht identisch) Beta-verteilt mit

$$f_{P_i}(p) = \frac{1}{B(a_i, b_i)} p^{a_i - 1}(1-p)^{b_i - 1}\,, \quad 0 < p < 1\,, \quad a_i, b_i > 0\,,$$

$$E(P_i) = \frac{a_i}{a_i + b_i} = \pi_i\,,$$

$$var(P_i) = \frac{a_i b_i}{(a_i + b_i)^2(a_i + b_i + 1)}$$

$$= \frac{\pi_i(1-\pi_i)}{a_i + b_i + 1} = \pi_i(1-\pi_i)\tau_i^2\,, \qquad \tau_i^2 = \frac{1}{a_i + b_i + 1}\,.$$

Logistische Regression: Überdispersion

Weiters sei $y_i|(P_i = p_i) \overset{ind}{\sim} \text{Binomial}(m_i, p_i)$. Als Randverteilung von y_i resultiert die Beta-Binomial-Verteilung mit

$$\Pr(y_i = y|a_i, b_i) = \binom{m_i}{y} \frac{B(a_i + y, m_i + b_i - y)}{B(a_i, b_i)}, \quad y = 0, 1, \ldots, m_i$$

und

$$\begin{aligned} \mathrm{E}(y_i) &= m_i \frac{a_i}{a_i + b_i} = m_i \pi_i , \\ \mathrm{var}(y_i) &= m_i \frac{a_i b_i}{(a_i + b_i)^2} \frac{a_i + b_i + m_i}{a_i + b_i + 1} \\ &= m_i \pi_i (1 - \pi_i)(1 + \tau_i^2 (m_i - 1)) , \qquad \tau_i^2 > 1 . \end{aligned}$$

Logistische Regression: Überdispersion

Beweis: Wir berechnen marginale Wahrscheinlichkeitsfunktion als

$$\begin{aligned}
\Pr(y_i = y | a_i, b_i) &= \binom{m_i}{y} \frac{1}{B(a_i, b_i)} \int_0^1 p^{a_i - 1 + y} (1-p)^{b_i - 1 + m_i - y} \, dp \\
&= \binom{m_i}{y} \frac{B(a_i + y, b_i + m_i - y)}{B(a_i, b_i)} \\
&= \binom{m_i}{y} \frac{\prod_{k=0}^{y-1} (a_i + k) \prod_{k=0}^{m_i - y - 1} (b_i + k)}{\prod_{k=0}^{m_i - 1} (a_i + b_i + k)} .
\end{aligned}$$

Logistische Regression: Überdispersion

Berechnung der marginalen Momente mittels:

$$\begin{aligned} \mathrm{E}(y) &= \mathrm{E}(\mathrm{E}(y|P)) \\ \mathrm{var}(y) &= \mathrm{E}(\mathrm{var}(y|P)) + \mathrm{var}(\mathrm{E}(y|P)) \,. \end{aligned}$$

Somit resultiert

$$\mathrm{E}(y_i) = \mathrm{E}(m_i P_i) = m_i \pi_i \,,$$

und

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{E}\big(m_i P_i (1 - P_i)\big) + \mathrm{var}(m_i P_i) \\ &= \mathrm{E}(m_i P_i) - \mathrm{E}(m_i P_i^2) + m_i^2 \,\mathrm{var}(P_i) \\ &= m_i \pi_i - m_i \mathrm{E}(P_i^2) + m_i^2 \tau_i^2 \pi_i (1 - \pi_i) \,. \end{aligned}$$

Aus $\mathrm{var}(P_i) = \mathrm{E}(P_i^2) - \mathrm{E}^2(P_i)$ folgt
$\mathrm{E}(P_i^2) = \mathrm{var}(P_i) + \mathrm{E}^2(P_i) = \tau_i^2 \pi_i (1 - \pi_i) + \pi_i^2$ und somit

$$\begin{aligned} \mathrm{var}(y_i) &= m_i \pi_i - m_i \big(\tau_i^2 \pi_i (1 - \pi_i) + \pi_i^2\big) + m_i^2 \tau_i^2 \pi_i (1 - \pi_i) \\ &= m_i \pi_i (1 - \pi_i) \big(1 + \tau_i^2 (m_i - 1)\big) \,. \end{aligned}$$

Logistische Regression: Überdispersion

Reparametrisiere Beta-Binomial mittels $\gamma_i = 1/(a_i + b_i) > 0$, $\pi_i = a_i/(a_i + b_i)$, d.h. $a_i = \pi_i/\gamma_i$, $b_i = (1 - \pi_i)/\gamma_i$, dann folgt

$$\Pr(y_i = y|\pi_i, \gamma_i) = \binom{m_i}{y} \frac{\prod_{k=0}^{y-1} \left(\frac{\pi_i}{\gamma_i} + k\right) \prod_{k=0}^{m_i-y-1} \left(\frac{1-\pi_i}{\gamma_i} + k\right)}{\prod_{k=0}^{m_i-1} \left(\frac{1}{\gamma_i} + k\right)}$$

$$= \binom{m_i}{y} \frac{\prod_{k=0}^{y-1} (\pi_i + k\gamma_i) \prod_{k=0}^{m_i-y-1} (1 - \pi_i + k\gamma_i)}{\prod_{k=0}^{m_i-1} (1 + k\gamma_i)} .$$

Für $\gamma_i \to 0$ resultiert Binomial(m_i, π_i), d.h.

$$\lim_{\gamma_i \to 0} \Pr(y_i = y|\pi_i, \gamma_i) = \binom{m_i}{y} \pi_i^y (1 - \pi_i)^{m_i - y} .$$

Logistische Regression: Überdispersion

Bei unterschiedlichen $\mathrm{E}(P_i)$ ist für $m_i > 1$ die Varianz der y_i gegenüber der binomialen Varianz um Faktor $\tau_i^2(m_i - 1)$ größer.

Fordern wir anstelle der Beta-Verteilung nur eine bestimmte Relation zwischen Erwartungswert und Varianz, so führt dies zu

Lemma

$P_1, \ldots, P_n$ seien stetig, auf $[0, 1]$ unabhängig verteilt mit $E(P_i) = \pi_i$ und $var(P_i) = \phi^2 \pi_i (1 - \pi_i)$, $0 < \phi^2 \leq 1$. Sei weiters $y_i | (P_i = p_i) \overset{ind}{\sim} Binomial(m_i, p_i)$, so folgt für die Momente der nicht bedingten Verteilung von y_i

$$
\begin{aligned}
E(y_i) &= m_i \pi_i \,, \\
var(y_i) &= m_i \pi_i (1 - \pi_i)(1 + \phi^2 (m_i - 1)) \,.
\end{aligned}
$$

(Beweis: verwende in Lemma zuvor ϕ^2 statt τ_i^2.)

Logistische Regression: Überdispersion

Beispiel: An 22 Kliniken wird Wirkung eines neuen Medikaments mit Standardtherapie verglichen.
In jeder Klinik werden Patienten mit einem der beiden Mitteln behandelt. Behandlung ist erfolgreich, wenn Nebenwirkung selten.

```
> clinics <- read.table("clinics.dat", header=TRUE)
> clinics$center   <- factor(rep(1:22, each=2))
> clinics$treatment<- factor(rep(c("new", "control"), times=22))
> attach(clinics)
> SF <- cbind(responses, size-responses)
> summary(glm(SF ~ treatment, family=binomial))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.0794     0.1405 -14.802  < 2e-16 ***
treatmentnew  -1.6463     0.3242  -5.079  3.8e-07 ***
---
    Null deviance: 129.051  on 43  degrees of freedom
Residual deviance:  95.317  on 42  degrees of freedom
AIC: 161.62
```

Logistische Regression: Überdispersion

Hochsignifikanter `treatment` Effekt mit negativem Vorzeichen (Hinweis auf bessere Wirkung des neuen Medikaments).

Aber Deviance ist 95.317, viel größer als Freiheitsgrad 42.

Würde man Klinikeffekt ins Modell nehmen, würde dies am Verhalten Deviance/Freiheitsgrad nicht sehr viel ändern.

Übersicht unter Annahme binomialer Varianz:

Modell	$\hat{\beta}_t$ (*s.e.*)	*Dev.* (*df*)
1		129.05 (43)
`treatment`	-1.646 (0.324)	95.32 (42)
`treatment+center`	-1.780 (0.339)	29.47 (21)
`treatment*center`		0.00 (0)

Interpretation: Hinweis auf Überdispersion!

Logistische Regression: Überdispersion

Betrachte Modell mit Varianzstruktur $\text{var}(y_i) = \phi m_i \pi_i (1 - \pi_i)$.
Dieser QL Ansatz kann in R einfach mittels `glm()` und `family=quasibinomial` angepasst werden.

```
> summary(qb <- glm(SF ~ treatment, family=quasibinomial))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.0794     0.2335  -8.904 3.19e-11 ***
treatmentnew  -1.6463     0.5389  -3.055   0.0039 **
---
    Null deviance: 129.051  on 43  degrees of freedom
Residual deviance:  95.317  on 42  degrees of freedom
AIC: NA
```

Logistische Regression: Überdispersion

Natürlich resultiert dieselbe Deviance und die gleichen Schätzer.

Aber Standardfehler sind um Faktor $\sqrt{2.76} = 1.66$ größer.

Dadurch reduziert sich p-Wert zum `treatment` Effekt geringfügig.

Dispersionsparameter ϕ wird durch mittleren Pearson-Statistik geschätzt, d.h.

```
> sum(residuals(qb, type="pearson")^2)/qb$df.residual
[1] 2.763604
```

Ein logistisches Modell unter der beta-binomialen Varianzstruktur kann mit `glm()` **nicht** geschätzt werden.

Diese Verteilung ist kein Mitglied der einparametrigen linearen Exponentialfamilie.

Logistische Regression: Überdispersion

Bibliothek `gamlss` (Generalized Additive Models for Location Scale and Shape).
Darin sind über 70 diskrete, stetige und gemischte Verteilungen implementiert, auch die Beta-Binomial mittels `family = BB`.
Link für Erwartung $g_\mu(\mu) = \eta$, und Link für Dispersion $g_\sigma(\phi) = \eta^*$. Default sind Logit-Link für μ und Log-Link für ϕ.

```
> library(gamlss)
> summary(gamlss(SF ~ treatment,
+         family=BB(mu.link="logit", sigma.link="identity")))
******************************************************************
Mu link function:  logit
Mu Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)    -2.063      0.2195   -9.396  6.960e-12
treatmentnew   -1.353      0.4173   -3.243  2.322e-03
```

Logistische Regression: Überdispersion

```
-------------------------------------------------------------------
Sigma link function:  identity
Sigma Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   0.07804     0.01979    3.943  0.0002922

-------------------------------------------------------------------
No. of observations in the fit:  44
Degrees of Freedom for the fit:  3
      Residual Deg. of Freedom:  41
                      at cycle:  7

Global Deviance:     139.3722
            AIC:     145.3722
            SBC:     150.7248
```

Logistische Regression: Überdispersion

Signifikanter `treatment` Effekt (etwas schwächer als beim Binomial-Modell).

Schätzung für ϕ ist 0.078 (s.e. 0.0198). Der p-Wert 0.00029 bewertet Hypothese $H_0 : \phi = 0$ (zu verwerfen).

Somit liegen entweder variierende Wahrscheinlichkeiten vor, oder die einzelnen Bernoullivariablen, die das Verhalten der Patienten derselben Klinik beschreiben, sind korreliert.

Logistische Regression: Überdispersion

Alternative aod (Analysis of Overdispersed Data)

```
> library(aod)
> betabin(cbind(responses, size-responses) ~ treatment, ~1,
+         data=clinics)
Beta-binomial model
-------------------
Fixed-effect coefficients:
               Estimate Std. Error    z value Pr(> |z|)
(Intercept)  -2.065e+00  2.303e-01 -8.968e+00 0.000e+00
treatmentnew -1.356e+00  4.245e-01 -3.195e+00 1.399e-03

Overdispersion coefficients:
                 Estimate Std. Error   z value   Pr(> z)
phi.(Intercept) 7.155e-02  3.298e-02 2.169e+00 1.503e-02

Log-likelihood statistics
   Log-lik    nbpar   df res.    Deviance         AIC        AICc
-6.969e+01        3        41  7.707e+01  1.454e+02   1.46e+02
```

Logistische Regression: Überdispersion

Schätzer von Intercept und `treatment` Effekt unterscheiden sich nur geringfügig vom Ergebnis aus `gamlss`.

Etwas größer ist der Unterschied bei der Schätzung von ϕ.

Vor allem der Standardfehler ist nun mit 0.033 deutlich größer als zuvor mit 0.020.

`aod()` verwendet die  Funktion `optim()` um die marginale Log-Likelihoodfunktion in $\boldsymbol{\beta}$ und ϕ zu maximieren.

Poisson Regression: Anzahlen

Binomialverteilte Responses: relative oder absolute Häufigkeiten.
Poissonverteilte Responses: Zählvariablen/Anzahlen (counts).
Annahme: Erwartung entspricht der Varianz, $\mathrm{E}(y_i) = \mu_i = \mathrm{var}(y_i)$
Kanonischer Link: Loglink
Loglineares Modell für Anzahlen:

$$y_i \overset{ind}{\sim} \text{Poisson}(\mu_i) \qquad \text{mit} \quad \log(\mu_i) = \eta_i \,.$$

Die (skalierte) Deviance ist wegen $\phi = 1$ gleich

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\} \,.$$

Falls Intercept im Modell, reduziert sich die Deviance zu

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log \frac{y_i}{\hat{\mu}_i} \,.$$

Devianceanteil verschwindet für $y_i = 0$ unabhängig von $\hat{\mu}_i$.

Poisson Regression: Anzahlen

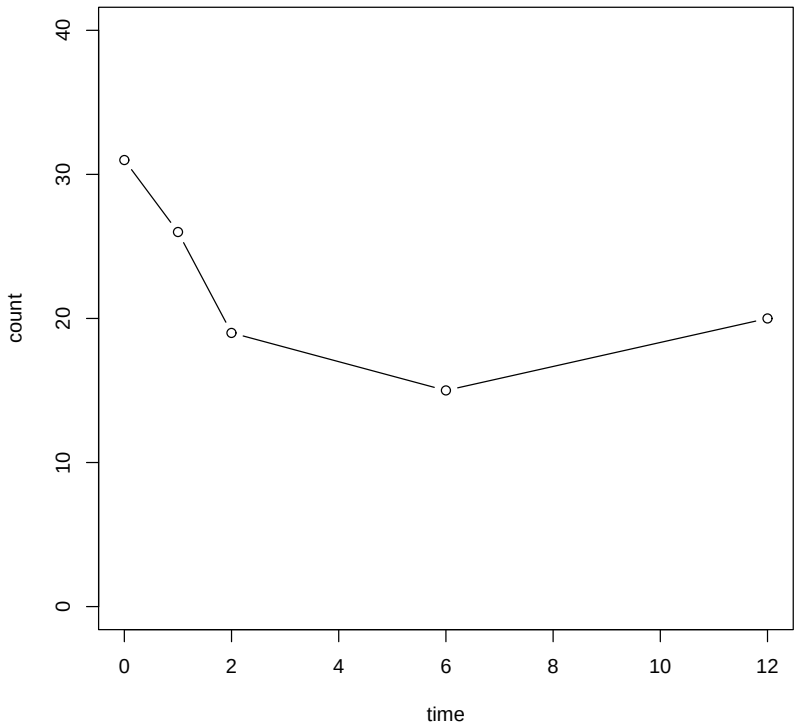
Beispiel: Studie über Aufbewahrbarkeit von Mikroorganismen im tiefgekühlten Zustand(-70^oC).
Bakterienkonzentrationen (Anzahl auf konstanter Fläche) am Anfang und nach 1, 2, 6, und 12 Monaten.

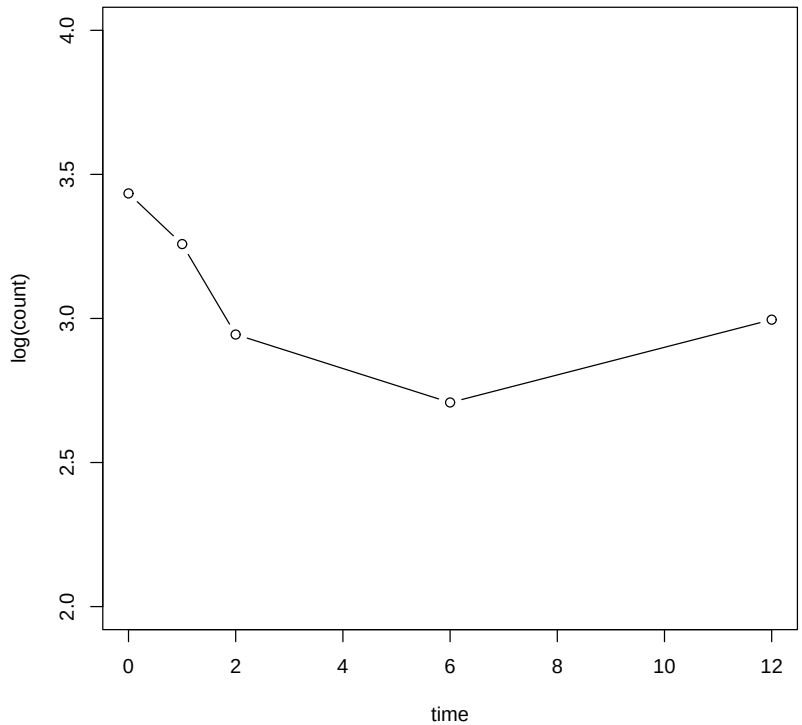
time	0	1	2	6	12
count	31	26	19	15	20

Gesucht: Modell für Konzentration in Abhängigkeit von der Zeit.
Vermutung: erwartete Anzahl μ proportional zu $1/\texttt{time}^\gamma$.

```
> time  <- c( 0, 1, 2, 6,12)
> count <- c(31,26,19,15,20)

> plot(time, count, type="b", ylim=c(0, 40))
> plot(time, log(count), type="b", ylim=c(2, 4))
```

Poisson Regression: Anzahlen





Poisson Regression: Anzahlen

Eigentlich wird exponentielle Abnahme erwartet (aber letzter Wert ist sogar größer als die beiden Werte zuvor).

Vielleicht weichen Beobachtungen nur wegen Messfehler ab.

Dann müsste log(Konzentration) linear zur Zeit sein.

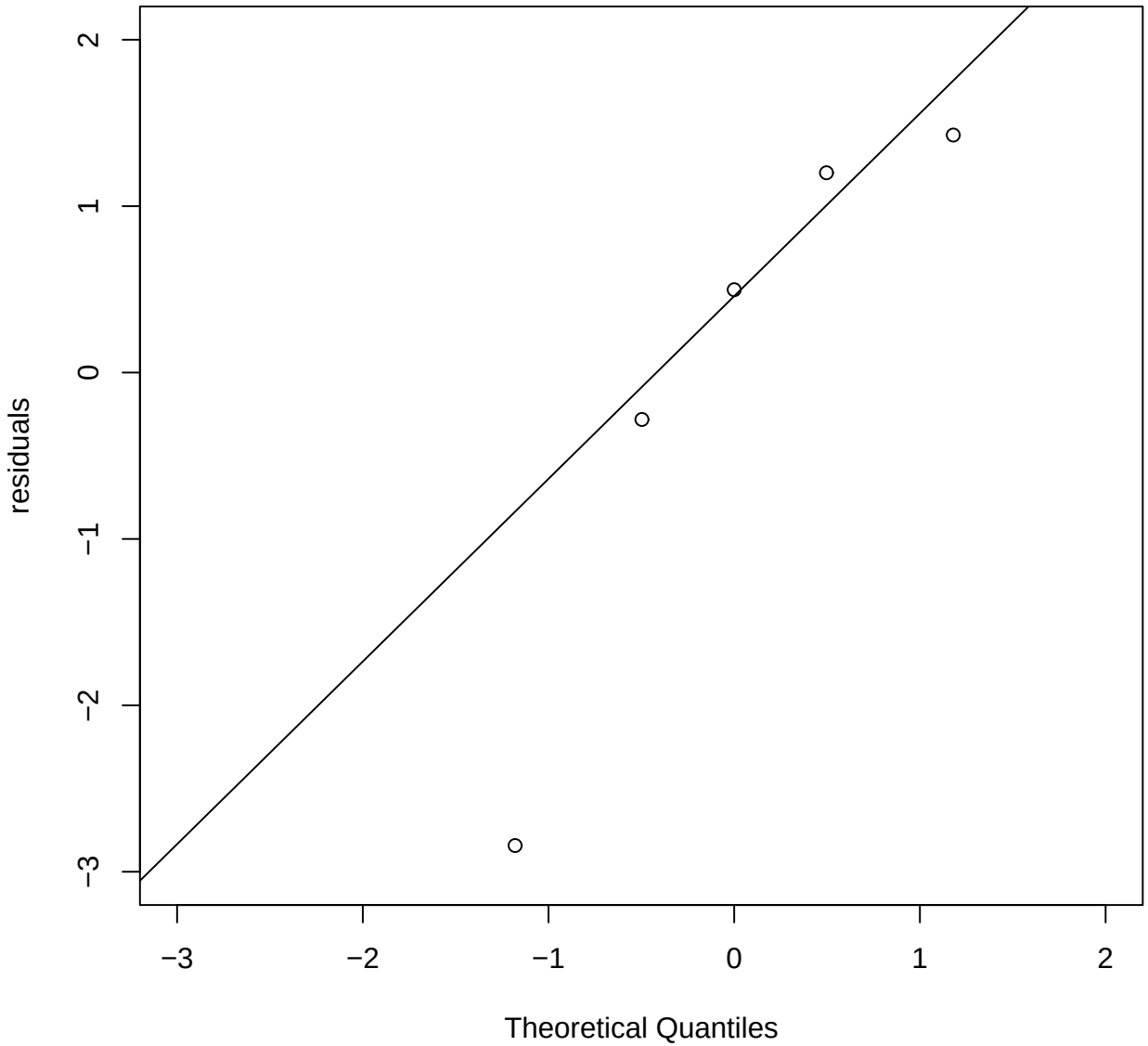
Test durch Modellierung mit quadratischen Zeitterm, ob Krümmung mehr als zufällig ist.

Erste Annahme, Anzahlen sind normalverteilt und genügen linearem Modell in `time` und $\texttt{time}^2$.

Poisson Regression: Anzahlen

```
> summary(mo.lm <- lm(count ~ time + I(time^2)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.80042    1.88294  15.827  0.00397 **
time        -4.61601    1.00878  -4.576  0.04459 *
I(time^2)    0.31856    0.08049   3.958  0.05832 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.438 on 2 degrees of freedom
Multiple R-squared:  0.9252,    Adjusted R-squared:  0.8503
F-statistic: 12.36 on 2 and 2 DF,  p-value: 0.07483

> qqnorm(residuals(mo.lm), ylab="residuals", xlim=c(-3,2),
+        ylim=c(-3,2), main="")
> qqline(residuals(mo.lm))
```

Poisson Regression: Anzahlen



Poisson Regression: Anzahlen

Quadratische Term scheint notwendig (p-Wert 0.058).

Q-Q Plot: Punkteverlauf weicht von Gerade ab
$\Rightarrow$ Normalverteilung scheint unpassend.

$\Rightarrow$ Versuche Poisson-Modell.

Für gewöhnlich werden Poisson-Erwartungen linear auf Log-Skala modelliert (d.h. exponentielle Abnahme der Erwartungen, aber Poissonverteilte Responses um diese Erwartungen betrachtet).

Noch immer quadratische Zeitterm im Modell notwendig?

Poisson Regression: Anzahlen

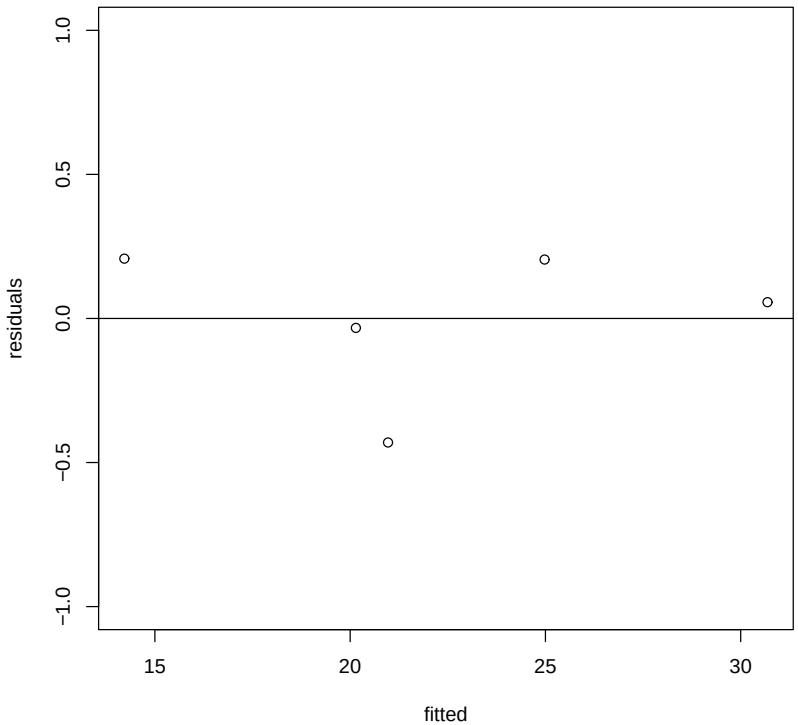
```
> summary(mo.P0 <- glm(count ~ time+I(time^2), family=poisson))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.423818   0.149027  22.975   <2e-16 ***
time        -0.221389   0.095623  -2.315   0.0206 *
I(time^2)    0.015527   0.007731   2.008   0.0446 *
---
    Null deviance: 7.0672  on 4  degrees of freedom
Residual deviance: 0.2793  on 2  degrees of freedom
AIC: 30.849

> r <- residuals(mo.P0, type="pearson"); sum(r^2)
[1] 0.2745424
```

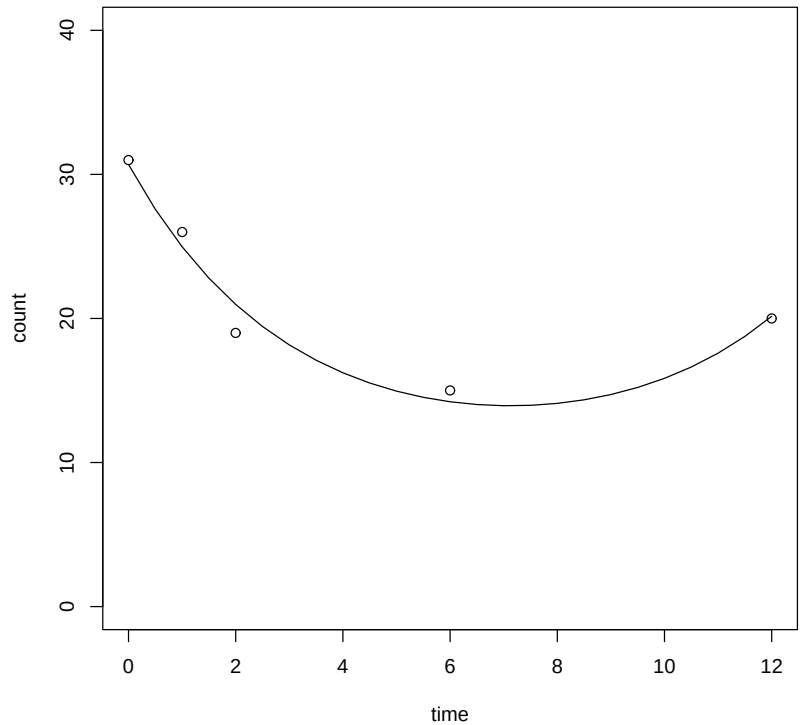
Deviance (0.2793) und $X^2 = 0.2745$ sollten unter wahrem Modell etwa $df = n - p = 2$ entsprechen (Test auf Güte der Anpassung). Da beide Werte klein sind, spricht nichts gegen Poisson-Annahme ($\text{var}(y_i) = \mu_i$).

Poisson Regression: Anzahlen

```
> f <- fitted(mo.P0)
> plot(f, r, ylab="residuals", xlab="fitted", ylim=c(-1,1))
> abline(0,0)

> plot(time, count, ylim=c(0,40))
> time.new <- seq(0, 12, 0.5)
> lines(time.new, predict(mo.P0, data.frame(time=time.new),
+                         type="response"))
```

Poisson Regression: Anzahlen





Poisson Regression: Anzahlen

Residuenplot: falls Varianz der Erwartung entspricht, ist ein Schätzer für die Varianz der Residuen unter dem wahren Modell gerade der geschätzte Erwartungswert. Daher sollten Pearson-Residuen r_i etwa Erwartung Null und Varianz Eins haben.

Residuenplot ist relativ ($n = 5$) unauffällig. Poisson-Annahme scheint passend zu sein.

Um die Modellgüte explorativ zu validieren, werden beobachtete und modellierte Werte gegen die Zeit aufgetragen. Natürlich muss sich das 3 Parameter Modell gut an 5 Beobachtungen anpassen.

Poisson Regression: Anzahlen

Messfehler können auch zum Zuwachs der Anzahl führen (in Realität aber unmöglich).
Wir testen Notwendigkeit des quadratischen Zeitterms mittels Modellvergleich.

```
> mo.P1 <- glm(count ~ time, family=poisson)
> anova(mo.P1, mo.P0, test="Chisq")
Analysis of Deviance Table
Model 1: count ~ time
Model 2: count ~ time + I(time^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3     4.5249
2         2     0.2793  1   4.2456  0.03935 *
```

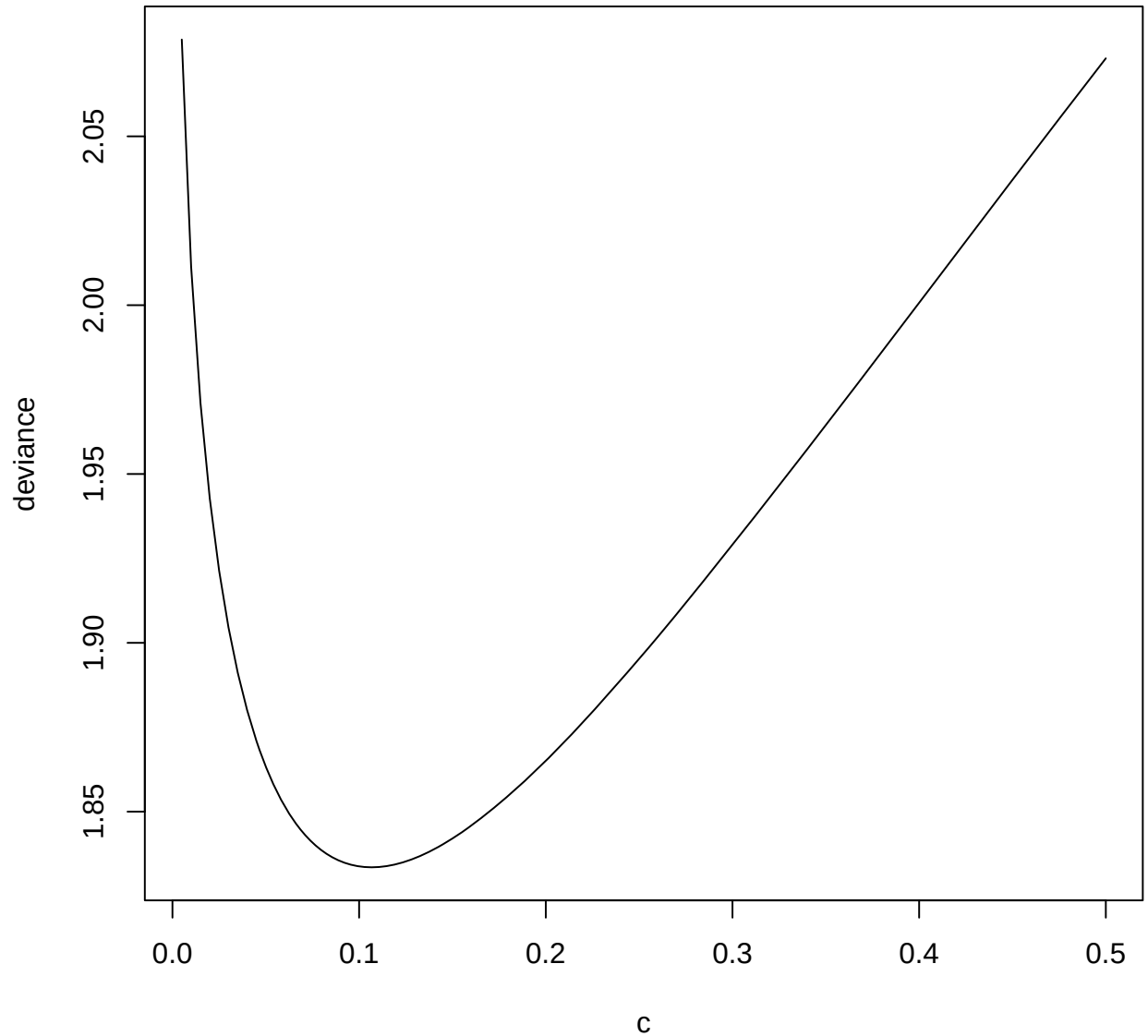
Weist auf Notwendigkeit dieses Effektes (p-Wert$=0.039$) hin.
Aber quadratisches Modell ist eher unsinnig.
Möglicherweise erreicht man realistischeres Modell durch log-Trafo der Zeitachse.

Poisson Regression: Anzahlen

Falls die Zeit multiplikativ wirkt, sollte sich das Modell auf $\log(\texttt{time})$ als Prädiktor beziehen.
Jedoch ist dafür die Anfangszeit, $\log(0)$, problematisch.
Betrachte daher die Transformation $\log(\texttt{time} + c)$ mit unbekanntem Shift $0 < c$.
Um c zu bestimmen, minimieren wir Deviance in c wie folgt:

```
> c <- d <- 1:100
> for (i in 1:100) {
+     c[i] <- i/200
+     d[i] <- deviance(glm(count ~ log(time+c[i]),
+                          family=poisson))
+     }
> plot(c, d, type="l", ylab="deviance")
> c[d==min(d)]
[1] 0.105
```

Poisson Regression: Anzahlen



Poisson Regression: Anzahlen

Optimale Wert für Modell $1 + \log(\texttt{time} + c)$ liegt um $c = 0.105$ und $\log(\texttt{time} + 0.105)$ wird ab jetzt als Prädiktor verwendet.

```
> time.c <- time + 0.105
> summary(mo.P3 <- glm(count ~ log(time.c), family=poisson))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.15110    0.09565  32.945   <2e-16 ***
log(time.c) -0.12751    0.05493  -2.321   0.0203 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7.0672  on 4  degrees of freedom
Residual deviance: 1.8335  on 3  degrees of freedom
AIC: 30.403
```

Poisson Regression: Anzahlen

Wieder ratsam, ein Modell mit quadratischen Zeitterm zu betrachten, um zu prüfen, ob verbliebene Krümmung vorhanden.

```
> mo.P2 <- glm(count ~ log(time.c)+I(log(time.c)^2),
+              family=poisson)
> anova(mo.P3, mo.P2, test="Chisq")
Analysis of Deviance Table

Model 1: count ~ log(time.c)
Model 2: count ~ log(time.c) + I(log(time.c)^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3     1.8335
2         2     1.7925  1  0.04109   0.8394
```

Quadratische Zeitterm nicht mehr signifikant. Es scheint, dass mit der transformierten Zeitachse ein linearer Trend im Prädiktor ausreicht.

Poisson Regression: Anzahlen

Gesucht: approximatives punktweises KIV für $\mu_0 = \exp(\eta_0)$.

Idee 1: verwende $\hat{\eta}_0 = \mathbf{x}_0^\top \hat{\boldsymbol{\beta}}$ mit $\widehat{s.e.}(\hat{\eta}_0)$. Das transformierte 95% Intervall ist

$$KIV(\mu_0) = \left(\exp\left(\hat{\eta}_0 \pm 1.96 \times \widehat{s.e.}(\hat{\eta}_0)\right)\right).$$

Idee 2: Delta Methode liefert

$$\log \hat{\mu} \approx \log \mu + (\hat{\mu} - \mu)\frac{\partial \log \mu}{\partial \mu},$$

somit approximative Varianz, bzw. Standardfehler

$$\operatorname{var}(\log \hat{\mu}) \approx \operatorname{var}(\hat{\mu})\frac{1}{\mu^2}$$

$$\widehat{\operatorname{var}}(\hat{\mu}) \approx \hat{\mu}^2 \operatorname{var}(\hat{\eta}) \qquad \Rightarrow \quad \widehat{s.e.}(\hat{\mu}_0) \approx \hat{\mu}_0 \widehat{s.e.}(\hat{\eta}_0).$$

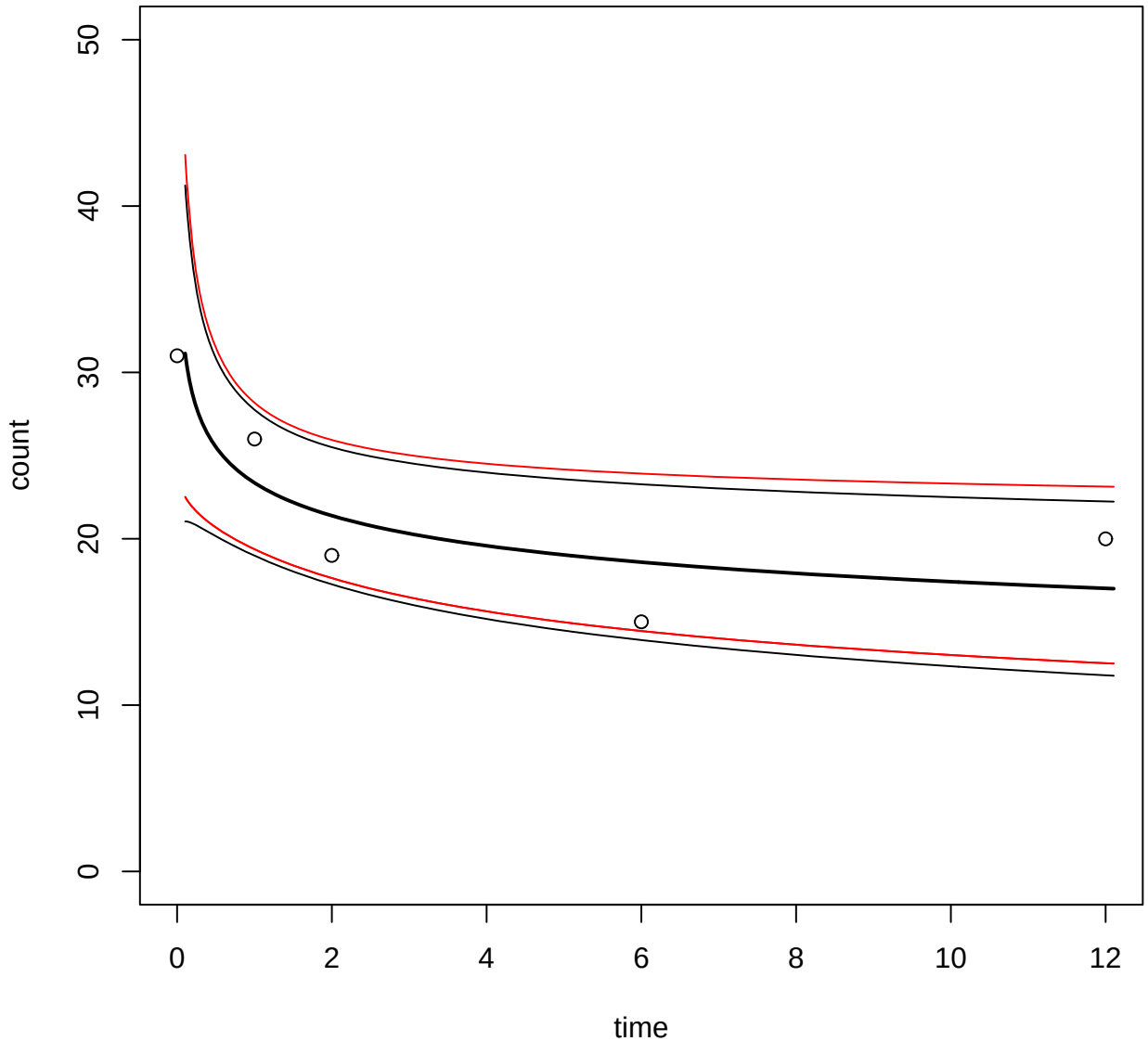
Als 95% KIV folgt

$$KIV_\Delta(\mu_0) = \left(\hat{\mu}_0 \pm 1.96 \times \hat{\mu}_0 \widehat{s.e.}(\hat{\eta}_0)\right).$$

Poisson Regression: Anzahlen

```
> # Delta-Method
> t.new <- data.frame(time.c = seq(0,12,.005) + 0.105)
> r.pred<-predict(mo.P3,newdata=t.new,type="response",se.fit=T)
> fit   <- r.pred$fit
> upper <- fit + qnorm(0.975)*r.pred$se.fit
> lower <- fit - qnorm(0.975)*r.pred$se.fit
> plot(time, count, type="p", xlab="time", ylab="count")
> lines(time.c.new[,1], upper)
> lines(time.c.new[,1], fit)
> lines(time.c.new[,1], lower)

> # using prediction of type="link"
> l.pred <- predict(mo.P3, newdata=t.new, type="link", se.fit=T)
> fit   <- exp(l.pred$fit)
> upper <- exp(l.pred$fit + qnorm(0.975)*l.pred$se.fit)
> lower <- exp(l.pred$fit - qnorm(0.975)*l.pred$se.fit)
> lines(time.c.new[,1], upper, col=2)
> lines(time.c.new[,1], lower, col=2)
```

Poisson Regression: Anzahlen



Poisson Regression: 2D Kontingenztafeln

Loglineare Modelle zur Analyse der Beziehungen zwischen Variablen verwendbar, wie stochastische Unabhängigkeit, konditionale Unabhängigkeit oder eben Abhängigkeit.
Im Speziellen wird hierbei keiner dieser Faktoren als Response definiert. Man spricht hierbei eher nur von **Klassifikatoren**.

Beispiel: Lebensraum von Eidechsen Gezählt wurde, wieviele Eidechsen welchen Aufenthaltsort (engl. perch) gewählt haben. Die Sitzstange ist charakterisiert durch zweistufige Faktoren, Höhe (`height`, ≥ 4.75, < 4.75) und Durchmesser (`diameter`, ≤ 4.0, > 4.0). Folgende Anzahlen sind beobachtet worden:

Perch		`diameter`		
		≤ 4.0	> 4.0	total
`height`	≥ 4.75	61	41	102
	< 4.75	73	70	143
total		134	111	245

Poisson Regression: 2D Kontingenztafeln

Frage: sind `diameter` und `height` Klassifikationen unabhängig? Assoziation mittels **odds-ratios** messbar. Bei Unabhängigkeit wäre odds-ratio Eins. Wir erhalten jedoch als Schätzung

$$\hat{\psi} = \frac{61/41}{73/70} = \frac{61/73}{41/70} = 1.43\,.$$

Deutet dies darauf hin, dass für den wahren Parameter $\psi \neq 1$ gilt?

Wir führen ein loglineares Modell zur Modellierung von 2×2 Tabellen ein und betrachten dazu folgende beobachtete Anzahlen:

	B		
A	1	2	total
1	y_{11}	y_{12}	$y_{1\bullet}$
2	y_{21}	y_{21}	$y_{2\bullet}$
total	$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet\bullet}$

mit $y_{\bullet\bullet} = n$, dem Stichprobenumfang.

Poisson Regression: 2D Kontingenztafeln

Falls wir für y_{kl} die Poissonverteilung annehmen und das Modell mit A und B als erklärende Größen und mit einem Log-Link definieren, entspricht dies einem loglinearen Modell.
Verteilungen von A und von B (Randverteilungen) nicht von Interesse.
Betrachte jetzt:

1. $A + B$ (Unabhängigkeitsmodell),
2. $A * B \equiv A + B + A : B$ (Abhängigkeitsmodell, saturiertes Modell).

Poisson Regression: 2D Kontingenztafeln

Unabhängigkeitsmodell:
Annahme: für alle beobachteten Paare (a_i, b_i), $i = 1, \ldots, n$, ist Wahrscheinlichkeit in Zelle (k, l) zu fallen gleich π_{kl}.
Dann gilt

$$\mathrm{E}(y_{kl}) = \mu_{kl} = n \cdot \pi_{kl} \,, \qquad k, l \in \{1, 2\} \,.$$

Falls stochastische Unabhängigkeit vorliegt, d.h. falls

$$\pi_{kl} = \Pr(A = k, B = l) = \Pr(A = k) \Pr(B = l) = \pi_k^A \pi_l^B \,,$$

dann liefert dafür das loglineare Modell gerade

$$\log \mu_{kl} = \log n + \log \pi_k^A + \log \pi_l^B \,.$$

Der Logarithmus der erwarteten Anzahl in Zelle (k, l) ist additive Funktion des k-ten Zeileneffekts und des l-ten Spalteneffekts.
Dieses Modell ist somit äquivalent mit dem Modell

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B \,, \qquad k, l \in \{1, 2\} \,.$$

Poisson Regression: 2D Kontingenztafeln

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B , \qquad k, l \in \{1, 2\} .$$

Wie sind Parameter zu definieren, und wieviele sind identifizierbar?
Definiere z.B. (bei Verwendung der Kontrast Parametrisierung)

$$\lambda_k^A = \log \pi_k^A - \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^A$$

$$\lambda_l^B = \log \pi_l^B - \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^B$$

$$\lambda = \log n + \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^A + \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^B .$$

Mit dieser Parametrisierung (Abweichungen von den Mitteln) gilt

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{k=1}^{2} \left\{ \log \pi_k^A - \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^A \right\} = 0 = \sum_{l=1}^{2} \lambda_l^B .$$

Poisson Regression: 2D Kontingenztafeln

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{k=1}^{2} \left\{ \log \pi_k^A - \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^A \right\} = 0 = \sum_{l=1}^{2} \lambda_l^B \, .$$

Außer λ nur noch 1 Zeilen- und 1 Spaltenparameter identifizierbar. Für die anderen gilt $\lambda_2^A = -\lambda_1^A$, $\lambda_2^B = -\lambda_1^B$.

Dieses Modell wird loglineares **Unabhängigkeitsmodell** genannt. Wir erhalten folgende lineare Prädiktoren

	B	
A	1	2
1	$\lambda + \lambda_1^A + \lambda_1^B$	$\lambda + \lambda_1^A - \lambda_1^B$
2	$\lambda - \lambda_1^A + \lambda_1^B$	$\lambda - \lambda_1^A - \lambda_1^B$

Poisson Regression: 2D Kontingenztafeln

Alternative Parametrisierung: **Referenzzelle** statt Kontrast. Zeichne dafür beliebige Zelle als Referenz aus und betrachte Parameter, welche die Abweichungen zu dieser beschreiben. Ist z.B. Zelle $(1,1)$ Referenz, dann liefert dies

$$\begin{aligned}\lambda_k^A &= \log \pi_k^A - \log \pi_1^A \\ \lambda_l^B &= \log \pi_l^B - \log \pi_1^B \\ \lambda &= \log n + \log \pi_1^A + \log \pi_1^B\end{aligned}$$

mit den Identifizierbarkeitsbedingungen

$$\lambda_1^A = \lambda_1^B = 0\,.$$

Dies liefert als Prädiktoren

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

Poisson Regression: 2D Kontingenztafeln

MLE einer Wahrscheinlichkeit (Binomialverteilung) ist die entsprechende relative Häufigkeit.

Bei stochastischer Unabhängigkeit gilt somit

$$\hat{\pi}_{kl} = \hat{\pi}_{k\bullet}\hat{\pi}_{\bullet l} \qquad \text{mit } \hat{\pi}_{k\bullet} = \frac{y_{k\bullet}}{y_{\bullet\bullet}} \text{ und } \hat{\pi}_{\bullet l} = \frac{y_{\bullet l}}{y_{\bullet\bullet}} .$$

Für die Erwartungswerte ergibt dies die Schätzer

$$\hat{\mu}_{kl} = n\hat{\pi}_{kl} = y_{\bullet\bullet}\frac{y_{k\bullet}}{y_{\bullet\bullet}}\frac{y_{\bullet l}}{y_{\bullet\bullet}} = \frac{1}{y_{\bullet\bullet}} y_{k\bullet}y_{\bullet l} .$$

Poisson Regression: 2D Kontingenztafeln

Als MLE unserer Parameter liefert dies wiederum sofort

$$\log \hat{\mu}_{11} = \hat{\lambda} = \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}}$$

$$\log \hat{\mu}_{21} = \hat{\lambda} + \hat{\lambda}_2^A = \log \frac{y_{2\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_2^A = \log \frac{y_{2\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{2\bullet}}{y_{1\bullet}}$$

$$\log \hat{\mu}_{12} = \hat{\lambda} + \hat{\lambda}_2^B = \log \frac{y_{1\bullet} y_{\bullet 2}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_2^B = \log \frac{y_{1\bullet} y_{\bullet 2}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{\bullet 2}}{y_{\bullet 1}}$$

Poisson Regression: 2D Kontingenztafeln

Bemerke, dass die Summe der geschätzten Erwartungen die Summe der Beobachtungen reproduziert, also

$$\begin{aligned}\hat{\mu}_{\bullet\bullet} &= \sum_{k=1}^{2}\sum_{l=1}^{2}\hat{\mu}_{kl} = e^{\hat{\lambda}} + e^{\hat{\lambda}+\hat{\lambda}_2^A} + e^{\hat{\lambda}+\hat{\lambda}_2^B} + e^{\hat{\lambda}+\hat{\lambda}_2^A+\hat{\lambda}_2^B} \\ &= e^{\hat{\lambda}}\left[1+e^{\hat{\lambda}_2^A}\right]\left[1+e^{\hat{\lambda}_2^B}\right],\end{aligned}$$

also

$$\begin{aligned}\log\hat{\mu}_{\bullet\bullet} &= \hat{\lambda} + \log\left[1+\frac{y_{2\bullet}}{y_{1\bullet}}\right] + \log\left[1+\frac{y_{\bullet 2}}{y_{\bullet 1}}\right] \\ &= \hat{\lambda} + \log\frac{y_{1\bullet}+y_{2\bullet}}{y_{1\bullet}} + \log\frac{y_{\bullet 1}+y_{\bullet 2}}{y_{\bullet 1}} \\ &= \log\frac{y_{1\bullet}y_{\bullet 1}}{y_{\bullet\bullet}} + \log\frac{y_{\bullet\bullet}}{y_{1\bullet}} + \log\frac{y_{\bullet\bullet}}{y_{\bullet 1}} \\ &= \log y_{\bullet\bullet}\,.\end{aligned}$$

Poisson Regression: 2D Kontingenztafeln

Beachte, dass unter dieser Parametrisierung gilt

$$\begin{aligned}\log \psi &= \log \frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}} \\ &= \log \mu_{11} - \log \mu_{12} - \log \mu_{21} + \log \mu_{22} \\ &= \lambda - (\lambda + \lambda_2^B) - (\lambda + \lambda_2^A) + (\lambda + \lambda_2^A + \lambda_2^B) \\ &= 0\,.\end{aligned}$$

In dieser Parametrisierung ist also ein odds-ratio von $\psi = 1$ mit Unabhängigkeit äquivalent.

Dies hält unabhängig von der Wahl der Referenzzelle.

Poisson Regression: 2D Kontingenztafeln

Saturiertes (volles) Modell:
Kann keine Unabhängigkeit angenommen werden, definieren wir

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}\,, \qquad k, l \in \{1, 2\}\,.$$

Die Interaktionsparameter λ_{kl}^{AB} beschreiben Abweichungen vom Unabhängigkeitsmodell.
Will man mit Kontrasten arbeiten, dann werden mit den linearen Prädiktoren $\eta_{kl} = \log \mu_{kl}$ die Parameter definiert. Seien dazu die mittleren (Zeilen-, Spalten-, overall-) Prädiktoren gleich

$$\eta_{k\bullet} = \frac{1}{2}\sum_{l=1}^{2} \eta_{kl}\,, \quad \eta_{\bullet l} = \frac{1}{2}\sum_{k=1}^{2} \eta_{kl}\,, \quad \eta_{\bullet\bullet} = \lambda = \frac{1}{2}\frac{1}{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \eta_{kl}\,.$$

Poisson Regression: 2D Kontingenztafeln

Definiere Zeileneffekt λ_k^A, Spalteneffekt λ_l^B und Interaktionseffekt (Wechselwirkung) λ_{kl}^{AB} als Abweichung vom mittleren Prädiktor

$$\begin{aligned}
\lambda_k^A &= \eta_{k\bullet} - \eta_{\bullet\bullet} \\
\lambda_l^B &= \eta_{\bullet l} - \eta_{\bullet\bullet} \\
\lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k\bullet} - \eta_{\bullet l} + \eta_{\bullet\bullet} = \underbrace{(\eta_{kl} - \eta_{\bullet\bullet})}_{\eta_{kl}-\lambda} - \underbrace{(\eta_{k\bullet} - \eta_{\bullet\bullet})}_{\lambda_k^A} - \underbrace{(\eta_{\bullet l} - \eta_{\bullet\bullet})}_{\lambda_l^B} .
\end{aligned}$$

λ_k^A, λ_l^B bezeichnen Abweichungen vom Prädiktormittel λ.
λ_{kl}^{AB} sind um Zeilen- und Spalteneffekt bereinigte Zelleneffekte.
Alle Parameter sind mittelwertsbereinigte Effekte, weshalb gilt

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{l=1}^{2} \lambda_l^B = 0 .$$

Wieder nur 1 freier Zeilen- bzw. 1 Spaltenparameter. Für $\lambda_k^A > 0$ ist mittlere (log) erwartete Anzahl für Zellen der k-ten Zeile größer als mittlere (log) erwartete Anzahl über gesamte Tabelle.

Poisson Regression: 2D Kontingenztafeln

Für Interaktionsparameter gilt

$$\sum_{k=1}^{2} \lambda_{kl}^{AB} = \sum_{k=1}^{2} \eta_{kl} - \sum_{k=1}^{2} \eta_{k\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet}$$

$$= 2\eta_{\bullet l} - 2\eta_{\bullet\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet} = 0 = \sum_{l=1}^{2} \lambda_{kl}^{AB} .$$

Somit ist die Summe aller Interaktionen in jeder Zeile und in jeder Spalte gleich Null.

Für eine 2×2 Tabelle gibt es deshalb auch nur 1 freien Interaktionsparameter.

Poisson Regression: 2D Kontingenztafeln

Das Unabhängigkeitsmodell ist Spezialfall des vollen Modells mit $\lambda_{kl}^{AB} = 0$ für alle (k, l).
Die zusätzlichen Parameter λ_{kl}^{AB} sind **Assoziationsparameter**, welche die Abweichungen von der Unabhängigkeit zwischen A und B beschreiben.
Gesamtanzahl freier Parameter ist 3 beim Unabhängigkeitsmodell, bzw. 4 beim Abhängigkeitsmodell.

In R arbeitet man generell mit `treatment` Kontrastierung, d.h. mit Referenzzelle $(1, 1)$. Möchte man mit Kontraste arbeiten, ist folgender Aufruf notwendig

```
> options(contrasts=c("contr.sum", "contr.poly"))
```

Die Default-Einstellung von R ist jedoch

```
> options(contrasts=c("contr.treatment", "contr.poly"))
```

Poisson Regression: 2D Kontingenztafeln

Wiederum ist es einfacher, mit Referenzzelle (z.B. $(1,1)$) zu arbeiten. Wir erhalten mit $\lambda = \eta_{11}$ alternativ

$$\begin{aligned}
\lambda_k^A &= \eta_{k1} - \eta_{11} \\
\lambda_l^B &= \eta_{1l} - \eta_{11} \\
\lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k1} - \eta_{1l} + \eta_{11} = \underbrace{(\eta_{kl} - \eta_{11})}_{\eta_{kl} - \lambda} - \underbrace{(\eta_{k1} - \eta_{11})}_{\lambda_k^A} - \underbrace{(\eta_{1l} - \eta_{11})}_{\lambda_l^B} .
\end{aligned}$$

Jetzt gilt $\lambda_1^A = \lambda_1^B = 0$. Weiters sind alle Interaktionen in der ersten Zeile und in der ersten Spalte Null und wir erhalten

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_1^B + \lambda_{22}^{AB}$

Poisson Regression: 2D Kontingenztafeln

Als MLE ergibt dies

$$\log \hat{\mu}_{11} = \hat{\lambda} = \log y_{11}$$

$$\log \hat{\mu}_{21} = \hat{\lambda} + \hat{\lambda}_2^A = \log y_{21} \Rightarrow \hat{\lambda}_2^A = \log y_{21} - \log y_{11} = \log \frac{y_{21}}{y_{11}}$$

$$\log \hat{\mu}_{12} = \hat{\lambda} + \hat{\lambda}_2^B = \log y_{12} \Rightarrow \hat{\lambda}_2^B = \log y_{12} - \log y_{11} = \log \frac{y_{12}}{y_{11}}$$

$$\begin{aligned} &\log \hat{\mu}_{22} = \hat{\lambda} + \hat{\lambda}_2^A + \hat{\lambda}_2^B + \hat{\lambda}_{22}^{AB} = \log y_{22} \\ &\Rightarrow \hat{\lambda}_{22}^{AB} = \log y_{22} - \log y_{11} - \log \frac{y_{21}}{y_{11}} - \log \frac{y_{12}}{y_{11}} = \log \frac{y_{11} y_{22}}{y_{12} y_{21}} . \end{aligned}$$

MLE des Interaktionseffekts ist das beobachtete log-odds-ratio, das Abweichung vom Unabhängigkeitsmodell schätzt.

Poisson Regression: 2D Kontingenztafeln

Beispiel: Lebensraum von Eidechsen

Um in R Zelle $(1, 1)$ als Referenz zu setzen, benötigen wir z.B.

```
> count <- c(61, 41, 73, 70)

> (hei <-   factor(c(">4.75", ">4.75", "<4.75", "<4.75")))
[1] >4.75 >4.75 <4.75 <4.75
Levels: <4.75 >4.75

> (height <- relevel(hei, ref = ">4.75"))
[1] >4.75 >4.75 <4.75 <4.75
Levels: >4.75 <4.75

> diameter <- factor(c("<4.0", ">4.0", "<4.0", ">4.0"))
```

Poisson Regression: 2D Kontingenztafeln

```
> summary(dep<-glm(count ~ height * diameter, family=poisson))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.1109     0.1280  32.107   <2e-16 ***
height<4.75                 0.1796     0.1735   1.035   0.3006
diameter>4.0               -0.3973     0.2019  -1.967   0.0491 *
height<4.75:diameter>4.0    0.3553     0.2622   1.355   0.1754
---

    Null deviance:  1.0904e+01  on 3  degrees of freedom
Residual deviance: -8.8818e-16  on 0  degrees of freedom
AIC: 31.726
```

Poisson Regression: 2D Kontingenztafeln

Die Deviance ist 0 bei $df = 0$. Modell reproduziert exakt Daten.
Geschätztes odds-ratio ist

```
> exp(dep$coef[4])
height<4.75:diameter>4.0
                1.426662
```

Unter dem Unabhängigkeitsmodell erhalten wir

```
> summary(ind<-glm(count ~ height + diameter, family=poisson))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.0216     0.1148  35.023  < 2e-16 ***
height<4.75    0.3379     0.1296   2.607  0.00913 **
diameter>4.0  -0.1883     0.1283  -1.467  0.14231
---
    Null deviance: 10.9036  on 3  degrees of freedom
Residual deviance:  1.8477  on 1  degrees of freedom
AIC: 31.574
```

Poisson Regression: 2D Kontingenztafeln

Odds-ratio ist jetzt Null und Deviance vergrößert sich um 1.85. Dies kann man als Teststatistik der Hypothese $H_0 : \psi = 1$ verwenden mit p-Wert

```
> pchisq(ind$deviance, 1, lower.tail = FALSE)
[1] 0.174055
```

was auf eine nicht-signifikante Änderung hindeutet (vgl. dies auch mit dem p-Wert 0.142 zur Wald-Statistik). Daher können wir auch nicht $H_0 : \psi = 1$ verwerfen, und `diameter` und `height` scheinen unabhängig zu klassifizieren.

Poisson Regression: 2D Kontingenztafeln

Mehrstufige Faktoren:
Ergebnisse können auf mehrstufige klassifizierende Faktoren verallgemeinert werden. Sei dafür A ein K-stufiger und B ein L-stufiger Faktor. Damit **Unabhängigkeitsmodell**

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B \,, \qquad k = 1, \ldots, K, \; l = 1, \ldots, L \,.$$

Soll Zelle $(1, 1)$ als Referenz dienen, definieren wir

$$\begin{aligned} \lambda_k^A &= \log \pi_k^A - \log \pi_1^A \\ \lambda_l^B &= \log \pi_l^B - \log \pi_1^B \\ \lambda &= \log n + \log \pi_1^A + \log \pi_1^B \,. \end{aligned}$$

und es gelten dieselben Identifizierbarkeitsbedingungen, d.h.

$$\lambda_1^A = \lambda_1^B = 0 \,.$$

Es sind $1 + (K - 1) + (L - 1)$ Parameter frei schätzbar.

Poisson Regression: 2D Kontingenztafeln

Dies liefert als Prädiktoren

			B			
A	1	2	$\cdots$	l	$\cdots$	L
1	λ	$\lambda+\lambda_2^B$	$\cdots$	$\lambda+\lambda_l^B$	$\cdots$	$\lambda+\lambda_L^B$
2	$\lambda+\lambda_2^A$	$\lambda+\lambda_2^A+\lambda_2^B$	$\cdots$	$\lambda+\lambda_2^A+\lambda_l^B$	$\cdots$	$\lambda+\lambda_2^A+\lambda_L^B$
$\vdots$						
k	$\lambda+\lambda_k^A$	$\lambda+\lambda_k^A+\lambda_2^B$	$\cdots$	$\lambda+\lambda_k^A+\lambda_l^B$	$\cdots$	$\lambda+\lambda_k^A+\lambda_L^B$
$\vdots$						
K	$\lambda+\lambda_K^A$	$\lambda+\lambda_K^A+\lambda_2^B$	$\cdots$	$\lambda+\lambda_K^A+\lambda_l^B$	$\cdots$	$\lambda+\lambda_K^A+\lambda_L^B$

Poisson Regression: 2D Kontingenztafeln

Die MLE sind jetzt für $k = 1, \ldots, K$ und $l = 1, \ldots, L$

$$\log \hat{\mu}_{11} = \hat{\lambda} = \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}}$$

$$\log \hat{\mu}_{k1} = \hat{\lambda} + \hat{\lambda}_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{k\bullet}}{y_{1\bullet}}$$

$$\log \hat{\mu}_{1l} = \hat{\lambda} + \hat{\lambda}_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{\bullet l}}{y_{\bullet 1}}$$

Poisson Regression: 2D Kontingenztafeln

Als **saturiertes Modell** für eine $K \times L$ Tabelle folgt

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}, \qquad k = 1, \ldots, K, \; l = 1, \ldots, L.$$

Mit Referenzzelle $(1, 1)$ folgt für alle $k = 1, \ldots, K$, $l = 1, \ldots, L$

$$\begin{aligned}
\lambda_k^A &= \eta_{k1} - \eta_{11} \\
\lambda_l^B &= \eta_{1l} - \eta_{11} \\
\lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k1} - \eta_{1l} + \eta_{11} = \underbrace{(\eta_{kl} - \eta_{11})}_{\eta_{kl} - \lambda} - \underbrace{(\eta_{k1} - \eta_{11})}_{\lambda_k^A} - \underbrace{(\eta_{1l} - \eta_{11})}_{\lambda_l^B},
\end{aligned}$$

wofür $\lambda_1^A = \lambda_1^B = 0$ gilt.
Wieder sind in Zeile 1 und in Spalte 1 alle Interaktionen Null.
Somit sind genau $1 + (K-1) + (L-1) + (K-1)(L-1) = K \times L$ Parameter frei schätzbar.

Poisson Regression: 2D Kontingenztafeln

Wir erhalten damit als Prädiktoren

			B			
A	1	2	$\cdots$	l	$\cdots$	L
1	λ	$\lambda+\lambda_2^B$	$\cdots$	$\lambda+\lambda_l^B$	$\cdots$	$\lambda+\lambda_L^B$
2	$\lambda+\lambda_2^A$	$\lambda+\lambda_2^A+\lambda_2^B+\lambda_{22}^{AB}$	$\cdots$	$\lambda+\lambda_2^A+\lambda_l^B+\lambda_{2l}^{AB}$	$\cdots$	$\lambda+\lambda_2^A+\lambda_L^B+\lambda_{2L}^{AB}$
$\vdots$						
k	$\lambda+\lambda_k^A$	$\lambda+\lambda_k^A+\lambda_2^B+\lambda_{k2}^{AB}$	$\cdots$	$\lambda+\lambda_k^A+\lambda_l^B+\lambda_{kl}^{AB}$	$\cdots$	$\lambda+\lambda_k^A+\lambda_L^B+\lambda_{kL}^{AB}$
$\vdots$						
K	$\lambda+\lambda_K^A$	$\lambda+\lambda_K^A+\lambda_2^B+\lambda_{K2}^{AB}$	$\cdots$	$\lambda+\lambda_K^A+\lambda_l^B+\lambda_{Kl}^{AB}$	$\cdots$	$\lambda+\lambda_K^A+\lambda_L^B+\lambda_{KL}^{AB}$

Saturierte Modell hat um $(K-1)(L-1)$ freie Parameter mehr als Unabhängigkeitsmodell.

Poisson Regression: 2D Kontingenztafeln

Beispiel: Zervixkarzinom

Liegt bei Daten zum Zervixkarinom stochastische Unabhängigkeit zwischen den Prädiktoren Grenzzonenbefall (GZ) und Anzahl befallener Lymphknotenstationen (LK) vor?
Wir betrachten somit die folgenden Häufigkeiten:

	LK-Stationen			
	0	1	2	≥ 3
GZ nicht befallen	124	21	16	13
GZ befallen	58	12	7	5
über GZ befallen	14	19	12	12

Wir passen zuerst das saturierte Modell den Daten an und prüfen damit die Notwendigkeit der Interaktion.

Poisson Regression: 2D Kontingenztafeln

```
> anova(glm(total ~ G*L, family=poisson), test="Chisq")
Analysis of Deviance Table
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    11    316.184
G     2   69.569          9    246.615 7.821e-16 ***
L     3  203.594          6     43.021 < 2.2e-16 ***
G:L   6   43.021          0      0.000 1.155e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hinweis, dass die 6 Interaktionsparameter nicht Null sind und somit die Unabhängigkeitshypothese verworfen werden kann.

Poisson Regression: 2D Kontingenztafeln

Alternativ kann dazu auch die Pearson-Statistik unter dem Unabhängigkeitsmodell betrachtet werden, d.h.

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

mit $\log \mu_{ij} = \lambda + \lambda_i^G + \lambda_j^K$. Diese Statistik realisiert in

```
> ind <- glm(total ~ G+L, family=poisson)
> r <- residuals(ind, type="pearson")
> sum(r^2)
[1] 43.83645
```

und entspricht der χ^2-Statistik bei der Kontingenztafelanalyse.

Poisson Regression: 2D Kontingenztafeln

Am einfachsten erhält man diese Pearson-Statistik mittels

```
> (N <- matrix(total, 3, 4, byrow=TRUE))
     [,1] [,2] [,3] [,4]
[1,]  124   21   16   13
[2,]   58   12    7    5
[3,]   14   19   12   12
> chisq.test(N)

        Pearson's Chi-squared test

data:  N
X-squared = 43.8365, df = 6, p-value = 7.965e-08
```

Poisson Regression: 2D Kontingenztafeln

Poisson- und binomialverteilte Responses scheinen bei $K \times 2$ Tabellen Gemeinsamkeiten aufzuweisen.
Bezeichne Tabelleneinträge mit N_{kl}, $k = 1, \ldots, K$ und $l \in \{1, 2\}$ (z.B. k unterschiedliche Experimentierumgebungen und l Responsekategorien "Erfolg/Misserfolg").
Für unabhängige Anzahlen $N_{kl} \overset{ind}{\sim} \text{Poisson}(\mu_{kl})$ liefert Konditionieren auf Gesamtanzahl der beiden Responses

$$N_{k\bullet} = N_{k1} + N_{k2}\,, \qquad k = 1, \ldots, K\,,$$

als konditionale Verteilung für jede Experimentierumgebung

$$N_{k2}|N_{k\bullet} \sim \text{Binomial}(N_{k\bullet}, \pi_k)\,, \qquad k = 1, \ldots, K\,,$$

mit

$$\pi_k = \frac{\mu_{k1}}{\mu_{k1} + \mu_{k2}}\,.$$

Poisson Regression: 2D Kontingenztafeln

Dieser Aspekt wird später noch im Detail besprochen. Wir diskutieren jetzt dieses Verhalten für eine $K \times 2$ Tafel.
Wegen der Annahme $N_{k1} \overset{ind}{\sim} \text{Poisson}(\mu_{k1})$ ist

$$\Pr(N_{k1} = n_{k1}) = \frac{\mu_{k1}^{n_{k1}}}{n_{k1}!} e^{-\mu_{k1}}$$

und speziell

$$\Pr(N_{k\bullet} = n_{k\bullet}) = \frac{\mu_{k\bullet}^{n_{k\bullet}}}{n_{k\bullet}!} e^{-\mu_{k\bullet}}$$

mit $\mu_{k\bullet} = \mu_{k1} + \mu_{k2}$ und $n_{k\bullet} = n_{k1} + n_{k2}$.
Bemerke, dass N_{k1} die Anzahl Erfolge beschreibt und $N_{k\bullet}$ die Gesamtanzahl der Versuche im k-ten Umfeld zählt.

Poisson Regression: 2D Kontingenztafeln

Damit ergibt sich bei gegebener (fixierter) Anzahl an Versuchen

$$
\begin{aligned}
\Pr(N_{k1} = n_{k1} | N_{k\bullet} = n_{k\bullet}) &= \frac{\Pr(N_{k1} = n_{k1}, N_{k2} = n_{k2})}{\Pr(N_{k\bullet} = n_{k\bullet})} \\
&= \frac{\frac{\mu_{k1}^{n_{k1}}}{n_{k1}!} e^{-\mu_{k1}} \frac{\mu_{k2}^{n_{k2}}}{n_{k2}!} e^{-\mu_{k2}}}{\frac{(\mu_{k1}+\mu_{k2})^{n_{k1}+n_{k2}}}{(n_{k1}+n_{k2})!} e^{-(\mu_{k1}+\mu_{k2})}} \\
&= \frac{(n_{k1}+n_{k2})!}{n_{k1}! n_{k2}!} \frac{\mu_{k1}^{n_{k1}}}{(\mu_{k1}+\mu_{k2})^{n_{k1}}} \frac{\mu_{k2}^{n_{k2}}}{(\mu_{k1}+\mu_{k2})^{n_{k2}}} \\
&= \binom{n_{k1}+n_{k2}}{n_{k1}} \left(\frac{\mu_{k1}}{\mu_{k1}+\mu_{k2}}\right)^{n_{k1}} \left(\frac{\mu_{k2}}{\mu_{k1}+\mu_{k2}}\right) \\
&= \binom{n_{k\bullet}}{n_{k1}} \pi_k^{n_{k1}} (1-\pi_k)^{n_{k2}} .
\end{aligned}
$$

Poisson Regression: 2D Kontingenztafeln

Man bemerke, dass hierbei

$$\text{logit}(\pi_k) = \log \frac{\pi_k}{1-\pi_k} = \log \frac{\mu_{k1}}{\mu_{k2}} = \log \mu_{k1} - \log \mu_{k2}$$

gilt, also dass der logit der Wahrscheinlichkeit eines Erfolgs mit einem Effekt auf der Poisson loglinearen Achse äquivalent ist.

Poisson Regression: 2D Kontingenztafeln

Beispiel: Rezidivbildung beim Zervixkarzinom

Untersuche, ob Lymphknotenbefall einen Einfluss auf die Wahrscheinlichkeit einer Rezidivbildung hat.

	$L = 0$	$L = 1$	$L = 2$	$L = 3$
$R = 0$	153	23	12	2
$R = 1$	43	29	23	28

Dies kann über ein loglineares Poissonmodell oder mittels eines logistischen Binomialmodells gemacht werden.

```
> count <- c(153, 23, 12, 2, 43, 29, 23, 28)
> L <- factor(c(0, 1, 2, 3, 0, 1, 2, 3))
> R <- c(0, 0, 0, 0, 1, 1, 1, 1)
```

Poisson Regression: 2D Kontingenztafeln

```
> summary(mod.P <- glm(count ~ R*L + L, family=poisson))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.03044    0.08085  62.223  < 2e-16 ***
R           -1.26924    0.17260  -7.354 1.93e-13 ***
L1          -1.89494    0.22364  -8.473  < 2e-16 ***
L2          -2.54553    0.29978  -8.491  < 2e-16 ***
L3          -4.33729    0.71171  -6.094 1.10e-09 ***
R:L1         1.50104    0.32826   4.573 4.81e-06 ***
R:L2         1.91983    0.39573   4.851 1.23e-06 ***
R:L3         3.90830    0.75200   5.197 2.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance:  3.0017e+02  on 7  degrees of freedom
Residual deviance: -4.4409e-15  on 0  degrees of freedom
AIC: 55.771
```

Poisson Regression: 2D Kontingenztafeln

```
> summary(mod.B <- glm(R ~ L, family=binomial, weight=count))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2692     0.1726  -7.354 1.93e-13 ***
L1            1.5010     0.3283   4.573 4.81e-06 ***
L2            1.9198     0.3957   4.851 1.23e-06 ***
L3            3.9083     0.7520   5.197 2.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 419.46  on 7  degrees of freedom
Residual deviance: 337.34  on 4  degrees of freedom
AIC: 345.34
```

Poisson Regression: 2D Kontingenztafeln

Spezifikation `R*L+L` für Prädiktor beim loglinearen Modell besteht aus 2 Teilen.

- `R*L` bezieht sich auf Interaktion zwischen binomialen `R` und Prädiktor unter dem logistischen Ansatz (`L`).
- `L` ist saturiert bezüglich des klassifizierenden Faktors und sichert somit, dass für jedes der vier Zellenpaare (Anzahlen zu `R=0` und `R=1`) die beobachtete Anzahl der geschätzten Anzahl entspricht (Reproduktion der marginalen binomialen Totalsummen).

Dadurch entsprechen im loglinearen Modell die vier Interaktionen mit `R` (das sind die Parameter `R`, `R:L1`, `R:L2`, sowie `R:L3`) den Parametern im logistischen Modell.

Poisson Regression: 3D Kontingenztafeln

Dreidimensionale Kontingenztafeln

Folgende Daten motivieren weitere Überlegungen zur Unabhängigkeit.

Beispiel: Diabetesstudie Diabetespatienten sind bezüglich

- Familien-Vorgeschichte betreffs Diabetes (vorhanden/nicht-vorhanden)
- Insulinabhängigkeit (ja/nein)
- Alter zum Zeitpunkt des Ausbruchs (< 45 / ≥ 45)

klassifiziert. Folgende Anzahlen liegen dazu vor:

	`history`			
	yes		no	
	`insulin`		`insulin`	
	yes	no	yes	no
`age` < 45	6	1	16	2
`age` ≥ 45	6	36	8	48

Poisson Regression: 3D Kontingenztafeln

Interesse an Abhängigkeitsstruktur der 3 Klassifikatoren.

Betrachte dazu verschiedene loglineare Modelle.

Im Folgenden sind Ergebnisse des **saturierten** Modells angeführt.

```
> count   <- c(6, 1, 16, 2, 6, 36, 8, 48)
> age     <- factor(c("<45", "<45", "<45", "<45",
+                     ">45", ">45", ">45", ">45"))
> hist    <- factor(c("yes", "yes", "no", "no",
+                     "yes", "yes", "no", "no"))
> history <- relevel(hist, ref = "yes")
> insu    <- factor(c("yes", "no", "yes", "no",
+                     "yes", "no", "yes", "no"))
> insulin <- relevel(insu, ref = "yes")
```

Poisson Regression: 3D Kontingenztafeln

```
> summary(mod1 <-glm(count~age*history*insulin, family=poisson))

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.792e+00  4.082e-01   4.389 1.14e-05
age>45                        8.955e-16  5.774e-01   0.000 1.00000
historyno                     9.808e-01  4.787e-01   2.049 0.04047 *
insulinno                    -1.792e+00  1.080e+00  -1.659 0.09715 .
age>45:historyno             -6.931e-01  7.217e-01  -0.960 0.33683
age>45:insulinno              3.584e+00  1.167e+00   3.072 0.00213 *
historyno:insulinno          -2.877e-01  1.315e+00  -0.219 0.82683
age>45:historyno:insulinno    2.877e-01  1.439e+00   0.200 0.84150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1.2516e+02  on 7  degrees of freedom
Residual deviance: 1.3323e-15  on 0  degrees of freedom
AIC: 47.626
```

Poisson Regression: 3D Kontingenztafeln

Suchen **passendes** Modell. Betrachten dazu Deviance zu verschiedenen Prädiktoren.

Modell	Deviance	df
A+H+I	51.93	4
A*I+H	1.95	3
A*H+I	50.03	3
I*H+A	51.02	3
A*I+I*H	1.04	2
A*I+A*H	0.05	2
I*H+A*H	49.12	2
A*I+I*H+A*H	0.04	1
A*I*H	0.00	0

Poisson Regression: 3D Kontingenztafeln

Saturierte Modell `A*I*H` hat Deviance Null bei $df = 0$.

Dreifachinteraktion scheint irrelevant zu sein (Verschlechterung der Deviance um 0.04), was zum Modell `A*I+I*H+A*H` führt.

Weglassen der Interaktion `A:I` verschlechtert Modell signifikant (Deviance Differenz von 49.08).

Jedoch kann man auf Interaktion `I:H` verzichten (Deviance Differenz von 0.01), was `A*I+A*H` ergibt. Jetzt kann man noch

auf Interaktion `A:H` verzichten (Deviance Differenz von 1.90).

Weitere Vereinfachungen würden dieses Modell `A*I+H` signifikant verschlechtern.

Poisson Regression: 3D Kontingenztafeln

Einzige verbliebene Interaktion ist `age:insulin`.

Also scheinen `age` und `insulin` unabhängig von `history` zu sein.

Die dazugehörenden log-odds-ratios (Interaktionen) scheinen nicht signifikant unterschiedlich von der Null zu sein.

Wir können somit Tabelle über `history` zusammen legen

	`insulin`	
	yes	no
`age` <45	22	3
`age` ≥ 45	14	84

Für diese Tabelle ergibt der Schätzer des odds-ratios

$$\hat{\psi} = \frac{22 \cdot 84}{14 \cdot 3} = 44\,,$$

was bedeutet, dass die Chance einer Insulinabhängigkeit bei den jüngeren Patienten das 44–fache von jener bei den älteren ist.

Poisson Regression: 3D Kontingenztafeln

Die beiden **konditionalen odds-ratios** bei den individuellen Tabellen sind unter Verwendung des saturierten Modells gerade

$\hat{\psi}_{\text{yes}} = (6 \cdot 36)/(6 \cdot 1) = 36$ bei der Gruppe mit Vorgeschichte

und

$\hat{\psi}_{\text{no}} = (16 \cdot 48)/(2 \cdot 8) = 48$ bei jenen ohne Vorgeschichte.

Somit hat das Zusammenklappen der Tabelle die Beziehung zwischen `insulin` und `age` nicht verzerrt.

Poisson Regression: 3D Kontingenztafeln

Generell definiert man **konditionale odds-ratios** zwischen zwei Klassifikatoren, indem man den dritten Klassifikator auf einer Stufe festhält.

Betrachten wir zuerst das saturierte Modell für 3 klassifizierende binäre Faktoren A, B und C

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \, .$$

Als log-odds-ratio zwischen A und B gegeben $C = 1$ folgt

$$\begin{aligned} \log \frac{\mu_{11(1)}\mu_{22(1)}}{\mu_{12(1)}\mu_{21(1)}} &= \left(\lambda\right) + \left(\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}\right) - \left(\lambda + \lambda_2^B\right) - \left(\lambda + \lambda_2^A\right) \\ &= \lambda_{22}^{AB} \, , \end{aligned}$$

weil in dieser Parametrisierung sämtliche Parameter Null sind, falls $i = 1$, $j = 1$ oder $k = 1$ ist.

Poisson Regression: 3D Kontingenztafeln

Konditionieren wir auf die zweite Stufe von C, so resultiert dafür

$$\begin{aligned}\log \frac{\mu_{11(2)}\mu_{22(2)}}{\mu_{12(2)}\mu_{21(2)}} &= \left(\lambda + \lambda_2^C\right) \\ &\quad + \left(\lambda + \lambda_2^A + \lambda_2^B + \lambda_2^C + \lambda_{22}^{AB} + \lambda_{22}^{BC} + \lambda_{22}^{AC} + \lambda_{222}^{ABC}\right) \\ &\quad - \left(\lambda + \lambda_2^B + \lambda_2^C + \lambda_{22}^{BC}\right) - \left(\lambda + \lambda_2^A + \lambda_2^C + \lambda_{22}^{AC}\right) \\ &= \lambda_{22}^{AB} + \lambda_{222}^{ABC} \,.\end{aligned}$$

Die dreifache Interaktion λ_{222}^{ABC} beschreibt daher gerade den Unterschied in den beiden konditionalen log-odds-ratios:

```
> exp(mod1$coef[6])
age>45:insulinno
              36
> exp(mod1$coef[6]+mod1$coef[8])
age>45:insulinno
              48
```

Poisson Regression: 3D Kontingenztafeln

Für das adäquate Modell erhalten wir folgenden Schätzer:

```
> summary(mod2<-glm(count~age*insulin+history, family=poisson))

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.1707     0.2403   9.034  < 2e-16 ***
age>45            -0.4520     0.3419  -1.322  0.18615
insulinno         -1.9924     0.6155  -3.237  0.00121 **
historyno          0.4122     0.1842   2.238  0.02520 *
age>45:insulinno   3.7842     0.6798   5.567  2.6e-08 ***
---
    Null deviance: 125.1626  on 7  degrees of freedom
Residual deviance:   1.9463  on 3  degrees of freedom
AIC: 43.572

> exp(mod2$coef[5])
age>45:insulinno
              44
```

Poisson Regression: 3D Kontingenztafeln

Interpretation derartige Modelle im Allgemeinen?
Für Modell mit 3 klassifizierenden Faktoren A, B und C gilt

1. `A*B*C` reproduziert das beobachtete Responsemuster
2. `A*B+B*C+C*A` ist das Modell ohne dreifacher Interaktion
3. `A*B+B*C` beschreibt konditionale Unabhängig von A und C gegeben B, wir schreiben $(A \perp C)|B$
   1. die geschätzten odds-ratios für A, B sind gleich auf sämtlichen Stufen von C
   2. die geschätzten odds-ratios für B, C sind gleich auf sämtlichen Stufen von A
   3. die geschätzten odds-ratios für A, C sind 1 auf sämtlichen Stufen von B, aber das marginale odds-ratio ist nicht 1.
4. `A*B+C` postuliert, dass die gemeinsame Verteilung von A, B für alle Stufen von C dieselbe ist. Ist dies der Fall, dann kann die Tabelle über C zusammengelegt werden.
5. `A+B+C` bedeutet vollständige Unabhängigkeit (mutual independence model).

Poisson Regression: Multinomiale Response Modelle

Interesse: Zusammenhang zwischen loglinearem Modell für Anzahlen und multinomialen (binomialen) Response Modell für Häufigkeiten.

Beziehung stammt von Tatsache, dass Multinomialverteilung aus einer Serie unabhängiger Poissonvariablen generiert werden kann, wenn deren Summe fixiert wird.

Multinomialverteilung
Objekte einer Population weisen eines von K Merkmalen auf, z.B. Haarfarbe oder Todesursache. Aus dieser Population zieht man Zufallsstichprobe und ist an der Anzahl Y_k jener Beobachtungen mit Ausprägung $k = 1, \ldots, K$ interessiert.

Poisson Regression: Multinomiale Response Modelle

Nehmen wir an, dass es sich bei den Anzahlen um unabhängige Poissonvariablen handelt mit Erwartungswerten $\mu_1, \ldots, \mu_K$, also

$$\Pr(Y_k = y_k) = \frac{\mu_k^{y_k}}{y_k!} \exp(-\mu_k), \qquad k = 1, \ldots, K .$$

Für deren Summe gilt $Y_\bullet = \sum_{k=1}^K Y_k \sim \text{Poisson}(\mu_\bullet = \sum_{k=1}^K \mu_k)$.
Bedingte Wahrscheinlichkeit der $(Y_1, \ldots, Y_K)$ gegeben $Y_\bullet$ ist

$$\Pr(Y_1 = y_1, \ldots, Y_K = y_K | Y_\bullet) = \frac{\prod_{k=1}^K \frac{\mu_k^{y_k}}{y_k!} \exp(-\mu_k)}{\frac{\mu_\bullet^{y_\bullet}}{y_\bullet!} \exp(-\mu_\bullet)} = \frac{y_\bullet! \prod_{k=1}^K \mu_k^{y_k}}{\mu_\bullet^{y_\bullet} \prod_{k=1}^K y_k!}$$

$$= \frac{y_\bullet!}{\prod_{k=1}^K y_k!} \prod_{k=1}^K \left(\frac{\mu_k}{\mu_\bullet}\right)^{y_k} = \frac{y_\bullet!}{\prod_{k=1}^K y_k!} \prod_{k=1}^K \pi_k^{y_k} .$$

Poisson Regression: Multinomiale Response Modelle

$$\Pr(Y_1 = y_1, \ldots, Y_K = y_K | Y_\bullet) = \frac{y_\bullet!}{\prod_{k=1}^{K} y_k!} \prod_{k=1}^{K} \pi_k^{y_k}$$

Multinomialverteilung auf K Zellen mit Wahrscheinlichkeiten beschrieben durch relative Erwartungswertanteile $\pi_k = \mu_k / \mu_\bullet$.

Der Erwartungswert μ_k bezeichnet bei der Poissonverteilung die theoretische Durchschnittsanzahl der Ereignisse vom Typ k, während π_k bei der Multinomialverteilung den theoretischen Anteil der Ereignisse dieses Typs beschreibt.

Durch diese Konstruktion ist sichergestellt, dass $\sum_k \pi_k = 1$ gilt.

Poisson Regression: Multinomiale Response Modelle

Stichprobe $y_1, \ldots, y_n$ von n nicht identischer aber unabhängiger Multinomialvektoren $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$, und sei die Zeilensumme $y_{i\bullet} = \sum_k y_{ik}$ fest für jedes i ("Stichprobenumfang" des i-ten Multinomialvektors).

Zu jedem $\mathbf{y}_i$ korrespondierte $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ für $i = 1, \ldots, n$.

Multinomiales Logitmodel bezieht sich also auf Daten der Form

Kovariablen-klasse	Erklärender Vektor	Response Kategorie 1	...	k	...	K	Zeilen-summe
1	x_1	y_{11}	...	y_{1k}	...	y_{1K}	$y_{1\bullet}$
2	x_2	y_{21}	...	y_{2k}	...	y_{2K}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_{i1}	...	y_{ik}	...	y_{iK}	$y_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_{n1}	...	y_{nk}	...	y_{nK}	$y_{n\bullet}$

Poisson Regression: Multinomiale Response Modelle

Der für die Schätzung von $\boldsymbol{\pi}_i$ relevante Beitrag der i-ten Beobachtung an der Log-Likelihood der Stichproben ist gerade

$$\ell(\boldsymbol{\pi}_i|\mathbf{y}_i) = \sum_{k=1}^{K} y_{ik} \log \pi_{ik} \,, \qquad \text{mit} \quad \sum_{k=1}^{K} \pi_{ik} = 1 \,.$$

Wegen Unabhängigkeit der $\mathbf{y}_i$ resultiert als Log-Likelihood

$$\ell(\boldsymbol{\pi}|\mathbf{y}) = \sum_{i=1}^{n} \ell(\pi_i|y_i) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log \pi_{ik} \,.$$

Für die Maximierung von $\ell(\boldsymbol{\pi}|\mathbf{y})$ folgt unter den n Bedingungen $\sum_k \pi_{ik} = 1$, $i = 1, \ldots, n$, als Scorefunktion

$$\frac{\partial}{\partial \pi_{ik}} \Big(\ell(\boldsymbol{\pi}|\mathbf{y}) - \sum_{i=1}^{n} \lambda_i \Big(\sum_{k=1}^{K} \pi_{ik} - 1 \Big) \Big) = \frac{y_{ik}}{\pi_{ik}} - \lambda_i \,.$$

Poisson Regression: Multinomiale Response Modelle

$$\frac{\partial}{\partial \pi_{ik}} \Big(\ell(\boldsymbol{\pi}|\mathbf{y}) - \sum_{i=1}^{n} \lambda_i \Big(\sum_{k=1}^{K} \pi_{ik} - 1 \Big) \Big) = \frac{y_{ik}}{\pi_{ik}} - \lambda_i \, .$$

Nullsetzen liefert $\hat{\pi}_{ik} = y_{ik}/\hat{\lambda}_i$.

Summieren über alle K geschätzten Wahrscheinlichkeiten ergibt $1 = \sum_k \hat{\pi}_{ik} = \sum_k y_{ik}/\hat{\lambda}_i = y_{i\bullet}/\hat{\lambda}_i$, also $\hat{\lambda}_i = y_{i\bullet}$.

Die Scorefunktion ist somit

$$\frac{\partial \ell(\boldsymbol{\pi}|\mathbf{y})}{\partial \pi_{ik}} = \frac{y_{ik} - y_{i\bullet}\pi_{ik}}{\pi_{ik}} \, .$$

Poisson Regression: Multinomiale Response Modelle

Im logistischen Modell hat man für binomialverteilte Responses (mit gerade 2 Kategorien) den Logarithmus des einen odds-ratios betrachtet und diesen linear modelliert, also

$$\text{logit}(\pi_i) = \log \frac{\pi_{i1}}{1-\pi_{i1}} = \log \frac{\pi_{i1}}{\pi_{i2}} = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \,.$$

Beim multinomialen Logitmodell sind zum Vergleich K Kategorien verfügbar. Für gewöhnlich wird 1. Kategorie dafür verwendet, d.h.

$$\log \frac{\pi_{ik}}{\pi_{i1}} = \log \frac{y_{i\bullet}\pi_{ik}}{y_{i\bullet}\pi_{i1}} = \log \frac{\mu_{ik}}{\mu_{i1}} = \eta_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta}_k \,, \qquad k = 2, \ldots, K$$

mit $\eta_{i1} = 0$ für $i = 1, \ldots, n$. Bemerke, dass $\boldsymbol{\beta}_k$ kategorienspezifisch ist um Effekt von $\mathbf{x}_i$ auf π_{ik} zu beschreiben. Die Verwendung der 1. Kategorie ist keinerlei Restriktion, da

$$\log \frac{\pi_{ik}}{\pi_{i1}} - \log \frac{\pi_{ij}}{\pi_{i1}} = \log \frac{\pi_{ik}}{\pi_{ij}} = \eta_{ik} - \eta_{ij} \,, \qquad k = 2, \ldots, K \,.$$

Poisson Regression: Multinomiale Response Modelle

Unter diesem Modell gilt für die Wahrscheinlichkeiten

$$\log \frac{\pi_{ik}}{\pi_{i1}} = \eta_{ik} \quad \Rightarrow \quad \pi_{ik} = \pi_{i1} e^{\eta_{ik}}$$

$$1 = \sum_{k=1}^{K} \pi_{ik} = \pi_{i1} \sum_{k=1}^{K} e^{\eta_{ik}} ,$$

also

$$\pi_{i1} = 1 \Big/ \sum_{k=1}^{K} e^{\eta_{ik}}$$

und

$$\pi_{ik} = e^{\eta_{ik}} \Big/ \sum_{k'=1}^{K} e^{\eta_{ik'}} , \qquad k = 2, \ldots, K .$$

Für $K = 2$ reduziert sich Modell auf binomiales logit-Link Modell. Für $K > 2$ sprengt Modell den GLM Rahmen. Parameter aber über loglineares Poisson GLM schätzbar. Schlüssel liegt in Spezifikation des linearen Prädiktors.

Poisson Regression: Multinomiale Response Modelle

Vergleich von Poisson-Erwartungen:

Wir betrachten unabhängige $Y_k \overset{ind}{\sim} \text{Poisson}(\mu_k)$ mit

$$\log \mu_k = \phi + \mathbf{x}_k^\top \boldsymbol{\beta}\,, \qquad k = 1, \ldots, K,$$

wobei $\mathbf{x}_k$ gegebene Konstanten sind und $\boldsymbol{\beta}$ einen unbekannten Parameter (-vektor ohne Intercept) bezeichnet. Der relevante Teil der Poisson-Log-Likelihoodfunktion

$$\ell(\boldsymbol{\mu}|\mathbf{y}) = \sum_{k=1}^{K} \Big(y_k \log \mu_k - \mu_k \Big)$$

entspricht hier

$$\ell(\phi, \boldsymbol{\beta}|\mathbf{y}) = \sum_{k=1}^{K} \Big(y_k(\phi + \mathbf{x}_k^\top \boldsymbol{\beta}) - \exp(\phi + \mathbf{x}_k^\top \boldsymbol{\beta}) \Big)\,.$$

Poisson Regression: Multinomiale Response Modelle

Definiere

$$\mu_\bullet = \sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} \exp(\phi + \mathbf{x}_k^\top \boldsymbol{\beta}) = \exp(\phi) \sum_{k=1}^{K} \exp(\mathbf{x}_k^\top \boldsymbol{\beta}) \,,$$

also

$$\log \mu_\bullet = \phi + \log\Big(\sum_{k=1}^{K} \exp(\mathbf{x}_k^\top \boldsymbol{\beta})\Big) \;\Rightarrow\; \phi = \log \mu_\bullet - \log\Big(\sum_{k=1}^{K} \exp(\mathbf{x}_k^\top \boldsymbol{\beta})\Big) \,.$$

Bezüglich $(\mu_\bullet, \boldsymbol{\beta})$ erhält man als Log-Likelihood

$$\begin{aligned}
\ell(\mu_\bullet, \boldsymbol{\beta}|\mathbf{y}) &= \sum_{k=1}^{K} \left\{ y_k \Big(\log \mu_\bullet - \log\Big(\sum_{k'=1}^{K} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta})\Big) + \mathbf{x}_k^\top \boldsymbol{\beta} \Big) \right\} - \mu_\bullet \\
&= \log \mu_\bullet \sum_{k=1}^{K} y_k - \sum_{k=1}^{K} y_k \log\Big(\sum_{k'=1}^{K} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta})\Big) + \sum_{k=1}^{K} y_k \mathbf{x}_k^\top \boldsymbol{\beta} - \mu_\bullet \,.
\end{aligned}$$

Poisson Regression: Multinomiale Response Modelle

$$\ell(\mu_\bullet, \boldsymbol{\beta}|\mathbf{y}) = \sum_{k=1}^{K} \left\{ y_k \left(\log \mu_\bullet - \log \left(\sum_{k'=1}^{K} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta}) \right) + \mathbf{x}_k^\top \boldsymbol{\beta} \right) \right\} - \mu_\bullet$$

$$= \log \mu_\bullet \sum_{k=1}^{K} y_k - \sum_{k=1}^{K} y_k \log \left(\sum_{k'=1}^{K} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta}) \right) + \sum_{k=1}^{K} y_k \mathbf{x}_k^\top \boldsymbol{\beta} - \mu_\bullet \,.$$

Mit $y_\bullet = \sum_k y_k$ folgt dafür

$$\ell(\mu_\bullet, \boldsymbol{\beta}|\mathbf{y}) = \left\{ y_\bullet \log \mu_\bullet - \mu_\bullet \right\} + \left\{ \sum_{k=1}^{K} y_k \mathbf{x}_k^\top \boldsymbol{\beta} - y_\bullet \log \left(\sum_{k=1}^{K} \exp(\mathbf{x}_k^\top \boldsymbol{\beta}) \right) \right\}$$

$$= \ell(\mu_\bullet | y_\bullet) + \ell(\boldsymbol{\beta}, \mathbf{y}|y_\bullet) \,.$$

Poisson Regression: Multinomiale Response Modelle

$\Rightarrow$ Log-Likelihood zerfällt in zwei Komponenten.

$\ell(\mu_\bullet | y_\bullet)$ entspricht Log-Likelihood zu $y_\bullet \sim \text{Poisson}(\mu_\bullet)$.

Um 2. Komponente zu untersuchen, verwende Parametrisierung

$$\pi_k = \frac{\mu_k}{\mu_\bullet} = \frac{\exp(\phi + \mathbf{x}_k^\top \boldsymbol{\beta})}{\sum_{k'} \exp(\phi + \mathbf{x}_{k'}^\top \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_k^\top \boldsymbol{\beta})}{\sum_{k'} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta})} .$$

Wegen

$$\log \pi_k = \mathbf{x}_k^\top \boldsymbol{\beta} - \log \left(\sum_{k'} \exp(\mathbf{x}_{k'}^\top \boldsymbol{\beta}) \right) ,$$

ist 2. Term äquivalent mit

$$\ell(\boldsymbol{\beta}, \mathbf{y} | y_\bullet) = \sum_{k=1}^{K} y_k \log \pi_k ,$$

einem multinomialen Log-Likelihood, also der konditionalen Poisson-Log-Likelihood von $\mathbf{y}$ gegeben die Gesamtanzahl $y_\bullet$.

Poisson Regression: Multinomiale Response Modelle

Erkenntnis:

$\ell(\mu_\bullet | y_\bullet)$ hängt nicht von $\boldsymbol{\beta}$ ab.

$\ell(\boldsymbol{\beta}, \mathbf{y} | y_\bullet)$ hängt nicht von $\mu_\bullet$ ab.

$\Rightarrow$ gesamte Information über $\boldsymbol{\beta}$ ist im zweiten Ausdruck enthalten.

$\Rightarrow$ MLE von $\boldsymbol{\beta}$ und dessen asymptotische Varianz basierend auf $\ell(\mu_\bullet, \boldsymbol{\beta} | \mathbf{y})$ (loglineares Poissonmodell) oder auf $\ell(\boldsymbol{\beta}, \mathbf{y} | y_\bullet)$ (multinomiales Logitmodell) sind dieselben.

Wir müssen einzig in passender Weise einen Nuisance Parameter $\mu_\bullet$ in das Modell aufnehmen.

Poisson Regression: Multinomiale Response Modelle

Zeige: bestimmte loglineare Modelle entsprechen multinomialen Modellen.
Ordne y_{ik}, mit Kovariablensituation $i = 1, \ldots, n$ und Kategorie $k = 1, \ldots, K$, Zellen einer Kontingenztafel zu und betrachte

$$\log \mu_{ik} = \phi_i + \mathbf{x}_i^\top \boldsymbol{\beta}_k \,, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K$$

mit $\mu_{ik} = \mathrm{E}(y_{ik})$ und Prädiktoren $\mathbf{x}_i$ ($p-1$ Elemente, ohne Intercept).
Prädiktoren sind beobachtungs- und nicht kategoriespezifisch.
Interesse: Parameter $\boldsymbol{\beta}_k$; die ϕ_i sind Nebenparameter ("nuisance parameters").
Dimension des Parameterraums ist $n + (p-1)$. Daher ist auch für $n \to \infty$ nicht garantiert, dass MLE effizient oder konsistent ist.

Poisson Regression: Multinomiale Response Modelle

Zeige: konditionale Likelihood hängt nur von $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)^\top$ und nicht von $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^\top$ ab.
Analog wie zuvor ist Log-Likelihood der Stichprobe

$$\begin{aligned}\ell(\boldsymbol{\phi}, \boldsymbol{\beta}|\mathbf{y}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\Big(y_{ik}(\phi_i + \mathbf{x}_i^\top\boldsymbol{\beta}_k) - \exp(\phi_i + \mathbf{x}_i^\top\boldsymbol{\beta}_k)\Big) \\ &= \sum_{i=1}^{n}\phi_i y_{i\bullet} + \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{x}_i^\top\boldsymbol{\beta}_k - \sum_{i=1}^{n}\sum_{k=1}^{K}\exp(\phi_i + \mathbf{x}_i^\top\boldsymbol{\beta}_k)\,.\end{aligned}$$

Fixiere i-te Responsesumme $y_{i\bullet}$ und betrachte Transformation

$$\mu_{i\bullet} = \sum_{k=1}^{K}\mu_{ik} = \exp(\phi_i)\sum_{k=1}^{K}\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_k)$$

$$\Rightarrow \phi_i = \log\mu_{i\bullet} - \log\Big(\sum_{k=1}^{K}\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_k)\Big)\,.$$

Poisson Regression: Multinomiale Response Modelle

Log-Likelihood als Funktion in $(\mu_\bullet, \boldsymbol{\beta})$ ist

$$\begin{aligned}\ell(\boldsymbol{\mu}_\bullet, \boldsymbol{\beta}|\mathbf{y}) &= \sum_{i=1}^{n}\left\{y_{i\bullet}\log\mu_{i\bullet} - \mu_{i\bullet}\right\} \\ &\quad + \sum_{i=1}^{n}\left\{\sum_{k=1}^{K} y_{ik}\mathbf{x}_i^\top\boldsymbol{\beta}_k - y_{i\bullet}\log\left(\sum_{k'=1}^{K}\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_{k'})\right)\right\} \\ &= \ell(\boldsymbol{\mu}_\bullet|\mathbf{y}_\bullet) + \ell(\boldsymbol{\beta}, \mathbf{y}|\mathbf{y}_\bullet)\,.\end{aligned}$$

1. Term: Likelihood zu n Summen $y_{i\bullet} \sim \text{Poisson}(\mu_{i\bullet})$.
2. Term: Likelihood zu $\mathbf{y}_i|y_{i\bullet}$; hängt von $\boldsymbol{\beta}$ und nicht von $\boldsymbol{\phi}$ ab.

Gesamte Information über $\boldsymbol{\beta}$ im 2. Term.

$\Rightarrow \hat{\boldsymbol{\beta}}$ und $\text{var}(\hat{\boldsymbol{\beta}})$ basierend auf $\ell(\boldsymbol{\beta}, \mathbf{y}|\mathbf{y}_\bullet)$ oder auf $\ell(\boldsymbol{\mu}_\bullet, \boldsymbol{\beta}|\mathbf{y})$ identisch.

$\Rightarrow$ loglineares Modell äquivalent mit dem multinomialen Modell

$$\pi_{ik} = \frac{\mu_{ik}}{\mu_{i\bullet}} = \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_k)}{\sum_{k'=1}^{K}\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_{k'})}\,.$$

Poisson Regression: Multinomiale Response Modelle

Beispiel: Rezidiv beim Zervixkarzinom
Analyse nun mittels Poissonvariablen und Log-Link. Betrachte dazu (binomiale) Datenstruktur

	befallene LK-Stationen							
	0		1		2		≥ 3	
	$R=0$	$R=1$	$R=0$	$R=1$	$R=0$	$R=1$	$R=0$	$R=1$
GZ nicht befallen	103	21	14	7	7	9	0	13
GZ befallen	40	18	6	6	2	5	0	5
über GZ befallen	10	4	3	16	3	9	2	10

Poisson Regression: Multinomiale Response Modelle

Um die Ergebnisse beim logistischen Modell mit Prädiktor $\mathtt{L} + \mathtt{G}$ zu erhalten, muss Summe der beiden Poisson-Beobachtungen in den von $\mathtt{L}*\mathtt{G}$ aufgespannten Zellen (in jeder Zelle befindet sich jeweils eine $\mathtt{R} = 0$ und eine $\mathtt{R} = 1$ Häufigkeit) fixiert werden.

$\Rightarrow$ Aufnahme der Interaktion $\mathtt{L}*\mathtt{G}$ in den Prädiktor (ergibt die Nebenparameter $\boldsymbol{\phi}$).

Eigentliches Modell für Erwartungswert geht als Interaktion mit $\mathtt{R}$ in den Prädiktor ein (rezidivspezifische Parameter), d.h.

$$\log \mu = \mathtt{R} * (\mathtt{L} + \mathtt{G}) + \mathtt{L} * \mathtt{G} \,.$$

Poisson Regression: Multinomiale Response Modelle

```
> count <- c(rez, total-rez)

> (R <- gl(2, 12, labels=1:0))
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
Levels: 1 0

> (L <- factor(rep(rep(0:3, 3), 2)))
 [1] 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3
Levels: 0 1 2 3

> (G <- factor(rep(rep(0:2, each=4), 2)))
 [1] 0 0 0 0 1 1 1 1 2 2 2 2 0 0 0 0 1 1 1 1 2 2 2 2
Levels: 0 1 2
```

Poisson Regression: Multinomiale Response Modelle

```
> summary(MN.glm <- glm(count ~ R*(L+G) + L*G, family=poisson))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.032932   0.203404  14.911  < 2e-16 ***
R0           1.604143   0.219202   7.318 2.51e-13 ***
L1          -0.852487   0.328546  -2.595 0.009467 **
L2          -0.870085   0.352170  -2.471 0.013487 *
L3          -0.573696   0.349225  -1.643 0.100431
G1          -0.196387   0.286157  -0.686 0.492530
G2          -1.387115   0.382373  -3.628 0.000286 ***
R0:L1       -1.287280   0.347971  -3.699 0.000216 ***
R0:L2       -1.778549   0.412032  -4.317 1.59e-05 ***
R0:L3       -3.797851   0.761597  -4.987 6.14e-07 ***
R0:G1       -0.728501   0.312710  -2.330 0.019825 *
R0:G2       -1.073534   0.381676  -2.813 0.004913 **
L1:G1       -0.007513   0.410137  -0.018 0.985385
 :
```

Poisson Regression: Multinomiale Response Modelle

```
    Null deviance: 434.130  on 23  degrees of freedom
Residual deviance:  10.798  on  6  degrees of freedom
AIC: 134.75
```

$\mathtt{L} * \mathtt{G}$ vergibt jeder Zelle einen Parameter (ϕ), während die Prädiktoren des Logit-Modells in den Wechselwirkungen mit `R` (hier $\mathtt{L} + \mathtt{G}$) zu finden sind (somit $p = 18$ Parameter).

Bemerke, dass die relevanten Parameter zu `R0`, `R0:L1`, `R0:L2`, `R0:L3`, `R0:G1` und `R0:G2` gehören und exakt jenen im Logit-Modell entsprechen.

```
> coefficients(rez.LG)
(Intercept)          L1          L2          L3          G1          G2
   1.604143   -1.287280 -1.778549 -3.797851 -0.728501 -1.073534
> deviance(rez.LG)
[1] 10.79795
```

Natürlich haben sich Deviance und Freiheitsgrade auch nicht geändert.

Poisson Regression: Multinomiale Response Modelle

Als Prognosewerte erhält man unter diesem Modell

```
> grid <- list(R=levels(R), L=levels(L), G=levels(G))
> MN.p <- predict(MN.glm, expand.grid(grid), type="response")
> (pred <- t(matrix(MN.p, 8)))
      [,1]    [,2]   [,3]   [,4]  [,5]  [,6]   [,7]   [,8]
[1,] 20.758 103.242  8.850 12.150 8.696 7.304 11.696 1.3041
[2,] 17.057  40.943  7.218  4.782 4.981 2.019  4.745 0.2553
[3,]  5.185   8.815 12.932  6.068 9.323 2.677 11.559 0.4405
```

Durch das Fixieren der Ränder ist in jeder Zelle (für jede `L*G` Kombination) die Summe der geschätzten Erwartungen $\sum_{r=1}^{2} \hat{\mu}_{ir}$ gleich der beobachteten Gesamtanzahl $\sum_{r=1}^{2} y_{ir}$.

Poisson Regression: Multinomiale Response Modelle

$\text{logit}(\mu) = 1$ entspricht z.B. Annahme, dass Zufallsstichprobe multinomialverteilter Anzahlen vorliegt. Spezifiziert wird es als

$$\log \mu = \mathtt{R} + \mathtt{L} * \mathtt{G} \,.$$

Modell mit $p = 13$ Parametern. `L*G` fixiert Ränder und $\mathtt{R} \equiv \mathtt{R}*\mathtt{1}$ steht nur in "Wechselwirkung" mit Intercept.
Wahrscheinlichkeit eines Rezidivs in jeder Zelle dieselbe:

```
> MN.glm1 <- glm(count ~ R + L*G, family=poisson)
> MN.p1 <- predict(MN.glm1, expand.grid(grid), type="response")
> (pred <- t(matrix(MN.p1, 8)))
      [,1]   [,2]  [,3]   [,4]  [,5]  [,6]  [,7]  [,8]
[1,] 48.728 75.272 8.252 12.748 6.288 9.712 5.109 7.891
[2,] 22.792 35.208 4.716  7.284 2.751 4.249 1.965 3.035
[3,]  5.502  8.498 7.466 11.534 4.716 7.284 4.716 7.284
> pred[1,2]/pred[1,1]
[1] 1.545
> pred[3,4]/pred[3,3]
[1] 1.545
```

Poisson Regression: Multinomiale Response Modelle

Alternative: `multinom` in Bibliothek `nnet` (multinomiale Daten werden direkt modelliert).

```
> library(nnet)
> MN.direct <- multinom(R ~ L+G, weights=count)

> summary(MN.direct)

Coefficients:
             Values Std. Err.
(Intercept)  1.6041    0.2192
L1          -1.2873    0.3480
L2          -1.7785    0.4120
L3          -3.7979    0.7616
G1          -0.7285    0.3127
G2          -1.0735    0.3817

Residual Deviance: 326.8
AIC: 338.8
```

Poisson Regression: Multinomiale Response Modelle

Deviance vergleicht mit Modell, das 313 einzelne Beobachtungen korrekt vorhersagt (und nicht mit vollem Modell zu den 12 Eintragungen der 3×4 Tafel).

Dieser Vergleich resultiert aus

```
> MN.direct.sat <- multinom(R ~ L*G, weights=count)
```

```
> anova(MN.direct, MN.direct.sat)
  Model Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1 L + G        -6    326.762             NA       NA        NA
2 L * G       -12    315.964 1 vs 2      6  10.7979 0.0948282
```

Zufällige Effekte: Prädiktoren

Bis jetzt: Prädiktoren der Form

$$\eta_i = g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \,.$$

Unbekannter (fester) Parameter $\boldsymbol{\beta}$, der geschätzt werden muss.

Ab jetzt: zusätzlich zum festen Effekt noch ein zufälliger Effekt.

Motivation A:
Fehlen relevante erklärende Variablen $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip'})^\top$, kann durch Hinzunahme eines **zufälligen** (skalaren) Effekts $z_i = \mathbf{u}_i^\top \boldsymbol{\gamma}$ die sonst resultierende Überdispersion berücksichtigt werden, d.h.

$$\eta_i = g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + z_i \,.$$

Zufällige Effekte: Prädiktoren

Motivation B:
Liegen n unabhängige Gruppen $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ vor, mit Abhängigkeit unter den y_{ij}. Zufälliger Effekt z_i für Responses derselben Gruppe, d.h.

$$\eta_{ij} = g(\mu_{ij}) = x_{ij}^\top \beta + z_i \,, \qquad j = 1, \ldots, n_i \,,$$

impliziert Korrelation (alle y_{ij} verwenden dasselbe z_i).

Nichtbeobachtbaren (**latenten**) z_i bilden Zufallsstichprobe (iid).

Zufällige Prädiktoren für konditionalen Erwartungswert $\mu_i = \mathrm{E}(y_i|z_i)$ (bzw. $\mu_{ij} = \mathrm{E}(y_{ij}|z_i)$).

Annahme: konditionale Verteilung der Response LEF Mitglied, gegeben die zufälligen Effekte.

MLE maximiere aber **marginale Likelihood** (nur selten analytisch berechenbar).
$\Rightarrow$ **EM-Schätzung**.

Zufällige Effekte: EM-Schätzer

Dempster, Laird & Rubin (1977): EM-Algorithmus

Annahme: Daten bestehen aus beobachtbaren Teil y und nicht beobachtbaren Teil z.

Gemeinsame Dichte einer vollständigen Beobachtung (y, z) sei $f(y, z|\boldsymbol{\theta})$ (alle Parameter zu $\boldsymbol{\theta} \in \Theta$ zusammengefasst mit Parameterraum Θ).

Somit gilt für die marginale Dichte $f(y|\boldsymbol{\theta})$ der Response y:

$$\ell(\boldsymbol{\theta}|y) = \log f(y|\boldsymbol{\theta}) = \log \int f(y, z|\boldsymbol{\theta}) dz \,.$$

Um MLE $\hat{\theta}$ zu bestimmen, wird marginale $\ell(\boldsymbol{\theta}|y)$ maximiert. Integral ist in der Praxis häufig problematisch.

Zufällige Effekte: EM-Schätzer

Mit

$$f(z|y;\boldsymbol{\theta}) = \frac{f(y,z|\boldsymbol{\theta})}{f(y|\boldsymbol{\theta})}\,,$$

folgt

$$\ell(\boldsymbol{\theta}|y) = \log f(y|\boldsymbol{\theta}) = \log f(y,z|\boldsymbol{\theta}) - \log f(z|y;\boldsymbol{\theta})\,.$$

MLE $\hat{\boldsymbol{\theta}}$ maximiert gerade $\ell(\boldsymbol{\theta}|y)$. Es gilt

$$\log f(y,z|\boldsymbol{\theta}) = \log f(z|y;\boldsymbol{\theta}) + \ell(\boldsymbol{\theta}|y)\,. \qquad (\clubsuit)$$

Da z nicht beobachtet, ersetze in ($\clubsuit$) fehlende Information durch ihren konditionalen Erwartungswert, gegeben das was beobachtet vorliegt (also gegeben y).
Für Berechnung verwende beliebigen zulässigen Parameterwert $\boldsymbol{\theta}_0 \in \Theta$, d.h. Erwartungswert bzgl. $f(z|y,\boldsymbol{\theta}_0)$.

Zufällige Effekte: EM-Schätzer

Der konditionale Erwartungswert von $\ell(\boldsymbol{\theta}|y)$, gegeben y, ist wiederum $\ell(\boldsymbol{\theta}|y)$, womit für (✪) folgt

$$\mathrm{E}\Big(\log f(y,z|\boldsymbol{\theta})\Big|y,\boldsymbol{\theta}_0\Big) = \mathrm{E}\Big(\log f(z|y,\boldsymbol{\theta})\Big|y,\boldsymbol{\theta}_0\Big) + \mathrm{E}\Big(\ell(\boldsymbol{\theta}|y)\Big|y,\boldsymbol{\theta}_0\Big)$$

$$\int \log f(y,z|\boldsymbol{\theta}) f(z|y,\boldsymbol{\theta}_0)\,dz = \int \log f(z|y,\boldsymbol{\theta}) f(z|y,\boldsymbol{\theta}_0)\,dz + \int \ell(\boldsymbol{\theta}|y) f(z|y,\boldsymbol{\theta}_0)\,dz$$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) + \ell(\boldsymbol{\theta}|y)\,. \qquad (✿)$$

Maximierung von $\ell(\boldsymbol{\theta}|y)$ ist äquivalent mit Maximierung von $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - H(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$. Bemerke, (✿) hält für beliebige $\boldsymbol{\theta} \in \Theta$, also auch für $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ womit folgt

$$Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) = H(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) + \ell(\boldsymbol{\theta}_0|y)$$

erhalten. Somit resultiert als Differenz zu (✿)

$$\ell(\boldsymbol{\theta}|y) - \ell(\boldsymbol{\theta}_0|y) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) - \Big[H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - H(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0)\Big]\,.$$

Zufällige Effekte: EM-Schätzer

Jensen Ungleichung für konkave Funktion ($g(x) = \log(x)$) liefert Abschätzung $\mathrm{E}(g(X)) \leq g(\mathrm{E}(X))$. Für beliebiges $\boldsymbol{\theta} \in \Theta$ folgt

$$
\begin{aligned}
H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - H(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) &= \int \log \frac{f(z|y,\boldsymbol{\theta})}{f(z|y,\boldsymbol{\theta}_0)} f(z|y,\boldsymbol{\theta}_0) dz \\
&= \mathrm{E}\Big(\log \frac{f(z|y,\boldsymbol{\theta})}{f(z|y,\boldsymbol{\theta}_0)} \Big| y, \boldsymbol{\theta}_0 \Big) \\
&\leq \log \mathrm{E}\Big(\frac{f(z|y,\boldsymbol{\theta})}{f(z|y,\boldsymbol{\theta}_0)} \Big| y, \boldsymbol{\theta}_0 \Big) \\
&= \log \int \frac{f(z|y,\boldsymbol{\theta})}{f(z|y,\boldsymbol{\theta}_0)} f(z|y,\boldsymbol{\theta}_0) dz \\
&= \log \int f(z|y,\boldsymbol{\theta}) dz = \log(1) = 0\,,
\end{aligned}
$$

also

$$
H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - H(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) \leq 0\,.
$$

Zufällige Effekte: EM-Schätzer

Dies hat zur Folge, dass wir Differenz schreiben können als

$$\ell(\boldsymbol{\theta}|y) - \ell(\boldsymbol{\theta}_0|y) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0)\,.$$

Sei $\boldsymbol{\theta}'$ jener Wert von $\boldsymbol{\theta}$, der $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ für gegebenes $\boldsymbol{\theta}_0$ maximiert. Somit gilt

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \geq 0 \text{ also auch } Q(\boldsymbol{\theta}'|\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) \geq 0\,.$$

Durch Maximierung von $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ wird Log-Likelihood zumindest nicht verkleinert, denn das letzte Ergebnis impliziert

$$\ell(\boldsymbol{\theta}'|y) - \ell(\boldsymbol{\theta}_0|y) \geq 0\,.$$

Zufällige Effekte: EM-Schätzer

Stationarität: Differenzieren von (✿) gibt

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) + \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|y)\,.$$

Nun gilt aber $H(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \leq H(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0)$ für alle $\boldsymbol{\theta} \in \Theta$. Unter dieser Extremalbedingung folgt

$$\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}|\boldsymbol{\theta}_0)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0\,.$$

$H(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ ist somit stationär in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Ist somit $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ stationär in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, dann ist dies dort auch $\ell(\boldsymbol{\theta}|y)$.

Zufällige Effekte: EM-Schätzer

EM-Algorithmus ist **zweistufig**.

E-Schritt: berechne bedingte Erwartung $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ für festes $\boldsymbol{\theta}_0$.

M-Schritt: maximiere $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ bezüglich $\boldsymbol{\theta}$.

Sei das Ergebnis dieser Maximierung $\boldsymbol{\theta}'$, so wird damit wieder E-Schritt mit aktualisierten $\boldsymbol{\theta}_0 = \boldsymbol{\theta}'$ berechnet.

Iteration bis Konvergenz im marginalen MLE $\hat{\boldsymbol{\theta}}$.

Zufällige Effekte: EM-Schätzer

Self-consistency des EM-Algorithmus: Falls MLE $\hat{\boldsymbol{\theta}}$ globales Maximum von $\ell(\boldsymbol{\theta}|y)$ darstellt, muss dieser auch

$$Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) \geq Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$$

genügen. Ansonsten würde es ja einen Parameterwert $\boldsymbol{\theta}^*$ geben mit der Eigenschaft

$$Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) < Q(\boldsymbol{\theta}^*|\hat{\boldsymbol{\theta}}),$$

was wiederum

$$\ell(\boldsymbol{\theta}^*|y) > \ell(\hat{\boldsymbol{\theta}}|y)$$

impliziert und somit einen Widerspruch darstellt zur Annahme, dass $\hat{\boldsymbol{\theta}}$ das globale Maximum von $\ell(\boldsymbol{\theta}|y)$ ist.

Zufällige Effekte: EM-Schätzer

Anstelle des Integrals $\ell(\boldsymbol{\theta}|y) = \log \int f(y, z|\boldsymbol{\theta})dz$ muss beim EM-Algorithmus das Integral $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ iterativ berechnet und maximiert werden.
Wie das folgende Beispiel zeigt, ist diese Berechnung in einigen Anwendungen möglich.

Zufällige Effekte: EM-Schätzer

Beispiel: Endliche diskrete Mischungen

Sei $y_1, \ldots, y_n$ Stichprobe aus Mischung von K Komponenten mit spezifischen Parametern $\theta_1, \ldots, \theta_K$ (z.B. Lokation der k-ten Gruppe), einem gemeinsamen Parameter ϕ (z.B. Variabilität – für alle Gruppen gleich) und Anteilen $\pi_1, \ldots, \pi_K$ mit $\sum_k \pi_k = 1$.

Marginale Dichte des i-ten Elements y_i daher

$$f(y_i|\boldsymbol{\theta}, \phi, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f(y_i|\theta_k, \phi)\,.$$

Likelihoodfunktion

$$\begin{aligned} L(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\mathbf{y}) &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i|\theta_k, \phi) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_{ik} \qquad \text{mit} \quad f_{ik} = f(y_i|\theta_k, \phi)\,. \end{aligned}$$

Zufällige Effekte: EM-Schätzer

Hier sind $\boldsymbol{\theta}$, ϕ und $\boldsymbol{\pi}$ unbekannt, also sind $K + 1 + (K - 1) = 2K$ Parameter zu schätzen.

Log-Likelihoodfunktion

$$\ell(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\mathbf{y}) = \log L(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\mathbf{y}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_{ik} \right)$$

Bezüglich ϕ erhält man

$$\begin{aligned} \frac{\partial}{\partial \phi} \ell(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\mathbf{y}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_k f_{ik}}{\sum_{l=1}^{K} \pi_l f_{il}} \frac{\partial \log f_{ik}}{\partial \phi} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \frac{\partial \log f_{ik}}{\partial \phi} \qquad \text{mit} \quad w_{ik} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K} \pi_l f_{il}} \, . \end{aligned}$$

Gewichte w_{ik} hängen von Parametern ab, also $w = w(\boldsymbol{\theta}, \phi, \boldsymbol{\pi})$.
Score ist gewichtete Summe von Komponenten-Scores.

Zufällige Effekte: EM-Schätzer

Gewichte w_{ik} auch aus Sicht eines formalen Bayes-Modells interpretierbar.

Da per Definition

$$\Pr(k|y_i) = \frac{\Pr(k)\Pr(y_i|k)}{\sum_{l=1}^{K}\Pr(l)\Pr(y_i|l)} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}} = w_{ik}$$

gilt, haben Gewichte folgende Bedeutung:

Wähle die Komponente k zufällig mit Wahrscheinlichkeit π_k. Ziehe y_i aus dieser Komponente, also aus einer Verteilung mit Dichte f_{ik}. Gegeben y_i ist die a posteriori Wahrscheinlichkeit, dass die Komponente k gewählt wurde, gleich w_{ik}.

Zufällige Effekte: EM-Schätzer

Bezüglich $\boldsymbol{\theta}$ folgt als Score

$$\frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}, \phi, \boldsymbol{\pi} | \mathbf{y}) = \sum_{i=1}^{n} \frac{\pi_k \partial f_{ik} / \partial \theta_k}{\sum_{l=1}^{K} \pi_l f_{il}} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log f_{ik}}{\partial \theta_k} ,$$

wiederum eine gewichtete Summe von Scoretermen.

Zufällige Effekte: EM-Schätzer

Bezüglich π benötigt man Randbedingung $\sum_k \pi_k = 1$. Mit Lagrange Multiplikator gibt dies

$$\frac{\partial}{\partial \pi_k}\left(\ell(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\mathbf{y}) - \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)\right) = \sum_{i=1}^{n}\frac{f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}} - \lambda$$
$$= \sum_{i=1}^{n} w_{ik}\frac{1}{\pi_k} - \lambda\,.$$

Nullsetzen gibt $\pi_k = \sum_i w_{ik}/\lambda$. Summieren über alle Komponenten liefert $1 = \sum_k \pi_k = \sum_i \sum_k w_{ik}/\lambda = n/\lambda$, also $\lambda = n$. Dies ergibt den Score

$$\frac{1}{\pi_k}\sum_{i=1}^{n} w_{ik} - n\,.$$

MLEs sind simultane Nullstellen aller 3 Scorefunktionen.

Zufällige Effekte: EM-Schätzer

Direkte Berechnung der MLEs meist ziemlich aufwendig, denn f_{ik} und w_{ik} sind komplexe Funktionen in sämtlichen Parametern.

Alternative Berechnung: EM-Schätzung
Betrachte y_i wegen unbekannter Gruppenzugehörigkeit als unvollständig.
Definiere zu y_i nichtbeobachtbaren Indikator $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$

$$z_{ik} = \begin{cases} 1 & \text{falls } y_i \text{ aus Gruppe } k \\ 0 & \text{sonst.} \end{cases}$$

gemeinsame Dichte von (y, z) somit

$$f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \phi, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i|\theta_k, \phi)^{z_{ik}} \pi_k^{z_{ik}} ,$$

also

$$\log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \phi, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) .$$

Zufällige Effekte: EM-Schätzer

E-Schritt: berechne für feste $\boldsymbol{\theta}_0$, ϕ_0, $\boldsymbol{\pi}_0$ Funktion $Q(\cdot|\cdot)$, d.h.

$$\mathrm{E}\Big(\log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \phi, \boldsymbol{\pi}) \Big| \mathbf{y}, \boldsymbol{\theta}_0, \phi_0, \boldsymbol{\pi}_0 \Big)$$

$$= \mathrm{E}\left(\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) \Big| y_i, \theta_0, \phi_0, \pi_0 \right) .$$

Durch Konditionieren sind im Argument des Erwartungswertes außer den z_{ik} nur feste Größen.
Wegen $\mathrm{E}(z_{ik}|y_i, \boldsymbol{\theta}_0, \phi_0, \boldsymbol{\pi}_0) = \Pr(z_{ik} = 1|y_i, \boldsymbol{\theta}_0, \phi_0, \boldsymbol{\pi}_0) = w_{ik} = w_{ik}(\boldsymbol{\theta}_0, \phi_0, \boldsymbol{\pi}_0)$ folgt

$$Q(\boldsymbol{\theta}, \phi, \boldsymbol{\pi}|\boldsymbol{\theta}_0, \phi_0, \boldsymbol{\pi}_0) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) .$$

$\boldsymbol{\theta}$, ϕ, $\boldsymbol{\pi}$ sind jene Größen, in denen im folgenden **M-Schritt** für (im Unterschied zur Berechnung der MLEs) **feste Gewichte** w_{ik} die Funktion Q maximiert wird.

Zufällige Effekte: EM-Schätzer

Beispiel: Mischung von Normalverteilungen

K Normalverteilungen mit Erwartungen μ_k und einheitlicher Varianz σ^2:

$$\log f_{ik} = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(y_i - \mu_k)^2}{2\sigma^2}\,.$$

Dafür folgt

$$\frac{\partial \log f_{ik}}{\partial \mu_k} = \frac{y_i - \mu_k}{\sigma^2}\,, \qquad \frac{\partial \log f_{ik}}{\partial \sigma^2} = -\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{(y_i - \mu_k)^2}{\sigma^4}\right).$$

Für geeignete Werte μ_k, σ^2 und π_k berechnet man im **E**-Schritt die Gewichte w_{ik}.

Zufällige Effekte: EM-Schätzer

Beispiel: Mischung von Normalverteilungen

Im **M**-Schritt mit diesen w_{ik} Maximierung durchführen, d.h.

$$\sum_{i=1}^{n} w_{ik} \frac{y_i - \mu_k}{\sigma^2} = 0 \implies \hat{\mu}_k = \frac{\sum_{i=1}^{n} w_{ik} y_i}{\sum_{i=1}^{n} w_{ik}} ,$$

sowie

$$\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \left(-\frac{1}{2\sigma^2} + \frac{(y_i - \mu_k)^2}{2\sigma^4} \right) = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} (y_i - \hat{\mu}_k)^2 .$$

Schließlich noch

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} w_{ik} .$$

Mit $\hat{\mu}_k$, $\hat{\sigma}^2$, $\hat{\pi}_k$ im nächsten **E**-Schritt w_{ik} aktualisieren und damit wieder **M**-Schritt durchführen.

Zufällige Effekte: Überdispersion

Zufälliger Effekt z mit Dichte $f(z)$.

Annahme: bedingte Dichte der Response y sei $f(y|z, \theta)$.

Somit gemeinsame Dichte $f(y, z|\theta) = f(y|z, \theta)f(z|\theta)$.

Gesucht: MLE zur marginalen Dichte der Response

$$f(y|\theta) = \int f(y, z|\theta)dz = \int f(y|z, \theta)f(z|\theta)dz\,.$$

Zufällige Effekte: Überdispersion

Damit folgt $f(z|y,\theta) = f(y|z,\theta)f(z|\theta)/f(y|\theta)$ und es resultiert

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) &= \sum_{i=1}^{n} \int \log f(y_i, z_i|\boldsymbol{\theta}) f(z_i|y_i, \boldsymbol{\theta}_0) dz_i \\ &= \sum_{i=1}^{n} \int \log f(y_i, z_i|\boldsymbol{\theta}) \frac{f(y_i|z_i, \boldsymbol{\theta}_0)}{f(y_i|\boldsymbol{\theta}_0)} f(z_i|\boldsymbol{\theta}_0) dz_i \\ &= \sum_{i=1}^{n} \frac{1}{f(y_i|\boldsymbol{\theta}_0)} \int \log f(y_i, z_i|\boldsymbol{\theta}) f(y_i|z_i, \boldsymbol{\theta}_0) f(z_i|\boldsymbol{\theta}_0) dz_i \,. \end{aligned}$$

Sowohl $f(y_i|\boldsymbol{\theta}_0)$ (ein Integral!) als auch das obige Integral selbst sind sehr unangenehm – nur selten analytisch geschlossen darstellbar!

Zufällige Effekte: Überdispersion

Zwei Ansätze bzgl. Handhabung der z_i:

- **Annahme**: $z_i \overset{iid}{\sim} \text{Normal}(0, \sigma_z^2)$ und beide Integrale durch K-Punkt **Gauss-Quadratur** approximieren.

- Falls Annahme nicht möglich, $f(z)$ durch **nicht-parametrischen Maximum-Likelihood (NPML)** Schätzer $\hat{f}(z)$ ersetzen und resultierende Zielfunktion maximieren.

Zufällige Effekte: Überdispersion

Normalverteilte zufällige Effekte:

Für $f(z_i) = \phi(z_i)$ liefert Gauss-Quadratur die Approximation

$$f(y_i|\boldsymbol{\theta}_0) = \int f(y_i|z_i, \boldsymbol{\theta}_0)\phi(z_i)dz_i \approx \sum_{k=1}^{K} f(y_i|z_k, \boldsymbol{\theta}_0)\pi_k$$

sowie

$$\int \log f(y_i, z_i|\boldsymbol{\theta}) f(y_i|z_i, \boldsymbol{\theta}_0) f(z_i|\boldsymbol{\theta}_0) dz_i \approx \sum_{k=1}^{K} \log f(y_i, z_k|\boldsymbol{\theta}) f(y_i|z_k, \boldsymbol{\theta}_0)\pi_k$$

mit "bekannten" Massen π_k auf bekannten Massepunkten z_k.

Zufällige Effekte: Überdispersion

Normalverteilte zufällige Effekte:

Damit resultiert

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \approx \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \log f(y_i, z_k|\boldsymbol{\theta}) f(y_i|z_k, \boldsymbol{\theta}_0)\pi_k}{\sum_{j=1}^{K} f(y_i|z_j, \boldsymbol{\theta}_0)\pi_j}$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Big(\log f(y_i|z_k, \boldsymbol{\theta}) + \log \pi_k\Big)$$

mit in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ ausgewerteten Gewichten

$$w_{ik} = \frac{\pi_k f(y_i|z_k, \boldsymbol{\theta}_0)}{\sum_{j=1}^{K} f(y_i|z_j, \boldsymbol{\theta}_0)\pi_j},$$

die somit für die folgende Maximierung feste Größen sind.

Zufällige Effekte: Überdispersion

Approximation durch Gauss-Quadratur liefert Funktion $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ wie beim Beispiel der diskreten Mischung!
Daher ist die Schätzung der Parameter im Modell mit zufälligen Effekten äquivalent der Schätzung in diskreten Mischmodellen.

Schreibe linearen Prädiktor als

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma_z z_i \,, \qquad \text{mit} \quad z_i \overset{iid}{\sim} \text{Normal}(0,1) \,.$$

Bei der Maximierung von $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ bezüglich $\boldsymbol{\beta}$ wird gewichtete Log-Likelihood einer "Stichprobe" mit nK Elementen maximiert.
Dazu ordne jedem y_i gerade K lineare Prädiktoren

$$\eta_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma_z z_k$$

mit bekannten Massen z_k und unbekannten $\boldsymbol{\beta}$ und σ_z zu.
D.h., jede Beobachtung K-fach betrachten und linearen Prädiktor jeweils mit z_k erweitern. Parameterschätzer zur neuen Spalte ist Schätzer von σ_z.

Zufällige Effekte: Überdispersion

Bei jeder EM-Iteration wird gewichtete ML-Schätzung eines GLMs berechnet. Als Daten verwendet man dazu die Struktur

y	w	β				σ_z
y_1	w_{11}	1	x_{11}	$\dots$	$x_{1,p-1}$	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{n1}	1	x_{n1}	$\dots$	$x_{n,p-1}$	z_1
y_1	w_{12}	1	x_{11}	$\dots$	$x_{1,p-1}$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{n2}	1	x_{n1}	$\dots$	$x_{n,p-1}$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_1	w_{1K}	1	x_{11}	$\dots$	$x_{1,p-1}$	z_K
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{nK}	1	x_{n1}	$\dots$	$x_{n,p-1}$	z_K

Zufällige Effekte: Überdispersion

NPML geschätzte zufällige Effekte:

Falls bzgl. $f(z)$ keine Annahmen möglich $\Rightarrow$ nichtparametrisch schätzen. Schätzer $\hat{f}(z)$ ist diskrete Wahrscheinlichkeitsfunktion auf K geschätzten Massepunkten z_k mit geschätzten Massen π_k.

Dadurch resultiert wie zuvor $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ als Zielfunktion, jetzt mit unbekannten z_k, π_k. Daher auch noch Schätzer $\hat{\pi}_k$ verwenden. Definiere linearen Prädiktor als

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + z_i \,, \qquad \text{mit} \quad z_i \overset{iid}{\sim} F(z) \,.$$

Bei NPML Schätzung wird daraus

$$\eta_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta} + z_k \,,$$

mit K unbekannten z_k. Mit K-stufigen Faktor umschreiben zu

$$\eta_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta} + z_2 \cdot 0 + \cdots + z_{k-1} \cdot 0 + z_k \cdot 1 + z_{k+1} \cdot 0 + \cdots + z_K \cdot 0 \,.$$

$z_2, \ldots, z_K$ sind "Parameter" zum $(K-1)$-elementigen Indikator. Intercept in $\mathbf{x}$ ist Parameter zu z_1 (Referenzklasse).

Zufällige Effekte: Überdispersion

Wieder erweiterte Daten mit nK Elementen. Für M-Schritt gewichtete ML-Schätzung eines GLMs berechnen, basierend auf

y	w	$\boldsymbol{\beta}$				$\mathbf{z}$			
y_1	w_{11}	1	x_{11}	$\dots$	$x_{1,p-1}$	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n1}	1	x_{n1}	$\dots$	$x_{n,p-1}$	0	0	$\dots$	0
y_1	w_{12}	1	x_{11}	$\dots$	$x_{1,p-1}$	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n2}	1	x_{n1}	$\dots$	$x_{n,p-1}$	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_1	w_{1K}	1	x_{11}	$\dots$	$x_{1,p-1}$	0	0	$\dots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{nK}	1	x_{n1}	$\dots$	$x_{n,p-1}$	0	0	$\dots$	1

Im folgenden E-Schritt werden Gewichte w_{ik} aktualisiert. Iteration bis Konvergenz eintritt (im NPMLE).

Zufällige Effekte: Überdispersion

Als nichtparametrischen Maximum-Likelihood (NPML) Schätzer der Dichte der Zufallseffekte erhält man bei Konvergenz die K Paare

$$\hat{f}(z) = (\hat{z}_1, \hat{\pi}_1), \ldots, (\hat{z}_K, \hat{\pi}_K)\,.$$

Dies entspricht einer Multinomialverteilung auf den geschätzten K Massestellen $\hat{\mathbf{z}}$ mit geschätzten Wahrscheinlichkeitsmassen $\hat{\boldsymbol{\pi}}$.

Zufällige Effekte: Überdispersion

Prädiktionen bei NPML Schätzung:
Schätzung für das **konditionale Modell** $\mu_{ik} = \mathrm{E}(y_i|z_k)$

$$g(\hat{\mu}_{ik}) = \hat{\eta}_{ik} = x_i^\top \hat{\beta} + \hat{z}_k \quad \text{also} \quad \hat{\mu}_{ik} = g^{-1}(\hat{\eta}_{ik})\,.$$

Das sind parallele Regresionsebenen im (η, x)-Raum.

Für Schätzung des **marginalen Modells** $\mathrm{E}(y_i)$, wobei

$$\begin{aligned}\mathrm{E}(y_i) &= \int y_i \int f(y_i|z_i) f(z_i)\, dz_i\, dy_i \\ &\approx \int y_i \sum_{k=1}^{K} f(y_i|z_k)\pi_k\, dy_i = \sum_{k=1}^{K} \pi_k \int y_i f(y_i|z_k)\, dy_i = \sum_{k=1}^{K} \pi_k \mathrm{E}(y_i|z_k)\end{aligned}$$

resultiert der Schätzer

$$\widehat{\mathrm{E}}(y_i) = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_{ik}\,.$$

Zufällige Effekte: Überdispersion

Posterior Mean von z_i, gegeben Response y_i, ist

$$\begin{aligned}
\mathrm{E}(z_i|y_i) &= \int z_i f(z_i|y_i) dz_i \\
&= \int z_i \, \frac{f(y_i|z_i) f(z_i)}{\int f(y_i|z_i) f(z_i) dz_i} \, dz_i \\
&\approx \sum_{k=1}^{K} z_k \, \frac{f(y_i|z_k)\pi_k}{\sum_{l=1}^{K} f(y_i|z_l)\pi_l} \, .
\end{aligned}$$

Empirischer Bayes Schätzer dafür ist

$$\widetilde{\mathrm{E}}(z_i|y_i) = \sum_{k=1}^{K} \hat{z}_k \hat{w}_{ik} \, ,$$

das posteriori-gewichtete Mittel der geschätzten Massepunkte.

Zufällige Effekte: Überdispersion

Damit ist es jetzt möglich, den empirischen Bayes Schätzer für den marginalen Erwartungswert zu berechnen. Wegen $\sum_k \hat{w}_{ik} = 1$ resultiert als Schätzer für den linearen Prädiktor

$$\widetilde{\eta}_i = x_i^\top \hat{\beta} + \widetilde{\mathrm{E}}(z_i|y_i) = \sum_{k=1}^{K} \hat{w}_{ik} \left(x_i^t \hat{\beta} + \hat{z}_k \right)$$

$$= \sum_{k=1}^{K} \hat{w}_{ik} \hat{\eta}_{ik}$$

und damit folgt wiederum

$$\widetilde{\mu}_i = g^{-1}(\widetilde{\eta}_i) \,.$$

Zufällige Effekte: Überdispersion

Beispiel: Matched Pairs

Bakterienkonzentrationen in der Luft bei Siedlungen ohne/mit Kompostieranlage (`site=6`/`site=7`) .
Anzahl kolonienbildender Mikroorganismen (cfu: colonies forming units) hängt vom Wetter ab.
26×2 Responses `bac`, Temperatur `temp` und Luftfeuchte `humi`.
Luftkeimsammler mit anschließender Auswertung der Proben im Labor.

6-stufiger Andersen Kaskadenimpaktor (Stufen filtern unterschiedlich große Keime). Für Menschen sind kleine Mikroorganismen (`stage` > 3) gefährlich (können bis Lunge vordringen).

Untersuche, ob Kompostieranlageneffekt relevant.

Zufällige Effekte: Überdispersion

Details zur Datenmanipulation siehe Skript!

Gesucht: Modell für `cfu` abhängig von `site`, `stage`, `temp`, `humi`.

Da Anzahlen (cfu's), **loglineares Poissonmodell** betrachten.
Interesse am Verhalten in beiden Orten auf unteren Stufen.
Zeige: Luftfeuchtigkeit zusätzlich zur Temperatur nicht relevant.

```
> g.t<-glm(cfu ~ stage*site + temp + I(temp^2), family=poisson)
> g.th<-update(g.t, . ~ . + humi + I(humi^2))

> anova(g.t, g.th, test="Chisq")
Analysis of Deviance Table

Model 1: cfu ~ stage*site + temp + I(temp^2)
Model 2: cfu ~ stage*site + temp + I(temp^2) + humi + I(humi^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       142     258.83
2       140     255.09  2    3.746   0.1537
```

Zufällige Effekte: Überdispersion

```
> summary(g.t)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3446015  0.1877857   1.835 0.066494 .
stage5        0.3976830  0.1971814   2.017 0.043712 *
stage6       -0.2954642  0.2334642  -1.266 0.205669
site7         0.0596360  0.2121516   0.281 0.778633
temp          0.0596588  0.0175674   3.396 0.000684 ***
I(temp^2)    -0.0021462  0.0006126  -3.504 0.000459 ***
stage5:site7 -0.4421347  0.2886983  -1.531 0.125652
stage6:site7  0.4180665  0.3089904   1.353 0.176053
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 284.23  on 149  degrees of freedom
Residual deviance: 258.83  on 142  degrees of freedom
AIC: 575.79
```

Zufällige Effekte: Überdispersion

Bedenklich ist die Überdispersion (Deviance 258.83 bei df=142).

Fehlen weitere Prädiktoren, z.B. Wind oder Luftdruck?

Modell mit zufälligen Effekten könnte dies kompensieren.

$\Rightarrow$ Modell mit zufälligem Intercepts aus Normalverteilung.

$\Rightarrow$ Paket `npmlreg` von J. Einbeck, R. Darnell und J. Hinde

Funktion `alldist` erlaubt Modellierung mit zufälligen Effekten.

`random = ~1` spezifiziert, dass Intercept zufällig ist.

Um Fehler durch Gauß-Hermite Quadratur gering zu halten, verwende sehr viele Quadraturpunkte (`k=100`).

Bemerke, dass die Option `data=` zwingend erforderlich ist.

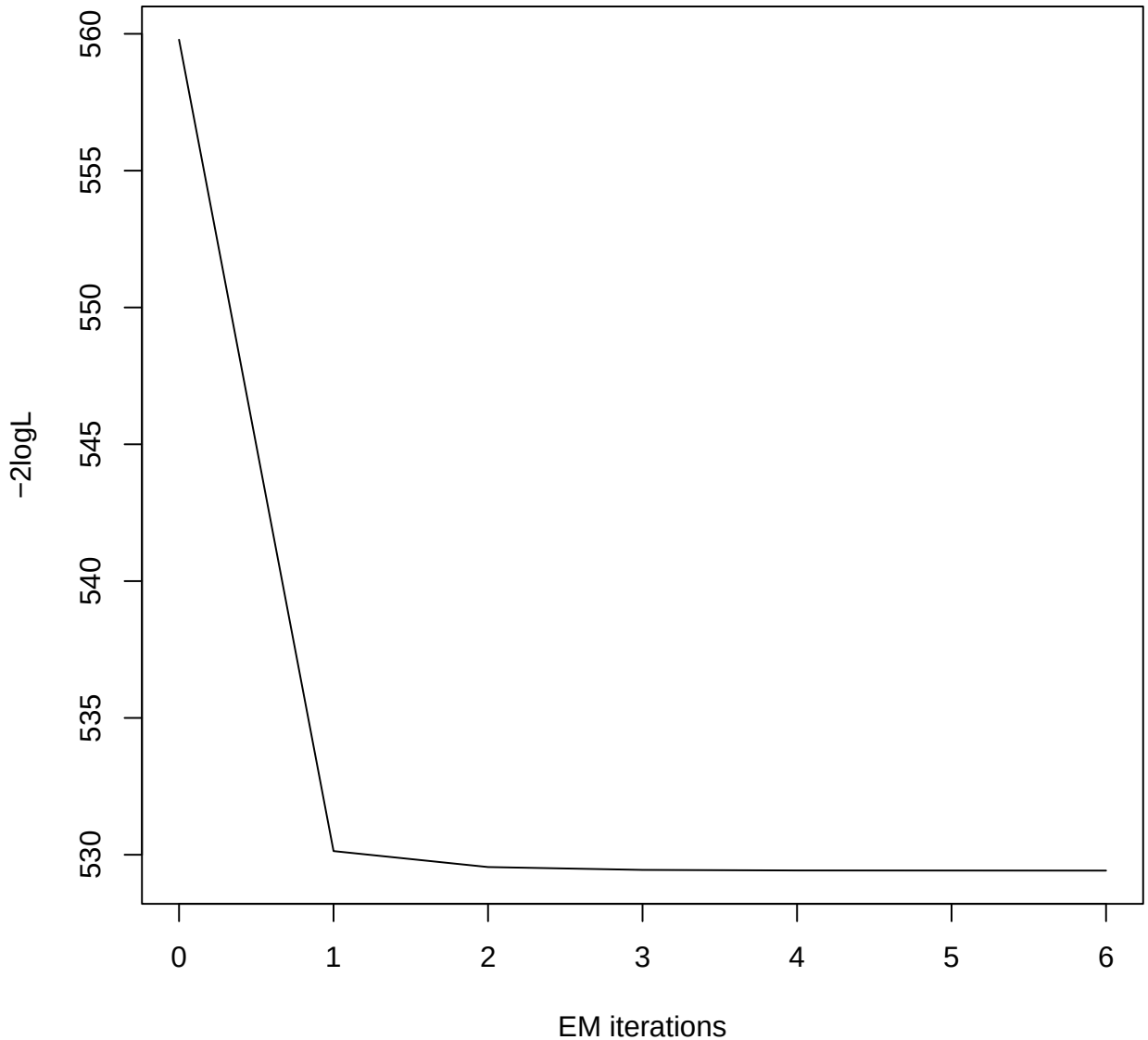
Zufällige Effekte: Überdispersion

```
> library(npmlreg)
> gq <- alldist(cfu ~ stage*site + temp + I(temp^2), data=bac,
+       random=~1,family=poisson,random.distribution="gq",k=100)
> summary(gq)
Coefficients:
             Estimate Std. Error t value
(Intercept)   0.17137   0.188758   0.908
stage5        0.30847   0.198028   1.558
stage6       -0.31104   0.233637  -1.331
site7         0.07377   0.212191   0.348
temp          0.06168   0.017711   3.483
I(temp^2)    -0.00218   0.000626  -3.493
stage5:site7 -0.36474   0.289138  -1.261
stage6:site7  0.40041   0.309218   1.295
z             0.58875   0.058675  10.034

Random effect distribution - standard deviation:          0.589
-2 log L:              529.4      Convergence at iteration  6
```

Zufällige Effekte: Überdispersion

Standardfehler und t-Werte stammen nur aus der letzten Iteration beim EM-Algorithmus (gewichtetes GLM bezüglich der erweiterten Datenstruktur) und sind nicht direkt interpretierbar.

Zufällige Effekte: Überdispersion



Disparität:
$-2 \log L(\hat{\theta}|y)$

Zufällige Effekte: Überdispersion

NPML Schätzung:

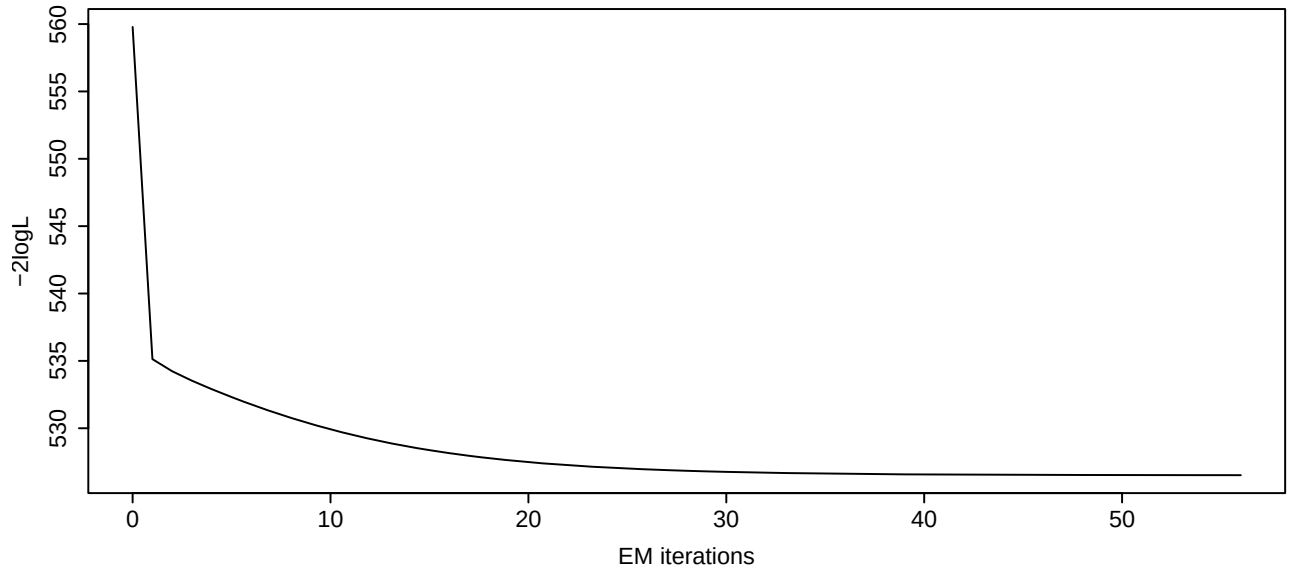
- $k = 1$ Massepunkt: entspricht dem loglinearen Poissonmodell!
- $k = 2$ Massepunkte erlaubt zusätzliche Variabilität im Prädiktor.

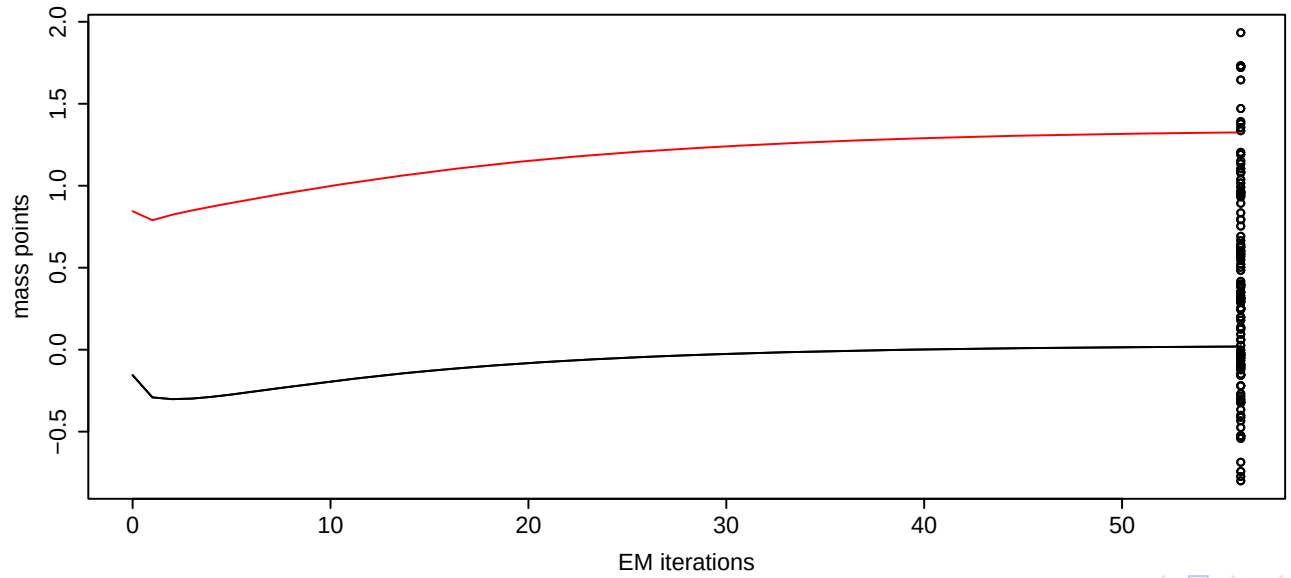
```
> np2 <- alldist(cfu ~ stage*site + temp + I(temp^2),
+                data=bac, random = ~1, family=poisson,
+                random.distribution = "np", k=2)

1 ..2 ..3 ..4 ..5 ..6 ..7 ..8 ..9 ..10 ..11 ..12 ..13 ..14 ..
EM algorithm met convergence criteria at iteration #  56
Disparity trend plotted.
EM Trajectories plotted
```

Zufällige Effekte: Überdispersion

```
> summary(np2)
Coefficients:
             Estimate Std. Error t value
stage5        0.39420   0.197224  1.9987
stage6       -0.25655   0.233736 -1.0976
site7         0.14508   0.212445  0.6829
temp          0.05846   0.017581  3.3254
I(temp^2)    -0.00205   0.000621 -3.3042
stage5:site7 -0.50573   0.288828 -1.7510
stage6:site7  0.23819   0.310004  0.7684
MASS1         0.01905   0.191331  0.0996
MASS2         1.32524   0.200956  6.5947

Mixture proportions:
 MASS1   MASS2
0.8702  0.1298
Random effect distribution - standard deviation:          0.439
-2 log L:              526.5     Convergence at iteration  56
```

Zufällige Effekte: Überdispersion





Zufällige Effekte: Überdispersion

- Disparität: nach gut 20 Iterationen endgültige Größenordnung.
- EM Trajektorien: stabile Verläufe der Massepunkte Schätzer.
- Punkte am rechten Rand stellen Residuen (fixed part residuals) auf Skala der linearen Prädiktoren dar, d.h.

$$r_i = g(y_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \,.$$

Bewertet Lage dieser Massepunkte zu den Daten.

- nur etwas geringere Disparität als bei normalverteilten Effekten.
- beide Massepunkte sind Intercepts $(0.02, 1.33)$ mit Massen $(0.87, 0.13)$ (deutet auf Abweichung von Normalverteilung hin).
- NPMLE: zweipunktige (diskrete) Verteilung mit Momenten

```
> (e <- sum(np2$masses * np2$mass.p))
[1] 0.189
> sqrt(sum(np2$masses * (np2$mass.p - e)^2))
[1] 0.439
```

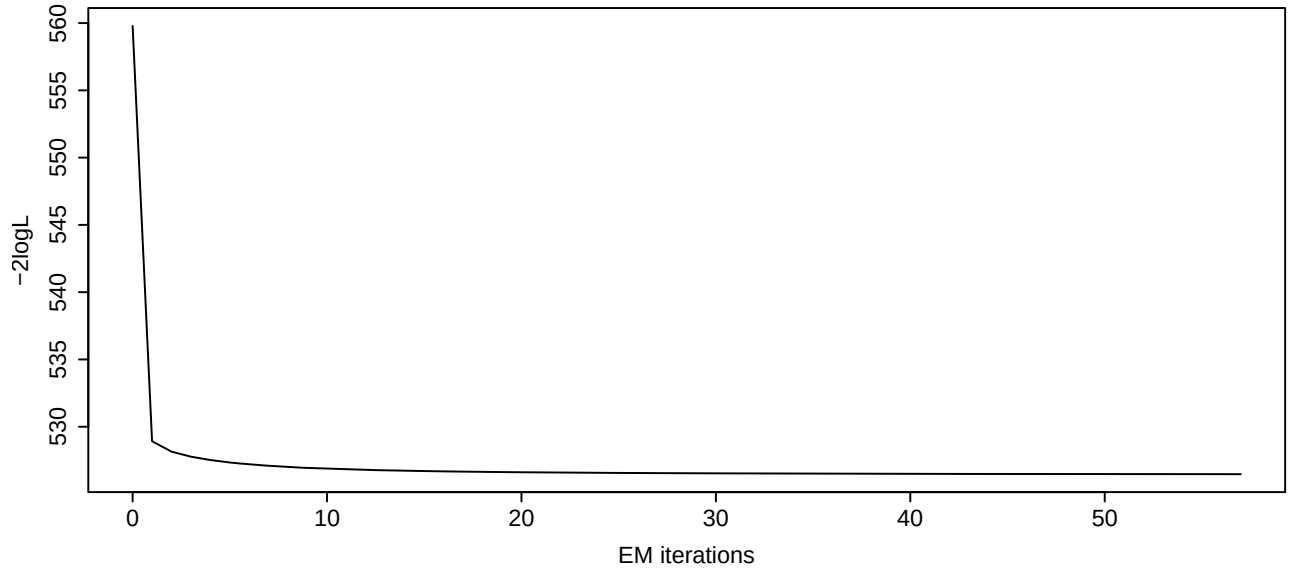
- Temperaturparameter $(0.0585, -0.0021)$ direkt vergleichbar mit Ergebnissen bei Gauß-Quadratur $(0.0617, -0.0022)$.

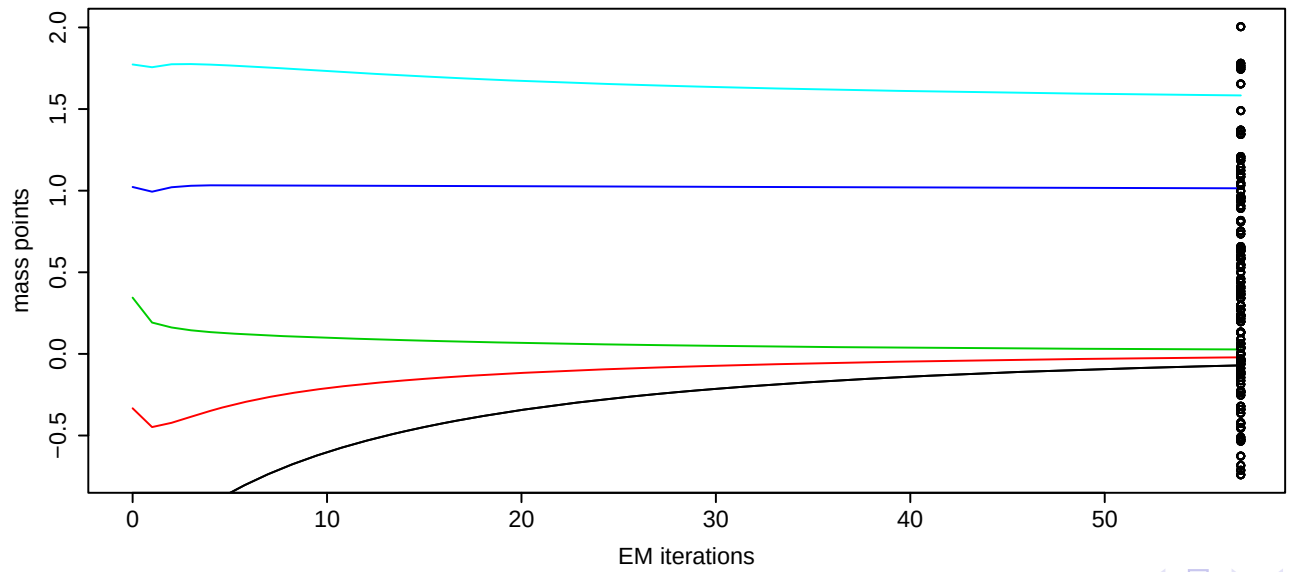
Zufällige Effekte: Überdispersion

```
> np3 <- alldist(cfu ~ stage*site+temp+I(temp^2), ... , k=3)
> np4 <- alldist(cfu ~ stage*site+temp+I(temp^2), ... , k=4)
> np5 <- alldist(cfu ~ stage*site+temp+I(temp^2), ... , k=5)

> np3$deviance
[1] 225.583
> np4$deviance
[1] 225.609
> np5$deviance
[1] 225.52
```

Deviance bleibt ziemlich konstant. Bemerke, dass Freiheitsgrade pro zusätzlichen Massepunkt um 2 (1 Massepunkt und 1 Masse) reduziert werden.
Wir bleiben bei Modell `np2` mit 2 Massepunkten.
Was passiert bei Erhöhung dieser Anzahl auf $K = 5$?

Zufällige Effekte: Überdispersion





Zufällige Effekte: Überdispersion

Kleinsten 3 Massepunkte nähern sich über Iterationen hinweg stark dem Wert Null an.

```
> np5$mass.points
   MASS1    MASS2    MASS3    MASS4    MASS5
-0.07033 -0.02099  0.02725  1.01455  1.58343
> np5$masses
  MASS1   MASS2   MASS3   MASS4   MASS5
0.01067 0.25720 0.56885 0.11513 0.04815
> sum(np5$masses[1:3])
[1] 0.8367
```

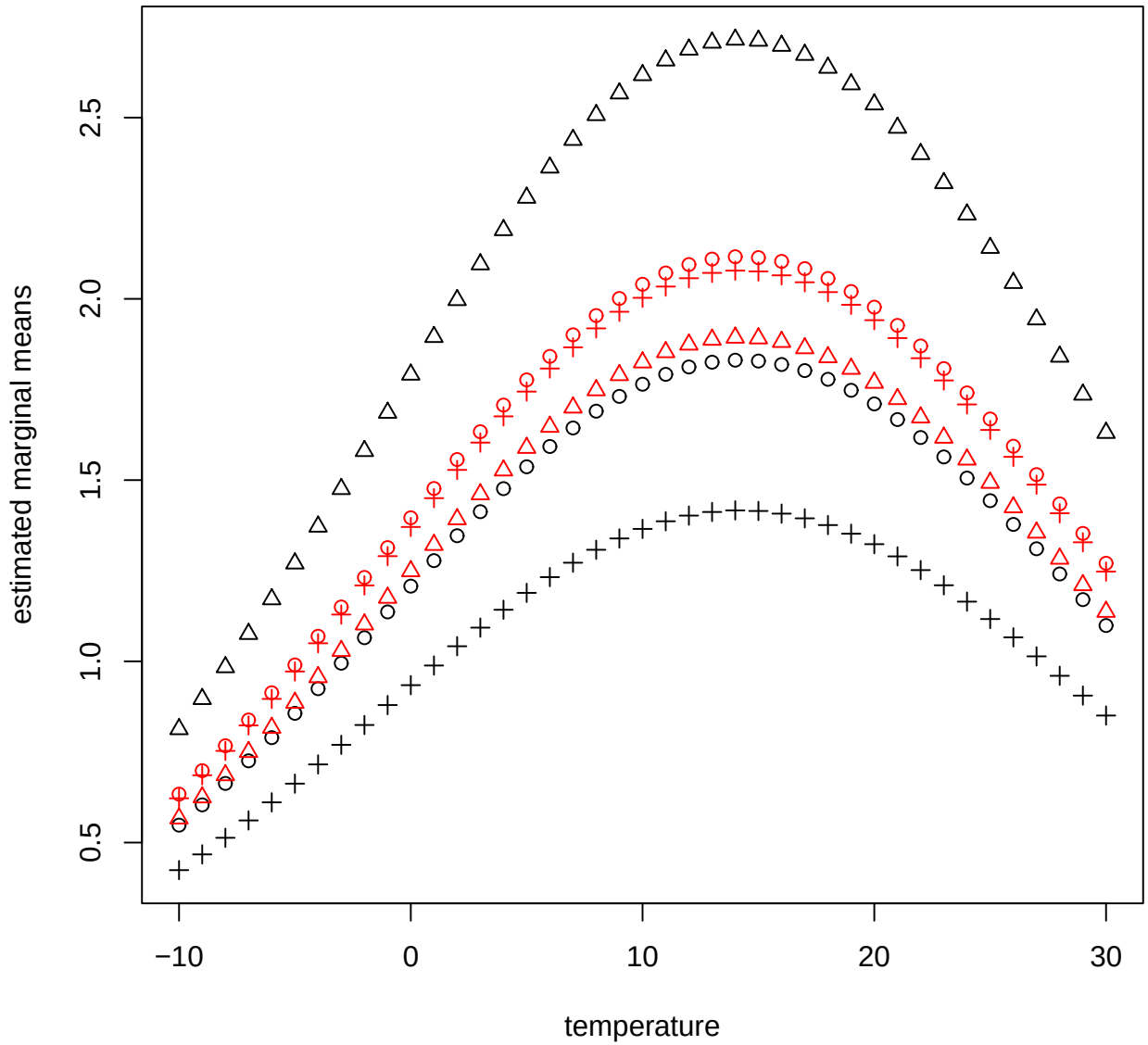
Summe der ersten 3 Massen ist 0.84 und somit ähnlich der ersten Masse 0.87 im Modell `np2`.
Beide anderen Punkte sind etwas kleiner und etwas größer als der zweite Punkt (1.3) im Modell `np2`.

Zufällige Effekte: Überdispersion

Geschätzte marginale Erwartungswerte:

```
> t <- -10:30
> newobs <- expand.grid(list(stage=levels(stage),
+                            site=levels(site), temp=t))

# estim. marginal means
> emm <- predict(np2, newdata=newobs, type="response")

> plot(newobs[ ,"temp"], emm, xlab="temperature",
+      ylab="estimated marginal means",
+      col=newobs[ ,"site"], pch=as.numeric(newobs[ ,"stage"]))
```

Zufällige Effekte: Überdispersion



Marginale Modelle abhängig von Temperatur, Ort mit(rot) und ohne(schwarz) Kompostieranlage, und Stufen 4 (○), 5 (△), 6 (+).

Zufällige Effekte: Überdispersion

Diagnostic Plot: Residuen gegen posteriori Gewichte.

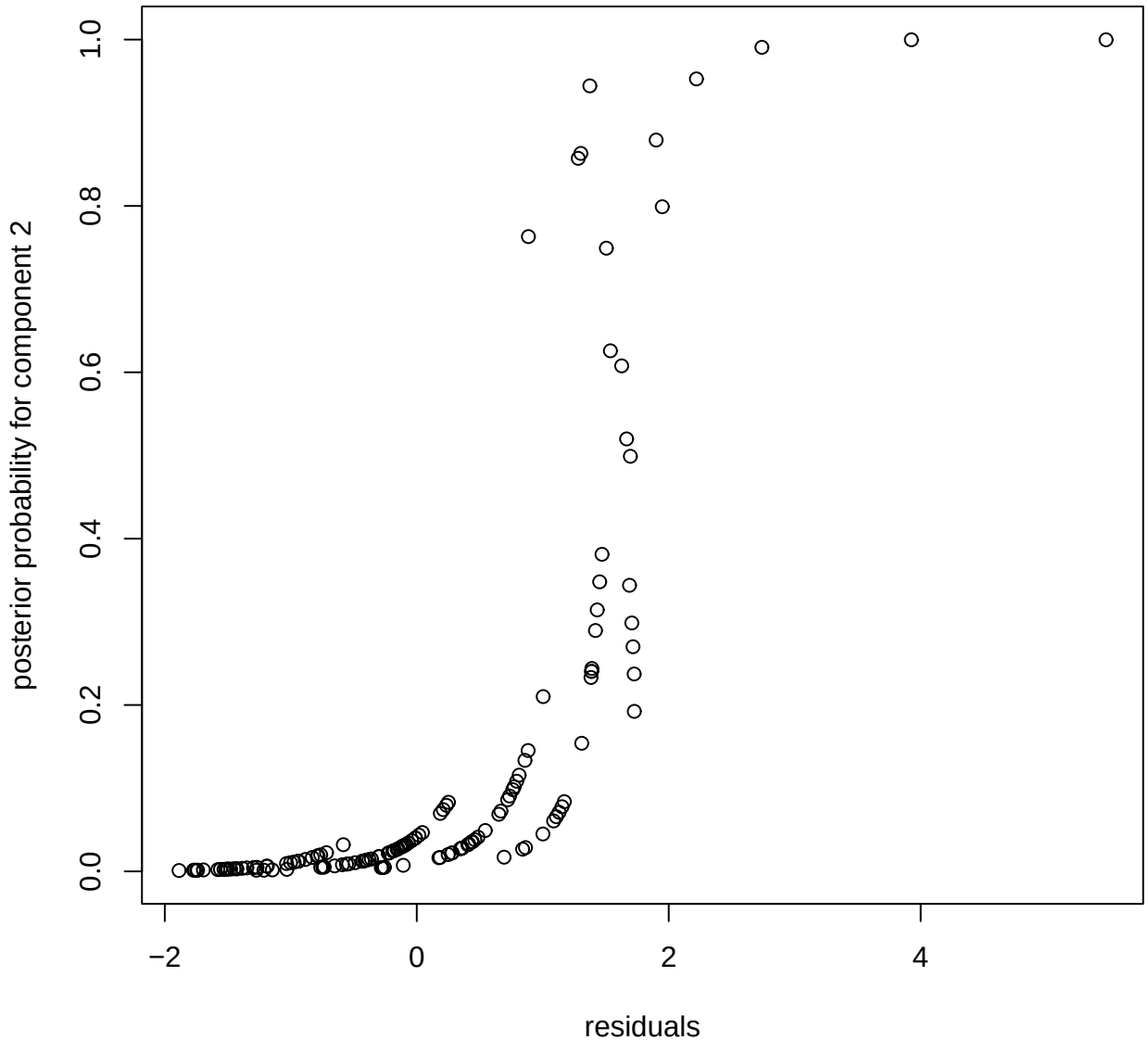
Residuen sind Differenzen zwischen beobachteten Responses und deren empirischen Bayesprädiktoren.

```
> plot(np2$residuals, np2$post.prob[ , 2], xlab="residuals",
+      ylab="posterior probability for component 2")
```

- Wenige Responses mit großen Residuen haben größere posteriori Wahrscheinlichkeiten für zweite Komponente ($w_{i2} > 0.5$).

- Für Mehrzahl gilt $w_{i1} > w_{i2}$ (Zugehörigkeit zu Komponente 1).

Zufällige Effekte: Überdispersion

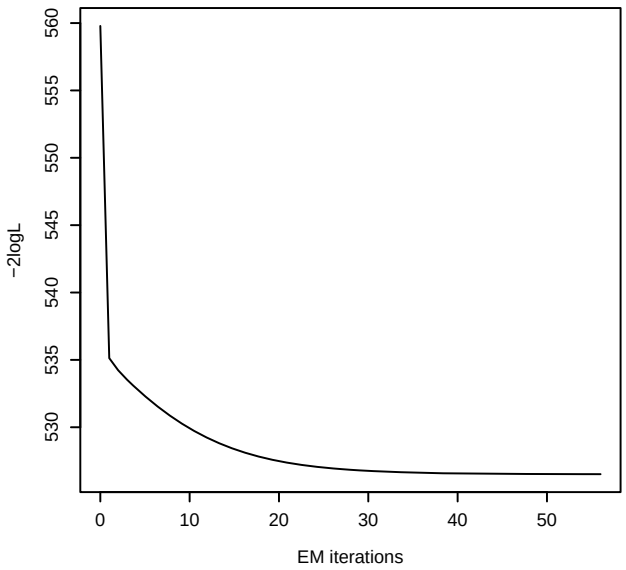


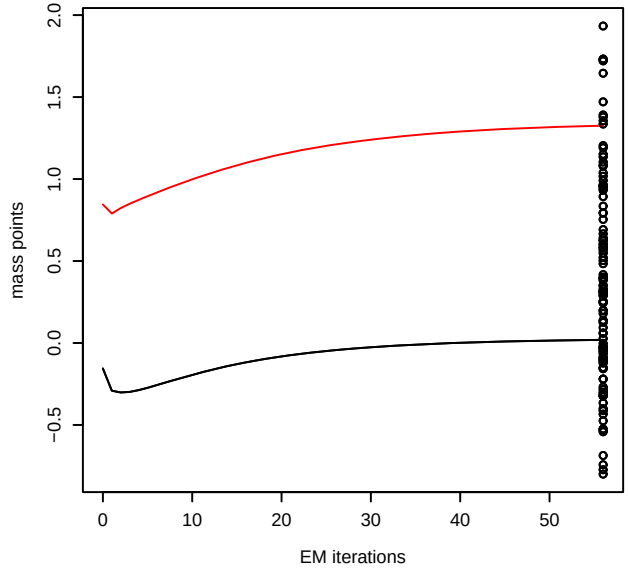
Zufällige Effekte: Überdispersion

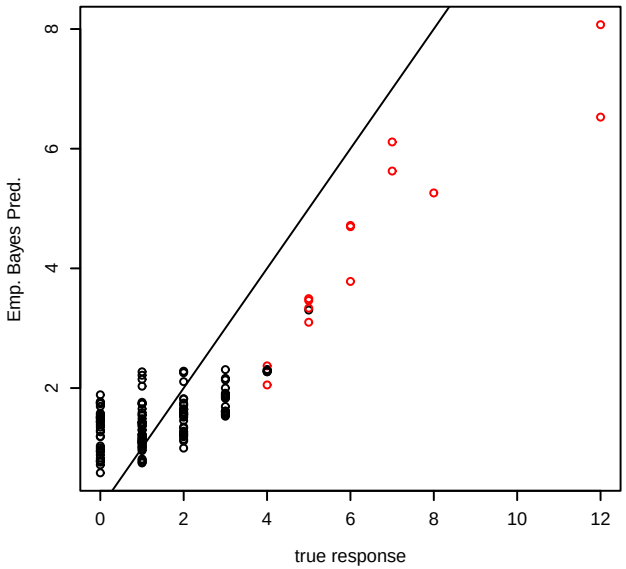
Etwas anders sieht diese Abbildung aus, wenn man sie durch Aufruf von `plot` erzeugt. Hierbei sind die Residuen sogenannte fixed part residuals ($y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$), also ohne Einbeziehen des Interceptterms).

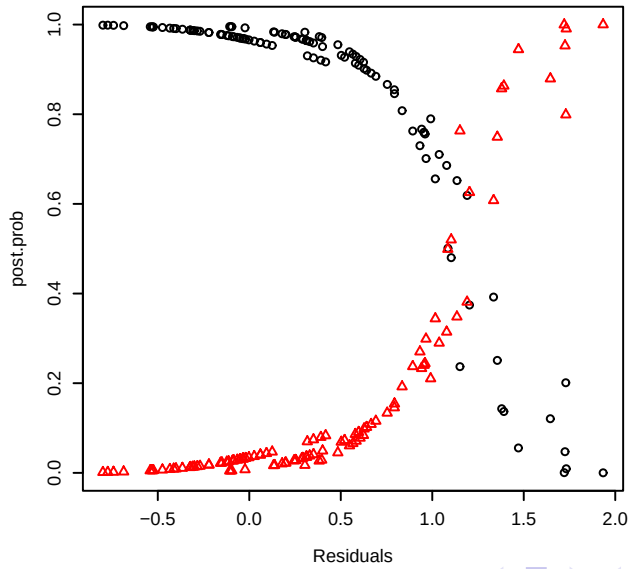
```
> plot(np2) # default plot.opt=15 (for details see manual)
```

Zufällige Effekte: Überdispersion



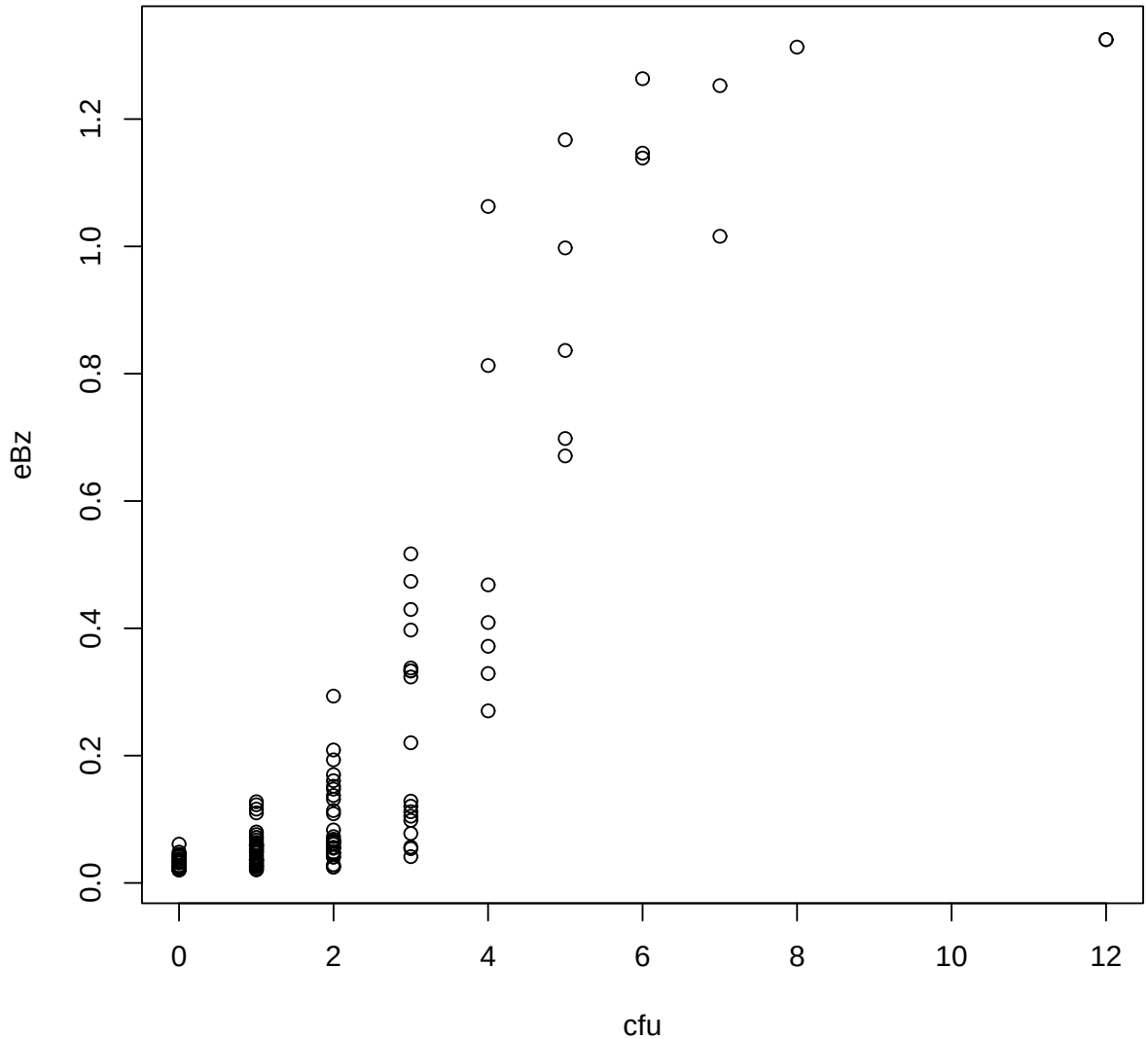






Zufällige Effekte: Überdispersion

```
> eBz <- np2$post.prob %*% np2$mass.points # equals np2$post.int
> plot(cfu, eBz)
```



Zufällige Effekte: Überdispersion

• Empirische Bayes Schätzer der z_i (posterior intercepts) hängen von den y_i ab.

• Deren Werte sind dann groß, wenn responses groß.

```
> mean(eBz)
[1] 0.1886
```

• Ihr globales Mittel 0.1886 entspricht Intercept beim Modell mit Gauss-Quadratur (0.1714) oder dem geschätzten Erwartungswert des zufälligen Effektes beim NPMLE Ansatz (0.1886).