Generalisierte Lineare Modelle

Herwig FRIEDL

Institut für Statistik
Technische Universität Graz

Oktober 2014

Inhaltsverzeichnis

Kapitel 1

Transformation auf Normalverteilung

Die statistische Analyse von Daten basiert häufig auf der Annahme, dass diese normalverteilt sind und konstante Varianz widerspiegeln. Falls die Daten diese Annahme nicht unterstützen, besteht die Möglichkeit der Verwendung einer Transformation, um dadurch eine bessere Approximation zu einer konstanten Varianz zu erzielen. Dann könnten auch klassische Methoden wie die Varianzanalyse oder die Lineare Regression auf solche Daten angewendet werden.

1.1 Box-Cox Transformationsfamilie

Die Verwendbarkeit der Normalverteilung wird erweitert, indem diese in eine größere Familie von Verteilungsfunktionen eingebettet wird, der Box-Cox Transformationsfamilie (Box und Cox, 1964). Deren allgemeine Form kann für eine positive Response $y > 0$ repräsentiert werden durch die Transformationsfunktion

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{falls } \lambda \neq 0, \\ \log y, & \text{falls } \lambda = 0, \end{cases} \tag{1.1}$$

wobei λ den Parameter der Transformation bezeichnet. Spezialfälle in dieser Familie sind $y(-1) = 1 - 1/y$ und $y(+1) = y - 1$. Darüberhinaus strebt für $\lambda \to 0$, $y(\lambda) \to \log y$, so dass $y(\lambda)$ eine stetige Funktion in λ ist.

Für Daten (y_i, x_i), $i = 1, \ldots, n$, nehmen wir nun an, dass es genau einen Wert von λ gibt, für den alle $y_i(\lambda)$ einer Normalverteilung mit konstanter Varianz genügen, d.h.

$$y_i(\lambda) \overset{ind}{\sim} \text{Normal}(\mu_i(\lambda), \sigma^2(\lambda)).$$

Unter dieser Annahme kann man auf die Dichtefunktion einer originalen Beobachtung y

schließen. Mit dem Transformationssatz für Dichten ist diese gerade

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}}\exp\left(-\frac{(y(\lambda)-\mu(\lambda))^2}{2\sigma^2(\lambda)}\right)\left|\frac{d}{dy}y(\lambda)\right|.$$

Die Transformationsfamilie (1.1) liefert dafür

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2(\lambda)}}\exp\left(-\dfrac{\left((y^\lambda-1)/\lambda-\mu(\lambda)\right)^2}{2\sigma^2(\lambda)}\right)y^{\lambda-1}, & \text{falls } \lambda\neq 0, \\ \dfrac{1}{\sqrt{2\pi\sigma^2(\lambda)}}\exp\left(-\dfrac{(\log y-\mu(\lambda))^2}{2\sigma^2(\lambda)}\right)y^{-1}, & \text{falls } \lambda = 0. \end{cases} \tag{1.2}$$

In der Regressionsanalyse verwendet man häufig ein lineares Modell der Form

$$y_i(\lambda) \overset{ind}{\sim} \text{Normal}(x_i^\top\beta(\lambda),\sigma^2(\lambda)),$$

wobei $\beta(\lambda) = (\beta_0(\lambda),\beta_1(\lambda),\ldots,\beta_{p-1}(\lambda))^\top$ den $p\times 1$ Vektor der unbekannten Parameter bezeichnet und $x_i = (1,x_{i1},\ldots,x_{i,p-1})^\top$ den $p\times 1$ Vektor mit den erklärenden Größen zur i-ten Beobachtung darstellt.
Für $\lambda\neq 0$ ergibt sich unter diesem Modell als Dichtefunktion einer Beobachtung y gerade

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

Somit ist es naheliegend, den reparameterisierten Vektor $\beta = (\beta_0,\ldots,\beta_{p-1})^\top$ bezüglich y^λ, mit Intercept $\beta_0 = 1+\lambda\beta_0(\lambda)$ und Slopeparametern $\beta_j = \lambda\beta_j(\lambda)$, $j = 1,\ldots,p-1$, sowie $\sigma^2 = \lambda^2\sigma^2(\lambda)$ zu definieren. Damit kann die Dichte (1.2) umgeschrieben werden zu

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

Ist hingegen $\lambda = 0$, so verwenden wir $\beta_j = \beta_j(\lambda)$, $j = 0,1,\ldots,p-1$, und $\sigma^2 = \sigma^2(\lambda)$ als Parameter bezüglich eines linearen Modells für $\log y$. Damit wird (1.2) zu

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

1.2 Maximum-Likelihood Schätzung

Liegen nun n unabhängige Beobachtungen (y_i, x_i) vor, die den Annahmen bei einer Box-Cox Transformation genügen, so ist deren Dichtefunktion gegeben durch

$$f(y_i|\lambda,\beta,\sigma^2) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y_i^\lambda - x_i^\top\beta)^2}{2\sigma^2}\right)|\lambda| y_i^{\lambda-1}, & \text{falls } \lambda \neq 0, \\ \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(\log y_i - x_i^\top\beta)^2}{2\sigma^2}\right) y_i^{-1}, & \text{falls } \lambda = 0. \end{cases}$$

Die Log-Likelihood Funktion ist daher

$$\begin{aligned} \ell(\lambda,\beta,\sigma^2|y) &= \sum_{i=1}^n \log f(y_i|\lambda,\beta,\sigma^2) \\ &= \begin{cases} -\dfrac{n}{2}\log\left(2\pi\sigma^2\right) - \dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^n \left(y_i^\lambda - x_i^\top\beta\right)^2 + n\log|\lambda| + (\lambda-1)\sum_{i=1}^n \log y_i, & \text{falls } \lambda \neq 0, \\ -\dfrac{n}{2}\log\left(2\pi\sigma^2\right) - \dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^n \left(\log y_i - x_i^\top\beta\right)^2 - \sum_{i=1}^n \log y_i, & \text{falls } \lambda = 0. \end{cases} \end{aligned} \tag{1.3}$$

Für einen festen Wert von λ lösen die Maximum-Likelihood Schätzer $\hat\beta_\lambda$ und $\hat\sigma^2_\lambda$ basierend auf (1.3) die Schätzgleichungen

$$\begin{aligned} \frac{\partial\ell(\lambda,\beta,\sigma^2|y)}{\partial\beta} &= \begin{cases} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n x_i(y_i^\lambda - x_i^\top\beta) = 0, & \text{falls } \lambda \neq 0, \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n x_i(\log y_i - x_i^\top\beta) = 0, & \text{falls } \lambda = 0, \end{cases} \\ \frac{\partial\ell(\lambda,\beta,\sigma^2|y)}{\partial\sigma^2} &= \begin{cases} -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (y_i^\lambda - x_i^\top\beta)^2 = 0, & \text{falls } \lambda \neq 0, \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (\log y_i - x_i^\top\beta)^2 = 0, & \text{falls } \lambda = 0\,. \end{cases} \end{aligned}$$

Mit der $n \times p$ Designmatrix $X = (x_1, \ldots, x_n)^\top$ folgt deshalb sofort

$$\begin{aligned} \hat\beta_\lambda &= \begin{cases} (X^\top X)^{-1} X^\top y^\lambda, & \text{falls } \lambda \neq 0, \\ (X^\top X)^{-1} X^\top \log y, & \text{falls } \lambda = 0, \end{cases} \\ \hat\sigma^2_\lambda &= \frac{1}{n}\mathrm{SSE}_\lambda(\hat\beta_\lambda) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^n (y_i^\lambda - x_i^\top\hat\beta_\lambda)^2, & \text{falls } \lambda \neq 0, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^n (\log y_i - x_i^\top\hat\beta_\lambda)^2, & \text{falls } \lambda = 0, \end{cases} \end{aligned}$$

wobei y^λ (resp. $\log y$) elementeweise gerechnet sind und $\mathrm{SSE}_\lambda(\hat\beta_\lambda)$ die Fehlerquadratsumme von y^λ (resp. $\log y$) an der Stelle $\hat\beta_\lambda$ für ein festes λ bezeichnet. Bemerke, dass wegen der obigen Reparameterisierung die Fehlerquadratsumme $\mathrm{SSE}_\lambda(\hat\beta_\lambda)$ in $\lambda = 0$ unstetig ist.

Substituiert man die Parameter (β, σ^2) in der Log-Likelihood Funktion (1.3) durch die beiden Schätzer $(\hat\beta_\lambda, \hat\sigma^2_\lambda)$ und lässt darin alle konstanten (von λ unabhängigen) Terme weg, so erhält man die **Profile (Log-) Likelihood Funktion**

$$p\ell(\lambda|y) = \ell(\lambda, \hat\beta_\lambda, \hat\sigma^2_\lambda|y) = \begin{cases} -\frac{n}{2}\log \mathrm{SSE}_\lambda(\hat\beta_\lambda) + n\log|\lambda| + (\lambda-1)\sum_{i=1}^n \log y_i, & \text{falls } \lambda \neq 0, \\ -\frac{n}{2}\log \mathrm{SSE}_0(\hat\beta_0) - \sum_{i=1}^n \log y_i, & \text{falls } \lambda = 0. \end{cases} \tag{1.4}$$

Für $\lambda = 1$ resultiert beispielsweise

$$p\ell(1|y) = -\frac{n}{2}\log \mathrm{SSE}_1(\hat\beta_1) = n\log\left(\sum_{i=1}^n (y_i - x_i^\top \hat\beta_1)^2\right)^{-1/2}.$$

Wegen

$$\begin{aligned} p\ell(\lambda|y) &= -\frac{n}{2}\log\sum_{i=1}^n \frac{(y_i^\lambda - x_i^\top\hat\beta_\lambda)^2}{\lambda^2} + (\lambda-1)\sum_{i=1}^n \log y_i \\ &= -\frac{n}{2}\log\sum_{i=1}^n \Big((y_i^\lambda - 1)/\lambda - x_i^\top\hat\beta(\lambda)\Big)^2 + (\lambda-1)\sum_{i=1}^n \log y_i, \end{aligned}$$

gilt $\lim_{\lambda\to 0} p\ell(\lambda|y) = p\ell(0|y)$. Obwohl $\mathrm{SSE}_\lambda(\cdot)$ in $\lambda = 0$ unstetig ist, ist die Profile Likelihood Funktion $p\ell(\lambda|y)$ also auch dort stetig.

Ignorieren wir aber auch noch den Term $-\sum_{i=1}^n \log y_i$ in der obigen Profile Likelihood Funktion, so lässt sich diese für $\lambda \neq 0$ schreiben als

$$\begin{aligned} p\ell(\lambda|y) &= -\frac{n}{2}\log\sum_{i=1}^n \Big((y_i^\lambda - 1)/\lambda - x_i^\top\hat\beta(\lambda)\Big)^2 + \lambda\sum_{i=1}^n \log y_i \\ &= -\frac{n}{2}\log\sum_{i=1}^n r_i^2 + \log\left(\exp\left[n\lambda\frac{1}{n}\sum_{i=1}^n \log y_i\right]\right) \\ &= -\frac{n}{2}\log\sum_{i=1}^n r_i^2 + \log\left(\exp\left[\frac{1}{n}\sum_{i=1}^n \log y_i\right]^{n\lambda}\right) \\ &= -\frac{n}{2}\log\sum_{i=1}^n r_i^2 + \log\left(c^{n\lambda}\right) \\ &= -\frac{n}{2}\log\sum_{i=1}^n r_i^2 + \frac{n}{2}\log\left((c^\lambda)^2\right) \\ &= -\frac{n}{2}\log\sum_{i=1}^n \left(\frac{r_i}{c^\lambda}\right)^2 \end{aligned} \tag{1.5}$$

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

Wir erinnern uns an die Likelihood Quotienten Teststatistik zum Testen der Hypothesen $H_0 : \lambda = \lambda_0$ gegen $H_1 : \lambda \neq \lambda_0$. Die Teststatistik war definiert als der Quotient

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

wobei $L(\cdot|y)$ die Likelihood Funktion der Stichprobe bezeichnet. Unter gewissen Regularitätsbedingung gilt nun dafür

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

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

Ein wichtiger Aspekt der Box-Cox Transformation ist, dass das Modell auf der transformierten Skala die Variation bezüglich des Erwartungswertes der (auf Normalverteilung) transformierten Variablen repräsentiert, während auf der Originalskala das Modell die Variation bezüglich des Medians der originalen Variablen darstellt. Dies sieht man recht einfach für die Log-Transformation ($\lambda = 0$). Seien $\log y_i \sim \text{Normal}(x_i^\top\beta, \sigma^2)$, dann gilt

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

Die originalen Beobachtungen y_i unterliegen selbst einer Lognormalverteilung mit

$$\begin{aligned}
\text{median}(y_i) &= \exp(x_i^\top\beta),\\
\text{E}(y_i) &= \exp(x_i^\top\beta + \sigma^2/2),\\
\text{var}(y_i) &= \left(\exp(\sigma^2) - 1\right)\exp(2x_i^\top\beta + \sigma^2).
\end{aligned}$$

Dies bedeutet, dass das additive Modell für den Erwartungswert (und daher auch für den Median) von $\log y_i$ ein multiplikatives Modell für den Median und für den Erwartungswert von y_i ist. Der Erwartungswert von y_i ist das $1 < \exp(\sigma^2/2)$-fache des Medians und die Varianz ist nicht mehr konstant für $i = 1, \ldots, n$.

Betrachtet man hingegen die Transformation $y(\lambda) = y^\lambda$ mit $\lambda \neq 0$, also $y_i^\lambda \sim \text{Normal}(\mu_i, \sigma^2)$, so folgt

$$\begin{aligned}
\text{median}(y_i) &= \mu_i^{1/\lambda},\\
\text{E}(y_i) &\approx \mu_i^{1/\lambda}\left(1 + \sigma^2(1-\lambda)/(2\lambda^2\mu_i^2)\right),\\
\text{var}(y_i) &\approx \mu_i^{2/\lambda}\sigma^2/(\lambda^2\mu_i^2).
\end{aligned}$$

Wiederum ist die offensichtliche Unstetigkeit zwischen $\lambda = 0$ und $\lambda \neq 0$ in der Verwendung von y^λ anstelle von $(y^\lambda - 1)/\lambda$ begründet.

1.3 Beispiel: Black Cherry Trees

Das verwendbare Holzvolumen V in feet^3 (1 foot = 30.48 cm) ist an $n = 31$ Black Cherry Bäumen erhoben worden. Von Interesse ist hierbei der Zusammenhang zwischen dem zu erwartenden Holzvolumen und der Baumhöhe H in feet und dem Durchmesser D in inches (1 inch = 2.54 cm), welcher auf einer Höhe von 4.5 feet über dem Boden gemessen wurde. Das Modell sollte also das verwendbare Holzvolumen V aus den leicht zu messenden Größen H und D vorhersagen.

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

> (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
```

Das lineare Regressionsmodell für das Volumen mit den beiden Prädiktoren Durchmesser und Höhe hat deutliche Schwächen (vgl. Abbildung 1.1). So scheint die Abhängigkeit des Volumens vom Durchmesser nicht wirklich linear zu sein und die Varianz des Volumens nimmt scheinbar mit der Baumhöhe zu. Besonders auffällige Hebelpunkte liegen zwar nicht wirklich vor aber der Verlauf der Residuen weist eine deutlich erkennbare Abhängigkeit vom Durchmesser auf. Dies alles motiviert die Verwendung der Box-Cox Transformation für das Volumen.

```
> 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

> # direct calculation of pl(lambda|y) values - does not 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)  # avoiding lambda=0
> res <- matrix(0, nrow=length(lambda), 2, dimnames=list(NULL,c("lambda","pl")))
> C <- exp(mean(log(V))) # scaling constant
```

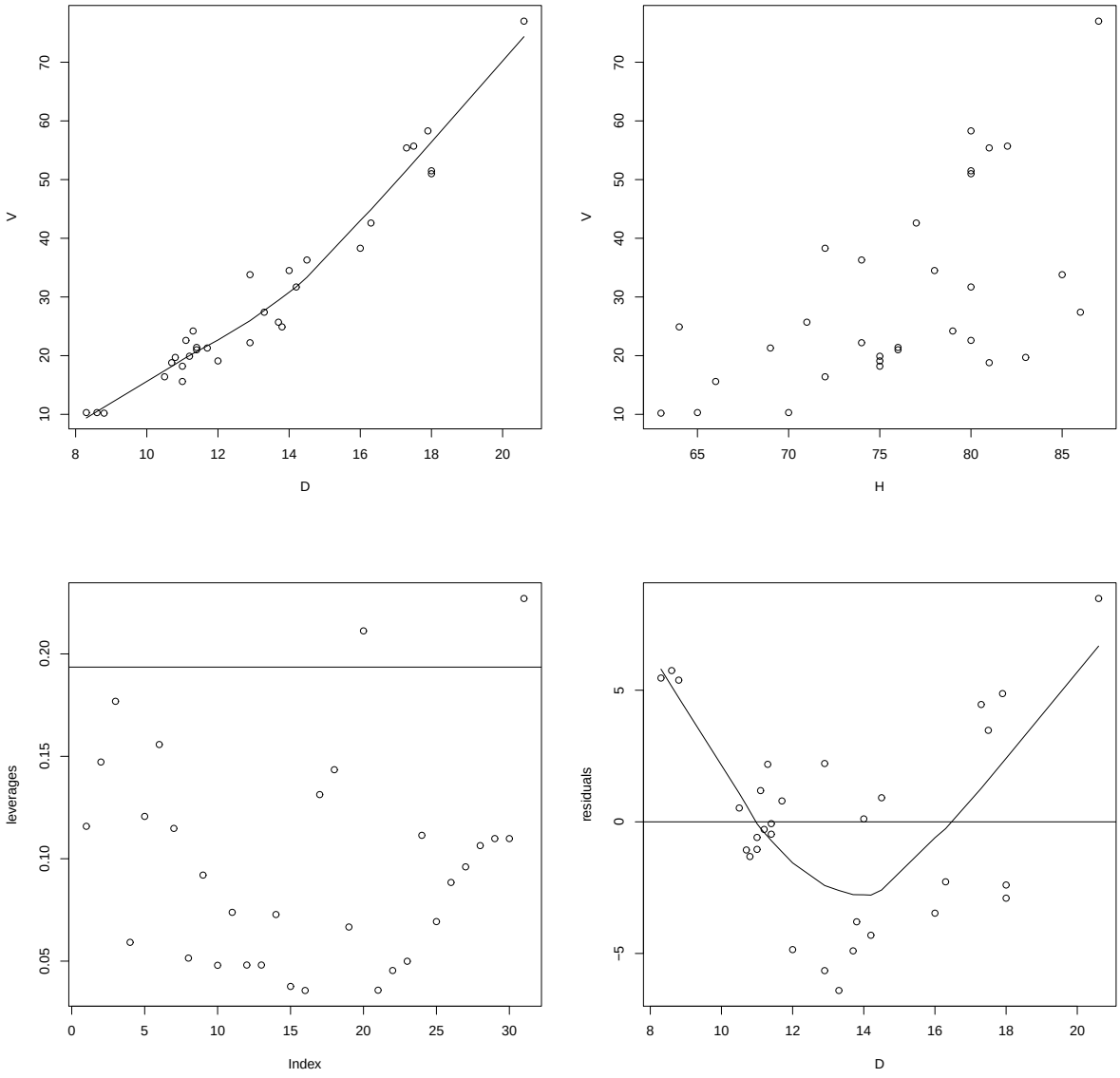

Abbildung 1.1: Oben: Volumen gegen Durchmesser (links) und gegen Höhe (rechts). Unten: Diagonalelemente der Hatmatrix (links) und Residuen gegen Durchmesser (rechts) unter dem linearen Modell für V.

```
> 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 on top to verify match
```

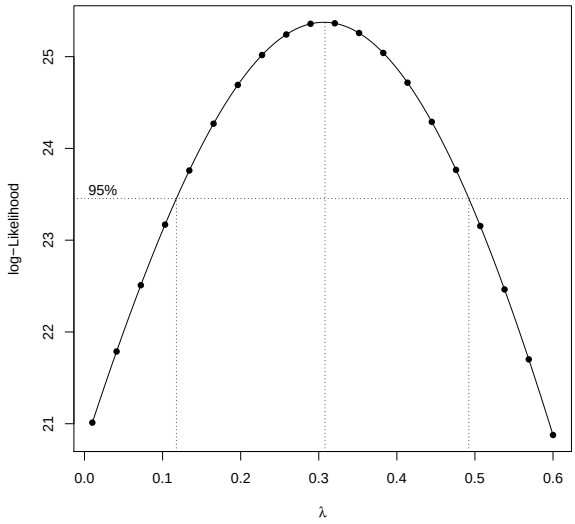

Abbildung 1.2: Profile Likelihood Funktion mit 95% Konfidenzintervall für λ.

Wie auch in der Abbildung 1.2 ersichtlich, tritt das Maximum der Profile Likelihood Funktion $p\ell(\lambda|y)$ in der Nähe von $\lambda = 0.31$ auf. Das approximative 95% Konfidenzintervall ist relativ klein, und erstreckt sich etwa auf $(0.12, 0.49)$. Dieses Intervall überdeckt weder die Null noch die Eins, jedoch scheint der Wert 1/3 (Kubikwurzel) äußerst plausibel. Es liegt in der Natur einer Volumsmessung, dass sich diese kubisch bezüglich der beiden linearen Prädiktoren Höhe und Durchmesser verhält. Daher erscheint es auch sinnvoll, die Kubikwurzel des Volumens als Response zu verwenden.

```
> 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
```

Der geschätzte Median von V unter diesem Modell mit festem $\lambda = 1/3$ ist $\hat{\mu}^3_{1/3}$, wobei $\mathrm{E}(V^{1/3}) = \mu_{1/3}$. Der Erwartungswert von V wird geschätzt durch $\hat{\mu}^3_{1/3}(1+3\hat{\sigma}^2_{1/3}/\hat{\mu}^2_{1/3})$. Diese beiden Schätzer für den Lageparameter können mit den originalen Volumsmessungen verglichen werden. Wir beschränken uns hier auf den Vergleich mit den Medianen.

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

Andere technische Überlegungen ergeben alternative Modelle. Die unerwünschte Krümmung in der Abbildung 1.1 kann vielleicht auch durch logarithmische Transformation aller Variablen reduziert werden. Dies legt eine Regression auf $\log(D)$ und $\log(H)$ nahe. Soll man jetzt jedoch auf der $\log(V)$ Achse modellieren?

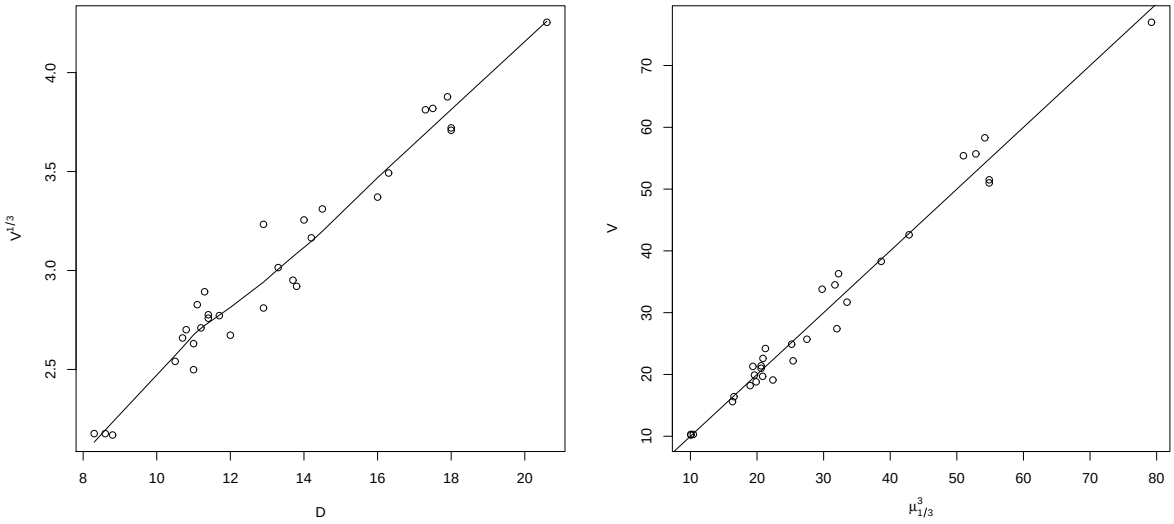

Abbildung 1.3: Kubikwurzel-transformierte Volumina gegen die Durchmesser (links) und Vergleich der geschätzten Mediane mit originalen Beobachtungen (rechts).

```
> 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))
```

Die Profile Likelihood Schätzung unter einer Box-Box Transformation liefert ein Maximum bei etwa $\lambda = -0.07$ und ein 95% Konfidenzintervall von $(-0.24, 0.11)$, welches zwar die Null (logarithmische Transformation), aber nicht mehr die Kubikwurzeltransformation $\lambda = 1/3$ oder die Identität $\lambda = 1$ beinhaltet.

Beide Modelle liefern annähernd dieselben Maxima der Profile Likelihood Funktionen. Welches der beiden ist nun *besser*? Wir können diese beiden Modelle mittels einen Likelihood Quotienten Test miteinander vergleichen. Dazu werden die Modelle eingebettet in die Modellfamilie

$$\begin{aligned} V^* &\sim \text{Normal}(\beta_0 + \beta_1 H^* + \beta_2 D^*, \sigma^2) \\ V^* &= (V^{\lambda_V} - 1)/\lambda_V \\ H^* &= (H^{\lambda_H} - 1)/\lambda_H \\ D^* &= (D^{\lambda_D} - 1)/\lambda_D \end{aligned}$$

Wir vergleichen nun den Wert der Profile Likelihood Funktion in $\lambda_V = 1/3$, $\lambda_H = \lambda_D = 1$ (entspricht dem Modell $\text{E}(V^{1/3}) = \beta_0 + \beta_1 H + \beta_2 D$), mit jenem in $\lambda_V = \lambda_H = \lambda_D = 0$

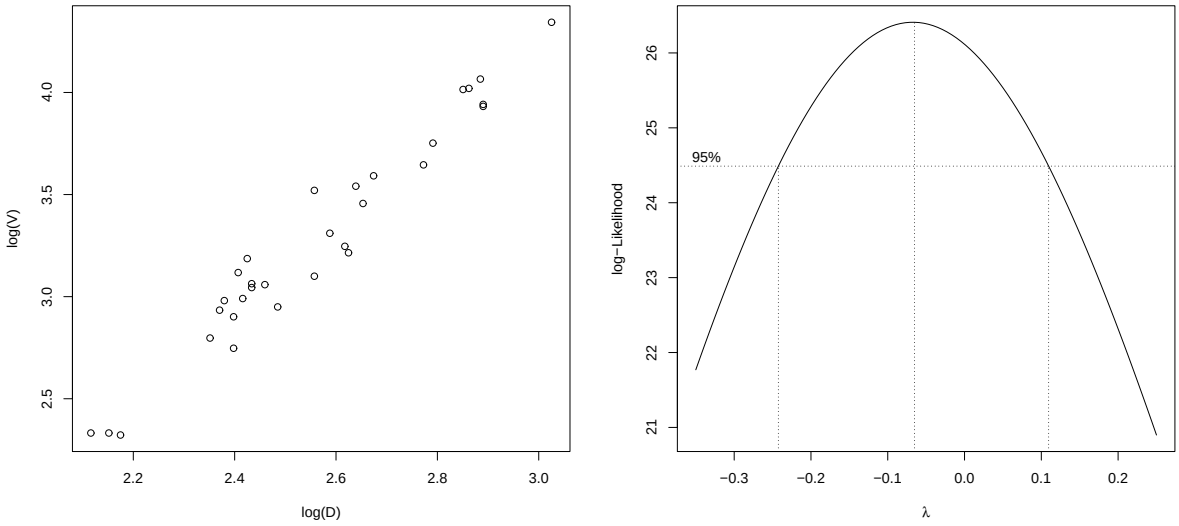

Abbildung 1.4: Lineare Abhängigkeit zwischen $\log(V)$ und $\log(D)$ (links) und Profile Likelihood Funktion für das Modell mit Prädiktor $\log H + \log D$ (rechts).

(entspricht dem Modell $\mathrm{E}(\log(V)) = \beta_0 + \beta_1 \log(H) + \beta_2 \log(D)$) und konzentrieren uns dabei auf die Werte dieser drei λ Parameter. Alle übrigen Parameter sind hierbei *nuisance Parameter* und die Profile Likelihood Funktion wird für feste Transformationsparameter über diese unter den beiden Modellen maximiert.

```
> bc1 <- boxcox(V ~ H + D, lambda = 1/3, plotit = FALSE)
> bc1$x
[1] 0.3333333
> bc1$y
[1] 25.33313

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

Die negative doppelte Differenz der beiden Profile Likelihoods ist nur $-2(25.333-26.116) = 1.566$, was nicht signifikant ist im Vergleich mit den Quantilen einer χ^2_3 Verteilung. Daher können wir auch nicht überzeugend eines der beiden Modelle vorziehen.

Bemerke aber, dass die Parameterschätzung zu $\log(H)$ nahe bei Eins liegt ($\hat{\beta}_1 = 1.117$) und die zu $\log(D)$ fast Zwei ist ($\hat{\beta}_2 = 1.983$). Nimmt man an, dass man einen Baum durch einen Zylinder oder durch einen Kegel beschreiben kann, so wäre sein Volumen $\pi h d^2/4$

(Zylinder) oder $\pi hd^2/12$ (Kegel). In beiden Fällen hätte man ein Ergebnis der Form

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

mit $c = \log(\pi/4)$ (Zylinder) oder $c = \log(\pi/12)$ (Kegel). Jedoch beziehen sich diese Überlegungen auf Größen, welche auf derselben Skala beobachtet sind. Wir konvertieren daher zuerst D von inches auf feet (1 foot entspricht 12 inches), d.h. wir betrachten $D/12$ als Prädiktor im Modell.

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

Natürlich hat diese Konvertierung nur einen Einfluss auf den Schätzer des Intercepts. In einem nächsten Schritt wollen wir die beiden Slopeparameter auf die Werte (1, 2) fixieren und nur noch den Intercept frei schätzen. Wir betrachten also das Modell

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

Hierbei bezeichnet man $1\log H + 2\log(D/12)$ als *offset* (Term mit festen Parametern) und es muss nur noch β_0 geschätzt werden.

```
> (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
```

Da -1.20 näher bei -1.34 liegt als -0.24, kann das verwendbare Holzvolumen dieser Baumart eher durch ein Kegelvolumen als durch das eines Zylinders beschrieben werden, hat jedoch ein etwas größeres Volumen als ein Kegel.

Kapitel 2

Die Lineare Exponentialfamilie

Beim Linearen Modell (LM) wird angenommen, dass die abhängigen Variablen (Responses) y_i stochastisch unabhängige, normalverteilte Größen sind mit Erwartungen $\mu_i = x_i^\top \beta$ und konstanter Varianz σ^2. In manchen Situationen ist die Annahme einer Normalverteilung sicherlich sehr künstlich und nur schwer zu vertreten. Man denke hierbei nur an Modelle für absolute Häufigkeiten, relative Anteile oder für Anzahlen. Weiters gibt es datengenerierende Mechanismen, die für größere Erwartungswerte auch größere Variabilität induzieren. Dazu zählen beispielsweise Modelle bei konstanten Variationskoeffizienten. Da bei einem LM die Erwartungswerte beliebig im p-dimensionalen Raum liegen können, ist ein solches LM sicherlich für nicht-negative oder speziell auch für binäre Responses unpassend.

Wir wollen nun wegen all dieser Schwachstellen die Klasse der Generalisierten Linearen Modelle (GLMs) betrachten, die gerade bezüglich der oben angeführten Restriktionen bei LMs einige flexible Verallgemeinerungen anbietet. So wird bei einem GLM statt der Normalverteilung eine Verteilung aus der einparametrigen linearen Exponentialfamilie (in kanonischer Form) angenommen und dadurch die Varianz in Termen des Erwartungswertes modelliert. Darüberhinaus wird der Erwartungswert nicht ausschließlich direkt linear modelliert, sondern der lineare Prädiktor $x_i^\top \beta$ entspricht einer bekannten Funktion $g(\mu_i)$, der Linkfunktion.

Wir definieren zuerst die einparametrige lineare Exponentialfamilie in kanonischer Form und besprechen dann einige wesentliche Eigenschaften dieser Familie.

Definition 2.1. Eine Zufallsvariable y sei aus einer Verteilung mit Dichte- oder Wahrscheinlichkeitsfunktion

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{2.1}$$

für spezielle bekannte Funktionen $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ mit $a(\phi) > 0$. Kann ϕ als feste Größe betrachtet werden, so bezeichnet man $f(y|\theta)$ als **einparametrige lineare Exponentialfamilie in kanonischer Form** mit kanonischem Parameter θ.

Bemerkung 2.1. Zur Erinnerung ist eine (allgemeine) Exponentialfamilie definiert durch

$$f(y|\theta) = h(y)p(\theta)\exp\left\{\sum_{j=1}^{k} t_j(y)w_j(\theta)\right\}, \tag{2.2}$$

wobei $h(y) \geq 0$ und $t_1(y), \ldots, t_k(y)$ reellwertige Funktionen der Beobachtung y (unabhängig von θ) und $p(\theta) \geq 0$ und $w_1(\theta), \ldots, w_k(\theta)$ reellwertige Funktionen des möglicherweise vektorwertigen Parameters θ (unabhängig von y) sind. Um darin (2.1) zu erkennen, schreiben wir diese Familie um zu

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

und setzen hierin $k = 1$ (einparametrig), $t(y) = y$ (linear), $w(\theta) = \theta$ (kanonische Parametrisierung), sowie $\log(p(\theta)) = -b(\theta)$ und $\log(h(y)) = c(y, \phi)$. Dies liefert bis auf die Skalierung durch $a(\phi)$ die Form (2.1).

Bemerkung 2.2. Für die allgemeine Exponentialfamilie (2.2) gilt

$$\mathrm{E}\left(\sum_{j=1}^{k} \frac{\partial w_j(\theta)}{\partial \theta_l} t_j(y)\right) = -\frac{\partial}{\partial \theta_l}\log(p(\theta)), \tag{2.3}$$

$$\mathrm{var}\left(\sum_{j=1}^{k} \frac{\partial w_j(\theta)}{\partial \theta_l} t_j(y)\right) = -\frac{\partial^2}{\partial \theta_l^2}\log(p(\theta)) - \mathrm{E}\left(\sum_{j=1}^{k} \frac{\partial^2 w_j(\theta)}{\partial \theta_l^2} t_j(y)\right). \tag{2.4}$$

Für den Fall $k = 1$, $w(\theta) = \theta$ und $t(y) = y$ liefert (2.3) das Ergebnis $\mathrm{E}(y) = b'(\theta)$ und mit (2.4) erhält man dafür $\mathrm{var}(y) = b''(\theta)$.

Bemerkung 2.3. Im Zusammenhang mit der Diskussion der Cramér-Rao Ungleichung wurde bereits gezeigt, dass unter gewissen Regularitätsbedingungen (welche für die Exponentialfamilie erfüllt sind) für die Scorefunktion und die Informationszahl folgendes gilt:

$$\mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = 0\,, \tag{2.5}$$

$$\mathrm{var}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = \mathrm{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2 = \mathrm{E}\left(-\frac{\partial^2 \log f(y|\theta)}{\partial \theta \partial \theta^\top}\right). \tag{2.6}$$

Wir wenden nun diese Eigenschaften auf unsere lineare Exponentialfamilie (2.1) an und erhalten die ersten beiden Momente der Responsevariablen unter diesem Verteilungsmodell. Mit (2.5) erhält man

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

also $\mathrm{E}(y) = b'(\theta)$ (wie bereits in der obigen Bemerkung hergeleitet), und mit (2.6) resultiert

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

Somit folgt für die beiden ersten Momente (Kumulanten) der linearen Exponentialfamilie

$$\begin{aligned}\mathrm{E}(y) &= b'(\theta)\\ \mathrm{var}(y) &= a(\phi)b''(\theta)\,.\end{aligned}$$

Sei $\mathrm{E}(y) = b'(\theta) = \mu$ und $\mathrm{var}(y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$. Die Varianz von y ist also ein Produkt zweier Funktionen: $V(\mu)$ hängt ausschließlich vom Erwartungswert μ ab und $a(\phi)$ ist von μ unabhängig. Wir nennen $V(\mu)$ **Varianzfunktion**, während ϕ als **Dispersionsparameter** bezeichnet wird. Die Funktion $b(\theta)$ wird auch manchmal **Kumulantenfunktion** genannt. Dies wird durch die folgende Diskussion noch besser motiviert.

Kumulanten höherer Ordnung bestimmt man einfacher mit der Kumulantenerzeugenden Funktion $K(t) = \log M(t)$, wobei $M(t)$ die Momentenerzeugende bezeichnet. Die k-te Kumulante κ_k ist gegeben durch $K^{(k)}(t)|_{t=0}$ und steht mit den Momenten in einer einfachen Beziehung, denn es gilt

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

Für die Exponentialfamilie folgt

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

woraus

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

folgt. Die Momentenerzeugende ist daher gegeben durch

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

und als Kumulantenerzeugende Funktion resultiert

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

Die k-te Kumulante von y, $\kappa_k(y)$, ist somit

$$\kappa_k(y) = K^{(k)}(t)|_{t=0} = a(\phi)^{k-1} b^{(k)}\Big(\theta + a(\phi)t\Big)\Big|_{t=0} = a(\phi)^{k-1} b^{(k)}(\theta)\,. \tag{2.7}$$

2.1 Maximum Likelihood Schätzung

Liegt eine n-elementige Zufallsstichprobe (iid) $y_1, \ldots, y_n$ aus der Exponentialfamilie (2.1) vor, so ist der Maximum Likelihood Schätzer (MLE) für den Erwartungswert μ die 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)

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

vereinfacht sich die obige Scorefunktion zu

$$\sum_{i=1}^{n} \frac{\partial \log f(y_i|\theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - \mu}{a(\phi)V(\mu)} = \sum_{i=1}^{n} \frac{y_i - \mu}{\mathrm{var}(y_i)}\,. \tag{2.8}$$

Diese recht simple Form resultiert bei der linearen Exponentialfamilie nur bezüglich der Ableitung nach μ. Sie entspricht der Ableitung der Fehlerquadratsumme beim klassischen Linearen Modell unter Normalverteilungsannahme mit $\mathrm{var}(y_i) = \sigma^2$.

Bemerkung 2.4. Generell könnten wir annehmen, dass es bei Vorliegen einer Stichprobe $y_1, \ldots, y_n$ beobachtungsspezifische Funktionen $a_i(\cdot)$ gibt, die jedoch nur von ein und demselben globalen Dispersionsparameter ϕ abhängen dürfen (ansonsten wäre die Anzahl der unbekannten Dispersionsparameter gleich n und somit nicht mehr schätzbar). Ein Beispiel dafür aus der Praxis ist die Modellierung von N arithmetischen Mitteln basierend auf Stichproben mit Umfängen $n_1, \ldots, n_N$. Seien dazu die N Gruppen von Stichprobenelementen als $y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{N1}, \ldots, y_{Nn_N}$ gegeben. Von jeder Gruppe sei aber ausschließlich das arithmetische Mittel $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki}$ verfügbar. Falls die y_{ki} eine (iid) Zufallsstichprobe darstellen mit $\mathrm{E}(y_{ki}) = \mu$ und $\mathrm{var}(y_{ki}) = \sigma^2$, dann gilt für die Mittelwerte $\mathrm{E}(\bar{y}_k) = \mu$ und $\mathrm{var}(\bar{y}_k) = \sigma^2/n_k = a_k \cdot \phi$ mit bekannten Gewichten $a_k = 1/n_k$ und unbekannter Dispersion $\phi = \sigma^2$.

Wir werden uns daher 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.

2.2 Mitglieder der Linearen Exponentialfamilie

Wir werden nun einige wichtige Mitglieder dieser Verteilungsfamilie kennen lernen. Dabei wird eine Parametrisierung verwendet, bei der der Erwartungswert immer durch μ bezeichnet wird. Die Varianz ist dadurch oft proportional zu einer Potenz von μ. Die Dispersionsfunktion $a(\phi)$ sei nun ausschließlich von der Form $a \cdot \phi$.

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

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

Setzen wir nun $\theta = \mu$ und $\phi = \sigma^2$, so führt dies zur linearen Exponentialfamilie mit

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

und mittels (2.7) zu

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

Hierbei wird die Dispersion $\phi = \sigma^2$ nicht als zu schätzender Parameter gesehen.

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

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

Mit $\theta = \log\mu$ und festem $\phi = 1$ führt dies zur linearen Exponentialfamilie mit

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

und mittels (2.7) zu den Kumulanten

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

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

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

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

Unter dieser Parametrisierung gilt $\text{E}(y) = a/\lambda$ und $\text{var}(y) = a/\lambda^2$. Die Reparametrisierung $\mu = \nu/\lambda$ mit $\nu = a$ liefert $\text{E}(y) = \mu$ und $\text{var}(y) = \mu^2/\nu$. Die entsprechende Dichtefunktion lautet damit

$$\begin{aligned} f(y|\mu,\nu) &= \exp\left(-\frac{\nu}{\mu}y\right)\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\frac{1}{\Gamma(\nu)} \\ &= \exp\left(-\frac{\nu}{\mu}y + \nu\log\nu - \nu\log\mu + (\nu-1)\log y - \log\Gamma(\nu)\right) \\ &= \exp\left(\frac{y\left(-\frac{1}{\mu}\right) + \log\frac{1}{\mu}}{1/\nu} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right)\,, \quad \mu,\nu,y>0\,. \end{aligned}$$

Mit $\theta = -1/\mu$ und $\phi = 1/\nu$ führt dies zur Exponentialfamilie mit

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

und mittels (2.7) zu den Kumulanten

$$\begin{aligned} \text{E}(y) &= b'(\theta) = -\frac{1}{\theta} = \mu \\ \text{var}(y) &= \phi b''(\theta) = \phi\frac{1}{\theta^2} = \frac{1}{\nu}\mu^2 \\ \kappa_k(y) &= (k-1)!\,\nu\left(\frac{\mu}{\nu}\right)^k \qquad \text{für } k>2\,. \end{aligned}$$

Wir haben hierbei eine Varianzstruktur, die sich proportional zum Quadrat des Erwartungswertes verhält. Die Dispersion spielt hierbei die Rolle des Proportionalitätsparameters.

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

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

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

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

und mittels (2.7) zu den Kumulanten

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

Hierbei wächst die Varianz sogar proportional zu μ^3.

• Die **standardisierte Binomialverteilung** $my \sim \mathrm{Binomial}(m,\pi)$:

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

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

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

und mittels (2.7) zu den Kumulanten

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

Die Zufallsvariable y beschreibt hierbei eine relative Häufigkeit. Das m-fache von y, die absolute Häufigkeit, ist eine binomialverteilte Größe. Man bemerke, dass in dieser Repräsentation der Dispersionsparameter bekannt Eins ist und die bekannte Anzahl m reziprok als Gewicht eingeht.

2.3 Die Quasi-Likelihoodfunktion

Betrachtet man die Scorefunktion (2.8) zur Exponentialfamilie, so erkennt man, dass der MLE $\hat{\mu}$ nur von der zugrundeliegenden Varianzannahme abhängt. In diesem Abschnitt wird nun untersucht, welche Eigenschaften ein Schätzer für μ aufweist, falls die Scorefunktion auch für eine Varianzannahme verwendet wird, die zu keinem Mitglied aus der Exponentialfamilie gehört. Generell spricht man dann von einer **Quasi-Scorefunktion**. Ohne Verlust der Allgemeinheit wollen wir annehmen, dass die Dispersion gegeben ist durch $a \cdot \phi = \phi$, also dass das Gewicht Eins ist.

Definition 2.2. Für eine Zufallsvariable y mit $\mathrm{E}(y) = \mu$ und $\mathrm{var}(y) = \phi V(\mu)$ und bekannter Varianzfunktion $V(\cdot)$ ist die **Quasi-Likelihoodfunktion** $q(\mu|y)$ (eigentlich Log-Quasi-Likelihoodfunktion) definiert über die Beziehung

$$\frac{\partial q(\mu|y)}{\partial \mu} = \frac{y-\mu}{\phi V(\mu)}\,, \tag{2.9}$$

oder äquivalent dazu durch

$$q(\mu|y) = \int^{\mu} \frac{y-t}{\phi V(t)} dt + \text{Funktion in } y \text{ (und } \phi)\,. \tag{2.10}$$

Die Ableitung $\partial q/\partial \mu$ wird als **Quasi-Scorefunktion** bezeichnet. Verglichen mit (2.5) und (2.6) hat sie folgende Eigenschaften mit der Scorefunktion gemeinsam

$$\mathrm{E}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = 0, \tag{2.11}$$

$$\mathrm{var}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = \frac{\mathrm{var}(y)}{\phi^2 V^2(\mu)} = \frac{1}{\phi V(\mu)} = -\mathrm{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right)\,. \tag{2.12}$$

Satz 2.1 (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

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

dann und nur 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.

Beweis: ($\Rightarrow$) Integration bezüglich μ liefert

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

($\Leftarrow$) Mit (2.7) folgt für die Kumulanten der linearen Exponentialfamilie $\mathrm{E}(y) = \mu = b'(\theta)$ und $\mathrm{var}(y) = \phi V(\mu) = \phi b''(\theta)$. Es gilt daher

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

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

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

2.3.1 Quasi-Likelihoodmodelle

Mit dieser Konstruktionsidee wird nun für einige Varianzfunktionen der assoziierte Parameter θ hergeleitet, sowie die Quasi-Likelihoodfunktion bestimmt.

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

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

- $V(\mu) = \mu$, $\phi = 1$, $0 < \mu$, $0 \le y$, (vgl. mit $y \sim \mathrm{Poisson}(\mu)$):

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

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

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

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

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

- $V(\mu) = \mu^k$, $\phi = 1$, $0 < \mu$, $0 \leq y$, $k \geq 3$:

$$\begin{aligned} \theta &= \int \frac{1}{\mu^k} d\mu = -\frac{1}{(k-1)\mu^{k-1}}\,, \\ q(\mu|y) &= \int^{\mu} \frac{y-t}{t^k} dt = \frac{1}{\mu^k}\left(\frac{\mu^2}{k-2} - \frac{y\mu}{k-1}\right)\,. \end{aligned}$$

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

$$\begin{aligned} \theta &= \int \frac{1}{\mu(1-\mu)} d\mu = \log\frac{\mu}{1-\mu}\,, \\ q(\mu|y) &= \int^{\mu} \frac{y-t}{t(1-t)} dt = y\log\frac{\mu}{1-\mu} + \log(1-\mu)\,. \end{aligned}$$

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

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

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

$$\begin{aligned} \theta &= \int \frac{1}{\mu + \mu^2/k} d\mu = \log\frac{\mu}{k+\mu}\,, \\ q(\mu|y) &= \int^{\mu} \frac{y-t}{t + t^2/k} dt = y\log\frac{\mu}{k+\mu} + k\log\frac{1}{k+\mu}\,. \end{aligned}$$

Während die ersten vier (Normal-, Poisson-, Gamma- und Inverse Gaußverteilung) und das sechste Beispiel (standardisierte Binomialverteilung) mit bereits bekannten Mitgliedern der Exponentialfamilie vergleichbar sind, stellen das fünfte sowie das siebente (speziell für Modelle für Prozentsätze) und achte Beispiel (Negativ-Binomialverteilung) neue (nicht in der einparametrigen linearen Exponentialfamilie inkludierte) Varianzfunktionen dar. Hängt die Varianzfunktion von einem weiteren Parameter k ab, so muss diese Größe beim Quasi-Likelihoodansatz als fest betrachtet werden. Es besteht (noch) keine Möglichkeit, k simultan mit μ zu schätzen.

2.3.2 Quasi-Dichten

Natürlich ist durch die Spezifikation einer Erwartungswert/Varianz-Beziehung auch eine Dichtefunktion spezifizierbar. Aus der (Log)-Quasi-Likelihoodfunktion folgt mit der Normalisierungsfunktion

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

als **Quasi-Dichte** (siehe dazu Nelder & Lee, 1992)

$$f_q(y|\mu) = \frac{\exp(q(\mu|y))}{\omega(\mu)}\,. \tag{2.13}$$

Die Größe $\omega(\mu)$ ist ungleich Eins, wenn die Varianz $\phi V(\mu)$ zu keiner Verteilung mit Dichte- oder Wahrscheinlichkeitsfunktion aus der linearen Exponentialfamilie gehört. Andererseits ist $\omega(\mu) = 1$, $\forall\mu$, falls zur Varianz eine Exponentialfamilie existiert.
Zur Quasi-Dichte (2.13) korrespondiert nun die Log-Likelihoodfunktion

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

und

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

Dieser Score unterscheidet sich vom Quasi-Score genau um

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

Hierbei bezeichnet

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

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

Kapitel 3

Das Generalisierte Lineare Modell

Unter Annahme der Existenz von $\mathrm{E}(y_i)$ und $\mathrm{var}(y_i)$ wird in der Klasse der Generalisierten Linearen Modelle (GLM) eine Parametrisierung der Form

$$\begin{aligned}&\text{stochastische Komponente: } y_i \overset{ind}{\sim} \text{Exponentialfamilie}(\theta_i),\ \mathrm{E}(y_i) = \mu_i = \mu(\theta_i)\\&\text{systematische Komponente: } \eta_i = x_i^\top \beta\\&\text{Linkfunktion: } g(\mu_i) = \eta_i\end{aligned}$$

betrachtet, wobei der Responsevektor $y = (y_1, \ldots, y_n)^\top$ aus unabhängigen Komponenten y_i aufgebaut ist mit $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{var}(y_i) = a_i \phi V(\mu_i)$. Die Dispersion sei wiederum gegeben durch das Produkt $a_i\phi$ von zuvor. Es bezeichnet im weiteren $x_i = (x_{i0}, x_{i1}, \ldots, x_{i,p-1})^\top$ den $p \times 1$ Vektor von bekannten erklärenden Variablen, zusammengefasst zu einer $n \times p$ Designmatrix $X = (x_1, \ldots, x_n)^\top$, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ den $p \times 1$ Vektor mit den unbekannten Parametern, $\eta = (\eta_1, \ldots, \eta_n)^\top$ den $n \times 1$ Vektor mit den Linearen Prädiktoren und $g(\cdot)$ eine bekannte (monotone und zweimal stetig differenzierbare) Linkfunktion.
Die wesentlichen Unterschiede zum herkömmlichen Linearen Modell sind:

- Es besteht keine allgemeine Additivität bezüglich nicht-beobachtbarer Fehlerterme ϵ_i wie beim Linearen Modell.
- Eine Abhängigkeit der Varianzstruktur auch vom Erwartungswert ist möglich.
- Eine Funktion des Erwartungswertes wird linear modelliert. Dies ist keinesfalls zu verwechseln mit einer einfachen Transformation der Responsevariablen.

Unser Hauptinteresse liegt nun in der Konstruktion eines Schätzers für den Parametervektor β, sowie an einem Maß für die Güte der Modellanpassung. Beides ist für Maximum-Likelihood-Schätzer besonders einfach und stellt im Wesentlichen nur eine Verallgemeinerung der Resultate bei den Linearen Modelle dar.

3.1 Maximum Likelihood Schätzung

Falls $y_1, \ldots, y_n$ unabhängige Responses sind und die y_i aus derselben Exponentialfamilie stammen mit Parameter (θ_i, ϕ), wobei der Vektor $\theta = (\theta_1, \ldots, \theta_n)^\top$ die unbekannten Parameter beschreibt welche geschätzt werden sollen, und ϕ (vorerst) als bekannte (nuisance) Komponente betrachtet wird, so ist die Log-Likelihoodfunktion der Stichprobe gegeben durch

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

Unter der recht allgemeinen Annahme $\mu = \mu(\beta)$ folgt aus (2.8) die Scorefunktion

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

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

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

und deshalb

$$\frac{\partial \ell(\theta(\beta)|y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i\phi V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} , \qquad j = 0, 1, \ldots, p-1 . \tag{3.1}$$

Es wurde bereits gezeigt, dass $b'(\theta_i) = \mu_i$ für die lineare Exponentialfamilie hält. Somit gilt wegen $g(\mu_i) = x_i^\top\beta$ für den kanonischen Parameter

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}(g^{-1}(x_i^\top\beta)) ,$$

und es ist naheliegend für die Linkfunktion die spezielle Wahl $g(\cdot) = b'^{-1}(\cdot)$ zu betrachten. Diesen speziellen Link $g(\mu_i) = \theta_i = x_i^\top\beta$ nennt man **kanonische Linkfunktion**. Hierbei wird der Parameter θ direkt durch den linearen Prädiktor η modelliert. In diesem Fall ist $g(\cdot)$ die Inverse von $b'(\cdot)$ und wegen $\mu = b'(\theta)$ folgt

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

Die Scorefunktion (3.1) vereinfacht sich für eine kanonische Linkfunktion zu

$$\frac{\partial \ell(\theta(\beta)|y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i\phi} x_{ij} , \qquad j = 0, 1, \ldots, p-1 . \tag{3.2}$$

Für $a_i = 1$ (ungewichtete Situation) gilt hier bei Modellen mit Intercept ($x_{i0} = 1,\ \forall i$) die bekannte Eigenschaft eines MLEs für den Erwartungswert, dass die Summe der Beobachtungen gleich der Summe der geschätzten Erwartungswerte ist, also dass

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

Falls für alle Beobachtungen $a_i = 1$ in der linearen Exponentialfamilie gilt und die Dispersion ϕ gegeben ist, dann folgt unter Verwendung des kanonischen Linkmodells als gemeinsame Dichte- oder Wahrscheinlichkeitsfunktion der Stichprobe die Aussage des Faktorisierungssatzes, nämlich

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

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

Der MLE $\hat{\beta}$ ist also für den allgemeinen Fall als Nullstelle von (3.1) oder im kanonischen Fall als Nullstelle von (3.2) definiert. Beide Gleichungssysteme sind i.A. nichtlinear in β ($\mu = g^{-1}(x^\top\beta)$) und können nur numerisch (iterativ) gelöst werden. Einzige Ausnahme bildet das klassische lineare Modell, in dem μ linear in β ist.

Die Newton-Raphson Methode liefert die Iterationsvorschrift

$$\beta^{(t+1)} = \beta^{(t)} + \left(-\frac{\partial^2\ell(\theta(\beta)|y)}{\partial\beta\partial\beta^\top}\right)^{-1} \frac{\partial\ell(\theta(\beta)|y)}{\partial\beta}\,, \qquad t = 0, 1, \ldots\,, \tag{3.3}$$

wobei beide Ableitungen der rechten Seite von (3.3) an der Stelle $\beta^{(t)}$ betrachtet werden. In Matrixnotation folgt für den Scorevektor

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

mit $D = \mathrm{diag}(d_1, \ldots, d_n)$ und $W = \mathrm{diag}(w_1, \ldots, w_n)$, wobei

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

Als negative Hessematrix der Log-Likelihoodfunktion resultiert somit

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

wegen $\partial\mu/\partial\eta = D^{-1}$. Weiters ist

$$\begin{aligned}\frac{\partial d_i w_i}{\partial \eta_i} &= -\frac{a_i V'(\mu_i)\frac{\partial \mu_i}{\partial \eta_i} g'(\mu_i) + a_i V(\mu_i) g''(\mu_i)\frac{\partial \mu_i}{\partial \eta_i}}{(a_i V(\mu_i) g'(\mu_i))^2} \\ &= -\frac{V'(\mu_i)g'(\mu_i) + V(\mu_i)g''(\mu_i)}{a_i V^2(\mu_i) g'^3(\mu_i)}\,. \end{aligned} \tag{3.4}$$

Fasst man die Elemente

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

zusammen zur Diagonalmatrix W^*, für die $\mathrm{E}(W^*) = W$ gilt, so resultiert als Newton-Raphson Vorschrift

$$\beta^{(t+1)} = \beta^{(t)} + (X^\top W^{*(t)} X)^{-1} X^\top D^{(t)} W^{(t)} (y - \mu^{(t)})\,, \qquad t = 0, 1, \ldots\,. \tag{3.5}$$

Bemerke, dass das Produkt von Scorevektor mit der inversen Hessematrix in dieser Iterationsvorschrift **unabhängig vom Dispersionsparameter** ϕ ist.
Mit sogenannten **Pseudobeobachtungen** (adjusted dependent variates)

$$z = X\beta + W^{*-1} D W (y - \mu) \tag{3.6}$$

kann (3.5) umgeschrieben werden in eine **Iterative (Re)Weighted Least Squares** Notation (IWLS- oder IRLS-Prozedur)

$$\beta^{(t+1)} = (X^\top W^{*(t)} X)^{-1} X^\top W^{*(t)} z^{(t)}\,, \tag{3.7}$$

wobei hier wiederum die rechte Seite in $\beta^{(t)}$ betrachtet wird.
Für **kanonische Links** ($g'(\mu) = 1/V(\mu)$, $g''(\mu) = -V'(\mu)/V^2(\mu)$) verschwinden die Ableitungen (3.4), da

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

und es gilt $W^* = W$. Die Pseudobeobachtungen (3.6) vereinfachen sich deshalb zu

$$z = X\beta + D(y - \mu) = X\beta + V^{-1}(y - \mu)\,, \tag{3.8}$$

mit $V = \mathrm{diag}(V(\mu_i))$, weshalb als Iterationsvorschrift

$$\beta^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)} \tag{3.9}$$

folgt.
Zur Erinnerung sei noch darauf hingewiesen, dass

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

beim linearen Modell den Least-Squares Schätzer definiert, welcher nicht iterativ gelöst werden muss ($W = I$). Weiters werden hier die beobachtbaren Responsevariablen y verwendet, während bei GLM's nicht-beobachtbare Pseudobeobachtungen z diese Rolle übernehmen.

Um auch für nicht-kanonische Linkfunktionen ein recht einfaches Schema zu haben, ist es üblich, anstelle der beobachteten negativen Hessematrix in (3.3) deren Erwartungswert (Informationsmatrix) zu verwenden. Da $\mathrm{E}(X^\top W^* X) = X^\top W X$ gilt, folgt bei dieser sogenannten **Fisher Scoring Technik** wie in (3.9) wiederum als Iterationsvorschrift

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

mit den Pseudobeobachtungen

$$z = X\beta + D(y - \mu)\,.$$

Dafür ist $\mathrm{E}(z) = X\beta$ und wegen

$$\mathrm{var}(y) = \phi(DWD)^{-1}$$

resultiert $\mathrm{var}(z) = D\mathrm{var}(y)D = \phi W^{-1}$. Dies bedeutet, dass die Elemente in der Gewichtsmatrix W gerade proportional zu den reziproken Varianzen der Pseudobeobachtungen sind.

3.2 Asymptotische Eigenschaften des MLEs

Um die asymptotischen Momente des Schätzers $\hat\beta$ herzuleiten, ist es angebracht, die Scorefunktion um den wahren Parameter β zu entwickeln. Aus

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

und nachdem die Hessematrix durch deren Erwartungswert $-X^\top W X$ ersetzt wurde, folgt

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

Da $\mathrm{var}(y) = \phi(DWD)^{-1}$ ergibt sich daher

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

wobei W hier im wahren Parameter β ausgewertet ist. Fahrmeir & Kaufmann (1985) zeigten sogar, dass

$$\sqrt{n}\Big(\hat\beta - \beta\Big) \xrightarrow{D} \mathrm{Normal}_p(0, n\phi(X^\top W X)^{-1})\,. \tag{3.10}$$

3.3 Pearson Statistik

Natürlich ist man nicht nur am asymptotischen Ergebnis für die Varianz-Kovarianzmatrix vom MLE $\hat{\beta}$ interessiert, sondern möchte vor allem diese Matrix auch schätzen. Standardmäßig wird dafür der Plug-in Schätzer verwendet, d.h. der unbekannte Parameter β in W wird durch $\hat{\beta}$ ersetzt, und man erhält

$$\widehat{\text{var}}(\hat{\beta}) = \phi(X^\top W(\hat{\beta})X)^{-1}\,. \tag{3.11}$$

Nun hängt dieses Ergebnis aber auch noch vom möglicherweise unbekannten Dispersionsparameter ϕ ab. Bei den linearen Regressionsmodellen unter Normalverteilungsannahme wird σ^2 durch die biaskorrigierte mittlere Fehlerquadratsumme

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

geschätzt. Unter Annahme einer linearen Exponentialfamilie kann der Dispersionsparameter dargestellt werden als $\phi = \text{var}(y_i)/a_iV(\mu_i)$, für alle $i = 1, \ldots, n$. Falls der Parameter β bekannt wäre, dann würden auch alle μ_i bekannt sein und der Schätzer

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

wäre erwartungstreu für ϕ. Naturgemäß ist β unbekannt. Deshalb verwendet man als Schätzer (Momentenmethode) die biaskorrigierte Größe

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{a_iV(\hat{\mu}_i)} = \frac{1}{n-p}X^2\,, \tag{3.12}$$

welche auch als mittlere (generalisierte) **Pearson Statistik** X^2 bezeichnet wird und in (3.11) Verwendung findet. Die einzelnen (nicht quadrierten) Summanden von X^2, also die Terme

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

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

3.4 Score- und Quasi-Scorefunktion

Liegen unabhängige Responsevariablen y_i vor ($i = 1, \ldots, n$), deren Verteilung ein Mitglied der linearen Exponentialfamilie ist, dann können über die Momente der Scorefunktion

(bzgl. β) die folgenden Aussagen gemacht werden. Sei dafür $\ell_i = \log f(y_i|\theta_i(\beta))$ die log-Likelihoodfunktion zu y_i und es gelte weiters für deren Erwartungswert das Modell $g(\mu_i) = \eta_i = x_i^\top \beta$, dann resultiert als Scorefunktion und korrespondierende Hessematrix

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

mit $0 \le j, k \le p-1$. Als Erwartungswerte folgen dafür die bekannten Likelihood-Eigenschaften

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial \ell_i}{\partial \beta_j} \right) &= 0\,, \\
\mathrm{E}\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} \right) &= -\frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k}\,, \\
\mathrm{E}\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k} \right) &= \frac{\mathrm{var}(y_i)}{(a_i \phi V(\mu_i))^2} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} = \frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} = -\mathrm{E}\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} \right)\,.
\end{aligned}$$

Wir wollen nun untersuchen, ob dieselben Eigenschaften resultieren, wenn anstelle eines konkreten Exponentialfamilienmitglieds *nur* ein Quasi-Likelihoodmodell betrachtet wird. Dazu nehmen wir jetzt ausschließlich an, dass $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{var}(y_i) = a_i \phi V(\mu_i)$ gilt, wobei wiederum $g(\mu_i) = x_i^\top \beta$ hält. Als Quasi-Scorefunktion resultiert mittels Definition 2.9

$$\frac{\partial q(\mu_i(\beta)|y_i)}{\partial \beta_j} = \frac{\partial q(\mu_i(\beta)|y_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i \phi V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}\,. \tag{3.13}$$

Dies entspricht genau dem i-ten Summanden in der Scorefunktion (3.1). Man spricht auch bei Quasi-Likelihood Ansätzen von der kanonischen Linkfunktion, falls $g'(\mu) = 1/V(\mu)$ gilt, was wiederum zum Score-Gleichungssystem (3.2) führt. Die Quasi-Likelihood Schätzung entspricht daher der Maximum-Likelihood Schätzung, vorausgesetzt dass zur angenommenen Varianz auch ein Mitglied aus der einparametrigen Exponentialfamilie existiert.

Sei nun $q_i = q(\mu_i(\beta)|y_i)$ für $i = 1, \ldots, n$, so gilt für jede Parameterkomponente

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial q_i}{\partial \beta_j} \right) &= \mathrm{E}\left(\frac{\partial q_i}{\partial \mu_i} \right) \frac{\partial \mu_i}{\partial \beta_j} = 0\,, \\
\mathrm{E}\left(\frac{\partial^2 q_i}{\partial \beta_j \partial \beta_k} \right) &= \mathrm{E}\left(\frac{\partial^2 q_i}{\partial \mu_i^2} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} + \frac{\partial q_i}{\partial \mu_i} \frac{\partial^2 \mu_i}{\partial \beta_j \partial \beta_k} \right) = -\frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k}\,, \\
\mathrm{E}\left(\frac{\partial q_i}{\partial \beta_j} \frac{\partial q_i}{\partial \beta_k} \right) &= \mathrm{E}\left(\left(\frac{\partial q_i}{\partial \mu_i} \right)^2 \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} \right) = \frac{1}{a_i \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} = -\mathrm{E}\left(\frac{\partial^2 q_i}{\partial \beta_j \partial \beta_k} \right)\,.
\end{aligned}$$

All die Resultate für eine Scorefunktion basierend auf einer Likelihoodfunktion halten jetzt auch beim Quasi-Likelihoodansatz für die Quasi-Scorefunktion.

3.5 Deviance und Quasi-Deviance

Um die Güte der Modellanpassung zu bewerten, betrachtet man die Likelihood-Quotienten Teststatistik zum Hypothesentest $H_0 : \mu = g^{-1}(x^\top \beta)$ gegen $H_1 : \mu \neq g^{-1}(x^\top \beta)$. D.h. wir postulieren unter der Nullhypothese, dass unser betrachtetes Modell mit den p Parametern das korrekte ist. Bekannterweise ist die Likelihood-Quotienten Teststatistik konstruiert als der Quotient, gebildet aus dem eingeschränkten Maximum der Likelihoodfunktion $L(\mu|y)$ unter der Nullhypothese und dem uneingeschränkten Maximum, also

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

Nun maximiert der MLE $\hat{\beta}$ gerade die Likelihoodfunktion unter H_0. Wird gar keine Modellstruktur bzgl. der Erwartungswerte $\mu_1, \ldots, \mu_n$ gefordert (uneingeschränktes Maximum), dann wird die Likelihoodfunktion maximal für $\hat{\mu}_i = y_i$ für alle $i = 1, \ldots, n$. Man spricht bei diesem Modell (in dem es gleichviele freie Parameter μ_i wie Beobachtungen y_i gibt) von einem vollen (saturierten) Modell. Das volle Modell kann daher als Modell mit n frei wählbaren Parametern interpretiert werden. Dies liefert unmittelbar den Quotienten

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

Wir bilden damit $-2 \log \Lambda(y)$ und erhalten dafür schlussendlich die Statistik

$$\frac{1}{\phi} D(y, \hat{\mu}) = -2\Big(\ell(\hat{\mu}|y) - \ell(y|y)\Big) . \tag{3.14}$$

Diese wird auch **skalierte Deviance** genannt und große Werte sprechen gegen das betrachtete Erwartungswertmodell. Diese Schreibweise ist für die lineare Exponentialfamilie berechtigt, denn es gilt

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

wobei $\tilde{\theta}$ jenen Wert von $\theta(\mu)$ bezeichnet, wenn für dessen Berechnung $\mu = y$ verwendet wird. Die unskalierte Deviance (bzw. die skalierte im Falle $\phi = 1$) kann daher auch geschrieben werden als

$$D(y, \hat{\mu}) = \sum_{i=1}^{n} d_i . \tag{3.15}$$

Dies legt eine alternative Definition von Residuen nahe. **Deviance Residuen** sind definiert als

$$r_i^D = \operatorname{sign}(y_i - \hat{\mu}_i) \sqrt{d_i} . \tag{3.16}$$

Nun gilt für den MLE $\hat{\mu}$, welcher $\ell(\mu|y)$ maximiert, dass dieser die Deviance minimiert, da $\ell(y|y)$ ein fester Wert unabhängig vom Parameter ist. Wie die folgenden Beispiele zeigen, ist das Maß der Deviance gerade eine Verallgemeinerung der Fehlerquadratsumme beim linearen Modell unter Annahme normalverteilter Responses.

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

$$\begin{aligned}
\ell(\hat{\mu}|y) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \right\} \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \\
\ell(y|y) &= -\frac{n}{2}\log(2\pi\sigma^2)\,,
\end{aligned}$$

wobei wiederum $\ell(y|y)$ den maximal möglichen Wert unter dem vollen Modell beschreibt. Die skalierte Deviance ist somit

$$\frac{1}{\phi}D(y,\hat{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2}\text{SSE}(\hat{\beta})\,,$$

also eine exakt χ^2_{n-p}-verteilte Zufallsvariable mit Erwartungswert $n-p$. Dadurch motiviert wird diese Verteilungseigenschaft auch gerne für die skalierte Deviance anderer Mitglieder der linearen Exponentialfamilie – zumindest im approximativen Sinn – verwendet (die herkömmliche Likelihood-Quotienten Test Asymptotik kann hier nicht angewandt werden, da der Freiheitsgrad $n-p$ gleich schnell wächst wie die Anzahl der Beobachtungen n).

Beispiel: Seien $m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)$ mit $y_i = 0, 1/m_i, 2/m_i, \ldots, 1$, dann folgt

$$\begin{aligned}
\ell(\hat{\mu}|y) &= \sum_{i=1}^{n}\left\{ m_i y_i \log\frac{\hat{\mu}_i}{1-\hat{\mu}_i} - m_i\log\frac{1}{1-\hat{\mu}_i} + \log\binom{m_i}{m_i y_i} \right\} \\
\ell(y|y) &= \sum_{i=1}^{n}\left\{ m_i y_i \log\frac{y_i}{1-y_i} - m_i\log\frac{1}{1-y_i} + \log\binom{m_i}{m_i y_i} \right\}\,.
\end{aligned}$$

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

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

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

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

$$D(y,\hat{\mu}) = -2\phi\Big(q(\hat{\mu}|y) - q(y|y)\Big) = -2\sum_{i=1}^{n}\int_{y_i}^{\hat{\mu}_i}\frac{y_i - t}{V(t)}dt\,, \tag{3.17}$$

mit

$$q(\hat{\mu}|y) = \sum_{i=1}^{n} q(\hat{\mu}_i|y_i)\,.$$

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

Beispiel: Liegen Prozentwerte y_i als Responses vor (d.h. man hat keine Informationen über die entsprechenden Gesamtanzahlen), für welche die Varianzfunktion $\mu(1-\mu)$ zu groß erscheint, so kann alternativ als Varianzfunktion $V(\mu) = \mu^2(1-\mu)^2$ verwendet werden. Als Quasi-Likelihood Funktion erhält man für $0 < \hat{\mu} < 1$, $0 \le y \le 1$ und $\phi = 1$

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

Die entsprechende Quasi-Deviance ist daher

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

Bemerke, dass jetzt in $y_i = 0$ und $y_i = 1$ die Quasi-Deviance nicht einmal definiert ist, der Maximum-Quasi-Likelihood Schätzer $\hat{\beta}$ in diesem Fall jedoch sehr wohl existiert. Dieser ist wiederum definiert als iterative Lösung des Quasi-Score Gleichungssystems $\partial q(\mu(\beta)|y)/\partial\beta = 0$, das für die beiden kritischen Beobachtungen unproblematisch ist.

3.6 Maximum Quasi-Likelihood Schätzung

Man bemerke, dass für die Herleitung der folgenden Ergebnisse nur eine Notation verwendet wird, die sich ausschließlich auf die Annahme bzgl. der ersten beiden Momente der Responsevariablen stützt.

Nehmen wir nun an, dass unabhängige Responses y_i vorliegen, für die nur $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{var}(y_i) = \phi V(\mu_i)$ gilt. Postulieren wir weiters das strukturelle Modell $g(\mu_i) = x_i^\top\beta$ für den Erwartungswert μ_i, so ist der Maximum Quasi-Likelihood Schätzer $\hat{\beta}$ definiert als Nullstelle der Quasi-Scorefunktion

$$U(\beta|y) = \sum_{i=1}^{n}\frac{y_i - \mu_i}{\phi V(\mu_i)}\frac{\partial\mu_i}{\partial\beta}\,.$$

Wir verwenden jetzt die Matrizen $V = \mathrm{diag}(V(\mu_1),\ldots,V(\mu_n))$ und $D = \partial\mu/\partial\beta^\top$, die $(n \times p)$ Ableitungsmatrix von $\mu(\beta)$ nach dem Parameter β, und schreiben damit die

Quasi-Scorefunktion als

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

Wie bereits im Abschnitt 3.4 gezeigt, gilt dafür

$$\begin{aligned} \mathrm{E}(U(\beta|y)) &= 0 \\ \mathrm{var}(U(\beta|y)) &= -\mathrm{E}\left(\frac{\partial}{\partial \beta^\top} U(\beta|y)\right) = \frac{1}{\phi} D^\top V^{-1} D\,. \end{aligned} \tag{3.18}$$

Für Quasi-Likelihoodfunktionen spielt die Matrix (3.18) dieselbe Rolle wie die Fisher-Information bei herkömmlichen Likelihoodfunktionen. Im Speziellen ist die asymptotische Varianz/Kovarianzmatrix von $\hat{\beta}$ gleich

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

Wiederum verwenden wir zum Finden des Maximum Quasi-Likelihood Schätzers $\hat{\beta}$ die Newton-Raphson Methode mit Fisher-Scoring. Wir starten dazu die Iteration in $\beta^{(0)}$ und berechnen das erste Update als

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

Der Maximum Quasi-Likelihood Schätzer $\hat{\beta}$ resultiert bei Konvergenz. Man bemerke, dass die Iterationsvorschrift wiederum unabhängig vom Dispersionsparameter ϕ ist.

In allen betrachteten Aspekten verhalten sich Quasi-Likelihoods bzgl. der Schätzung von β genau so wie herkömmliche Log-Likelihoods. Eine Ausnahme bildet jedoch die Schätzung von ϕ. Daher verwenden wir auch in diesem Fall wieder die mittlere Pearson Statistik

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

3.7 Parametertests

Wie schon bei den linearen Modellen können auch jetzt wiederum die Parameter über das Konzept der **nested models** (ineinandergeschachtelte Submodelle) getestet werden. Dies entspricht dem Testen von Hypothesen der Form

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

$(q < p)$ oder äquivalent dazu einem (multiplen) Test der Hypothese

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

Sei M das betrachtete Modell (unter H_1 spezifiziert), und M_0 jenes unter H_0, das man erhält indem die letzten $p-q$ Parameter Null gesetzt werden. Somit ist M_0 ein Untermodell von M und wir schreiben $M_0 \subset M$. Bezeichnen $\hat{\mu}_0$ und $\hat{\mu}$ die geschätzten Erwartungen unter den beiden Modellen, dann erhalten wir als korrespondierende Deviancen

$$\begin{aligned} D(M_0) &= D(y, \hat{\mu}_0) = -2\phi\Big(\ell(\hat{\mu}_0|y) - \ell(y|y)\Big) \\ D(M) &= D(y, \hat{\mu}) = -2\phi\Big(\ell(\hat{\mu}|y) - \ell(y|y)\Big)\,. \end{aligned}$$

Die Differenz dieser beiden Deviancen ist gerade

$$D(M_0|M) = D(M_0) - D(M) = D(y, \hat{\mu}_0) - D(y, \hat{\mu}) = -2\phi\Big(\ell(\hat{\mu}_0|y) - \ell(\hat{\mu}|y)\Big) \qquad (3.19)$$

und beschreibt hierbei den Unterschied in der Anpassungsgüte dieser beiden Modelle. Die Differenz der skalierten Deviancen $D(M_0|M)/\phi$ entspricht der **Likelihood-Quotienten Teststatistik** um H_0 zu testen.
Ähnlich der Zerlegung der Fehlerquadratsumme bei linearen Modellen können wir die Deviance des eingeschränkten Modells zerlegen in

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

Der Term $D(M_0|M)$ beschreibt den Zuwachs in der Diskrepanz zwischen den Daten und der Modellanpassung, wenn das Modell M_0 anstatt des weniger restriktiven Modells M verwendet wird. Im Falle **normalverteilter Responses** entspricht dies der Zerlegung der Fehlerquadratsumme

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

Unter H_0 gilt hier, dass die beiden Terme auf der rechten Seite stochastisch unabhängig sind. Ferner gilt dafür $\mathrm{SSE}(M)/\sigma^2 \sim \chi^2_{n-p}$ und $(\mathrm{SSE}(M_0) - \mathrm{SSE}(M))/\sigma^2 \sim \chi^2_{p-q}$. Um H_0 zu testen verwenden wir bei einem klassischen linearen Modell die F-Statistik

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

Bei einem GLM betrachten wir im Falle eines bekannten Dispersionsparameter ϕ die Devianceredukti on

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

welche (unter geeigneten jedoch schwächeren Regularitätsbedingungen) asymptotisch χ^2-verteilt ist mit $p-q$ Freiheitsgraden. Ist ϕ unbekannt, so verwendeten wir wie bei den linearen Modellen die Teststatistik

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

wobei ϕ unter dem Alternativmodell H_1 geschätzt wird. Anstelle von $\hat{\phi} = D(y, \hat{\mu})/(n-p)$ kann auch die mittlere Pearson Statistik $X^2/(n-p)$ unter dem Alternativmodell zur Dispersionsschätzung verwendet werden.

Einen alternativen Zugang zu Hypothesentests bei statistischen Modellen stellt der sogenannte **Wald Test** dar. Dieser verwendet für den Fall einer simplen Hypothese der Form $H_0 : \beta_j = 0$, $j = 1, \ldots, p-1$, die auf den MLE $\hat{\beta}$ basierende Teststatistik

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

Dieses Konzept kann man aber auch verallgemeinern auf **allgemeine lineare Hypothesen** der Form

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

Hierbei ist C eine vorgegebene $s \times p$ Matrix (mit vollem Zeilenrang $s \leq p$) und ξ ein fester $s \times 1$ Vektor. Wir erinnern uns, dass für den MLE asymptotisch $\hat{\beta} \sim \text{Normal}(\beta, F^{-1}(\beta))$ gilt, mit der erwarteten Information (Fisher Information)

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

Somit gilt unter der Nullhypothese auch, dass $C\hat{\beta}$ erwartungstreu für $C\beta = \xi$ ist mit Varianz $\text{var}(C\hat{\beta}) = CF^{-1}(\hat{\beta})C^\top$. Es liegt somit nahe, die Statistik

$$(C\hat{\beta} - \xi)^\top \Big(\widehat{\text{var}}(\hat{\beta}) \Big)^{-1} (C\hat{\beta} - \xi) = (C\hat{\beta} - \xi)^\top \Big(CF^{-1}(\hat{\beta})C^\top \Big)^{-1} (C\hat{\beta} - \xi) \overset{H_0}{\sim} \chi^2_s \qquad (3.20)$$

zu betrachten. Diese Statistik misst daher die Distanz zwischen $C\hat{\beta}$ und $C\beta$, wobei mit der inversen (asymptotischen) Kovarianzmatrix $CF^{-1}(\hat{\beta})C^\top$ gewichtet wird.

Natürlich kann auch für die obige Hypothesensituation sofort die Idee des **Likelihood-Quotienten Tests** angewendet werden. Sei dazu die Statistik definiert als

$$-2\Big(\ell(\tilde{\beta}|y) - \ell(\hat{\beta}|y) \Big) \overset{H_0}{\sim} \chi^2_s \,, \qquad (3.21)$$

wodurch die Abweichung zwischen dem unrestringierten Maximum $\ell(\hat{\beta}|y)$ und dem unter H_0 restringierten Maximum $\ell(\tilde{\beta}|y)$ beschrieben ist, wobei $\tilde{\beta}$ der MLE unter der Restriktion $C\beta = \xi$ ist. Es ist hierbei also notwendig, die beiden Modelle (unter H_0 und unter H_1) anzupassen, um die Likelihood-Quotienten Teststatistik berechnen zu können.

Ein weiterer Ansatz verwendet die sogenannte **Score Teststatistik**. Wir erinnern uns, dass die Score Funktion $s(\beta) = (1/\phi)X^\top DW(y-\mu)$ für das unrestringierte Modell (unter H_1 spezifiziert) gerade der Nullvektor ist, falls diese im unrestringierten MLE $\hat{\beta}$ ausgewertet ist, also $s(\hat{\beta}) = 0$. Falls wir jedoch $\hat{\beta}$ durch den restringierten MLE $\tilde{\beta}$ ersetzen (unter H_0 berechnet), so wird $s(\tilde{\beta})$ sich genau dann signifikant vom Nullvektor unterscheiden, wenn

die Nullhypothese nicht zutrifft. Nun ist wegen $\mathrm{E}(s(\beta)) = 0$ die Varianz/Kovarianzmatrix des Score Vektors durch $\mathrm{var}(s(\beta)) = \mathrm{E}(s(\beta)s(\beta)^\top) = F(\beta)$ gegeben. Somit liegt es nahe, als Score Teststatistik

$$s(\tilde{\beta})^\top F^{-1}(\tilde{\beta}) s(\tilde{\beta}) \overset{H_0}{\sim} \chi^2_s \tag{3.22}$$

zu verwenden.

Ein Vorteil von der Wald- und Score-Teststatistik liegt darin, dass beide nur von den ersten beiden Momente abhängen. Alle drei Teststatistiken haben dieselbe asymptotische χ^2-Verteilung. Falls die Statistiken stark differieren, dann ist dies ein Hinweis darauf, dass die notwendigen Regularitätsbedingungen für die asymtotischen Resultate nicht halten und wohl verletzt sind.

3.8 Beispiel: Konstante Varianz

Im Beispiel über das verwendbare Holzvolumen von Black Cherry Bäumen aus Kapitel 1 wurde die Responsevariable V einer Box-Cox Transformation unterzogen. Für beide Modelle wurde dann angenommen, dass die transformierte Response $V^{1/3}$ oder $\log V$ aus einer Normalverteilung stammt. Alternativ dazu könnte man auch annehmen, dass V selbst bereits normalverteilt ist mit Erwartungswert μ, aber dass dieser Erwartungswert einem GLM folgt, d.h. dass $g(\mu) = \eta$ mit $g(\mu) \neq \mathrm{id}(\mu)$ gilt.

Unter der Annahme einer Normalverteilung (konstante Varianz) für V betrachten wir zuerst das Modell mit der Linkfunktion $g(\mu) = \mu^{1/3} = \eta$ (bzw. $\mu = \eta^3$), wobei der lineare Prädiktor durch $\eta = \beta_0 + \beta_1 H + \beta_2 D$ spezifiziert ist.

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

Call:  glm(formula = 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
```

Mittels `family` wird das konkrete Mitglied aus der linearen Exponentialfamilie spezifiziert. Erlaubt sind `binomial`, `gaussian`, `Gamma`, `inverse.gaussian`, `poisson`, oder `quasi`. Für jedes dieser fünf speziellen Mitglieder kann auch die Linkfunktion spezifiziert werden. Zur Verfügung stehen dazu beispielsweise für normalverteilte Responses `identity`, `log` und `inverse`. Eine weitere Alternative um damit die Linkfunktion zu spezifizieren, bietet die Funktion `power(lambda = 1)` an. Bei `family = quasi` muss zusätzlich noch die Vari-

anzfunktion spezifiziert werden. Für ein klassisches lineares Modell lautet beispielsweise der Aufruf `quasi(link = "identity", variance = "constant")`.

Unter dem `Null` Modell versteht man das Modell mit nur dem Intercept (entspricht einer iid Annahme). Als Freiheitsgrad resultiert dafür $n - 1 = 30$ bei einer Deviance von 8106. Diese Deviance verbessert sich auf 184.2 wenn die beiden Prädiktoren ins Modell aufgenommen werden, wodurch jedoch der Freiheitsgrad sich auf $n - p = 28$ verringert.

Unter AIC versteht man das Akaike Informationskriterium, definiert als

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

wobei k die Anzahl der Parameter im betrachteten Modell bezeichnet. Für unser Modell ist $k = 4$ (3 lineare Parameter β_0, β_1, β_2, und ein Dispersionsparameter ϕ).

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

> logLik(powermodel) # maximized log-likelihood function
'log Lik.' -71.60508 (df=4)
> sum(log(dnorm(V, powermodel$fit, sqrt(summary(powermodel)$dispersion*28/31))))
[1] -71.60508
> -2*logLik(powermodel) + 2*4
'log Lik.' 151.2102 (df=4)

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

Die Schätzer sind fast identisch mit jenen aus dem Beispiel in Abschnitt 1.3.

Wir können auch einen Log-Link anstelle der Annahme einer Lognormal-Verteilung für V verwenden, also die Linkfunktion $g(\mu) = \log \mu = \eta$ (bzw. $\mu = \exp(\eta)$) betrachten, jetzt für den linearen Prädiktor $\eta = \beta_0 + \beta_1 \log H + \beta_2 \log D$.

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

Call:  glm(formula = 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
```

Auch diese Schätzer entsprechen in etwa jenen unter der Lognormal-Verteilung in Abschnitt 1.3. Die Deviance ist auch nun wieder etwas geringer als beim obigen Ansatz.

3.9 Beispiel: Konstanter Variationskoeffizient

Das folgende Beispiel ist ausschließlich in Kontext eines **Quasi-Likelihood Modells** zu sehen. Wir werden zuerst nach einer passenden Varianzstruktur suchen und diese dann als Modellannahme für Anzahlen verwenden. Zur Erinnerung ist die Standardannahme bei derartigen Responses die Poissonverteilung. Diese ist Mitglied der linearen Exponentialfamilien und es ist die Equivarianzeigenschaft $\mathrm{E}(y) = \mathrm{var}(y)$ erfüllt. Beim folgenden Beispiel jedoch führen manche überzeugende Aspekte zur Annahme einer Varianz, die sich quadratisch hinsichtlich der Erwartung verhält, d.h. wir werden mit $\mathrm{var}(y) = \phi\mu^2$ (vergleichbar mit der Varianzstruktur unter einem Gammamodell) die Analysen durchführen. Für die Variationskoeffizienten unter einer derartigen Modellannahme ergibt sich

$$\frac{\sqrt{\mathrm{var}(y_i)}}{\mathrm{E}(y_i)} = \frac{\sqrt{\phi\mu_i^2}}{\mu_i} = \sqrt{\phi}\,, \qquad i = 1, \ldots, n\,,$$

also sind diese konstant für alle Beobachtungen.

Ein Feldversuch über Verfahren zur Kontrolle von Insekten produziert $n = 140$ Insektenanzahlen (`counts`) von zwei unabhängigen Wiederholungen (`plots`) in jedem der 10 Blöcke (`block`) für jedes der sieben Verfahren (`treatment`). Dabei können die plots als Replikationen innerhalb der $(7 \times 10 = 70)$ Zellen der `treatment:block` Klassifikation (Interaktion, Wechselwirkung) betrachtet werden.

```
> 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)

> i.lm <- lm(counts ~ treatment*block)      # calculate 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
```

```
3  7.0 21.0 25.5  7.0 11 17.0 31.5 33.0 27.5 22.5
4 13.0 15.0 12.0  6.5  8 12.0 27.0 12.0 28.0 12.0
5 15.0 12.0 14.0  4.5 16 13.0 29.5 22.5 17.5 34.0
6 15.0  9.5 25.0  4.5 14 11.5 18.0 15.5 20.0 19.5
7 40.5 64.5 41.5 15.5 25 35.5 36.5 48.0 47.5 51.5

> 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)))

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

Welche Verteilung liegt den Responses zugrunde? In einer ersten Prüfung betrachten wir in jeder der 70 Zellen das Mittel der beiden `counts` und deren Standardabweichungen. Wir prüfen dazu die Residuen r eines linearen Models unter Normalverteilungsannahme für die `counts`. Das Histogramm der Residuen sieht gut aus (symmetrisch), während der QQ-Plot etwas längere Schwänze in der empirischen Verteilung zeigt als es bei der Normalverteilung der Fall ist (siehe Abbildung 3.1 oben). Auch im Scatterplot der Residuen gegen die angepassten Werte sieht man eine leicht zunehmende Dispersion der Residuen. Da gerade zwei Beobachtungen je Zelle vorliegen, ist dieser Plot symmetrisch um Null. Einfacher ist dies im Scatterplot der empirischen Momente zu erkennen (siehe Abbildung 3.1 unten/rechts).

Deutlich ist ein linearer Zuwachs der Standardabweichungen für wachsende Mittelwerte zu sehen, wenn man diese Schätzungen `treatment`-spezifisch durchführt und die resultierenden sieben Punkte darstellt (wie in der Abbildung 3.2). Da die Standardabweichungen proportional zu den Mittelwerten zu sein scheinen, legt dies ein quadratisches Modell für die Varianz der `counts` nahe. Dies erfüllt gerade die Gammaverteilung für die $\mathrm{var}(y) = \phi\mu^2$, und somit $\mathrm{sd}(y) = \sqrt{\phi}\,\mu$, gilt.

Die reparametrisierte Form der Gamma(μ, ϕ)-Verteilung hat außer dem Erwartungswert μ auch einen Dispersionsparameter ϕ, der durch die mittlere Pearsonstatistik geschätzt werden kann. Der kanonische Link der Gammaverteilung (in den Programm-Paketen als Defaultlink gesetzt) ist die reziproke (inverse) Funktion. Die Verwendung des Log-Links hat aber demgegenüber wesentliche Vorteile, vor allem was die Interpretation der Parameter betrifft.

Die Stufen der beiden kategoriellen Prädiktoren im Modell sollen derart kodiert werden, dass deren Parameter gerade die Abweichungen von der Basisstufe (`treatment=1`, `block=1`) beschreiben. Wir beginnen mit einem Modell, in dem jede Zelle durch eine Parameterkombination eindeutig identifiziert werden kann. Ein Prädiktor mit der Interaktion `treatment * block` erlaubt dies und es werden dadurch 70 Parameter generiert.

```
> summary(i.glmmax <- glm(counts ~ treatment * block, family = Gamma(link=log)))
```

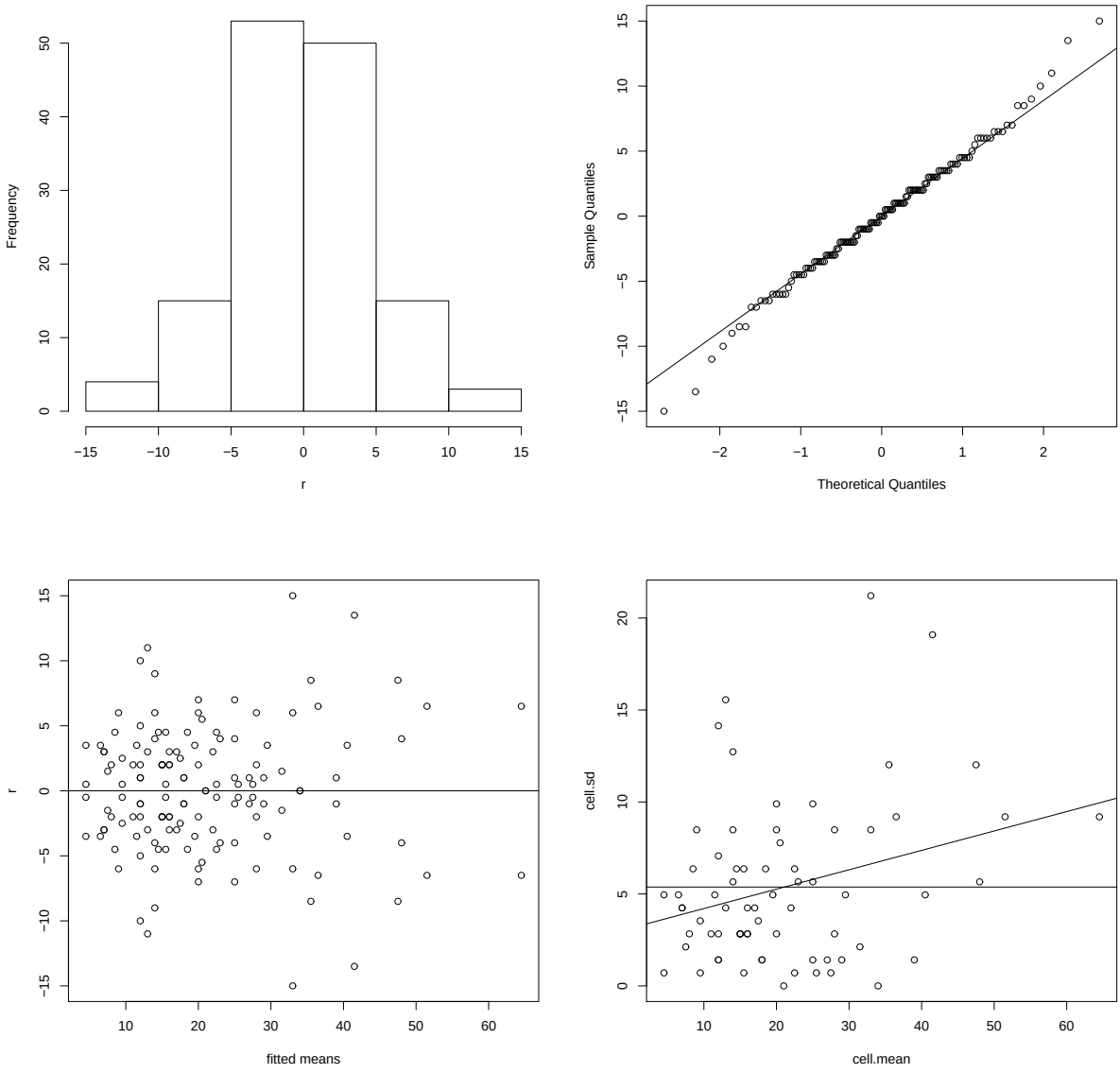

Abbildung 3.1: Oben: Histogramm der Residuen aus dem LM (links) und deren QQ-Plot (rechts). Unten: Residuen gegen arithmetische Mittel in den Zellen (links) und zellspezifische Standardabweichungen gegen Mittelwerte (rechts).

```
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 **
treatment3   -1.27297    0.42004   -3.03   0.0034 **
treatment4   -0.65393    0.42004   -1.56   0.1240
```

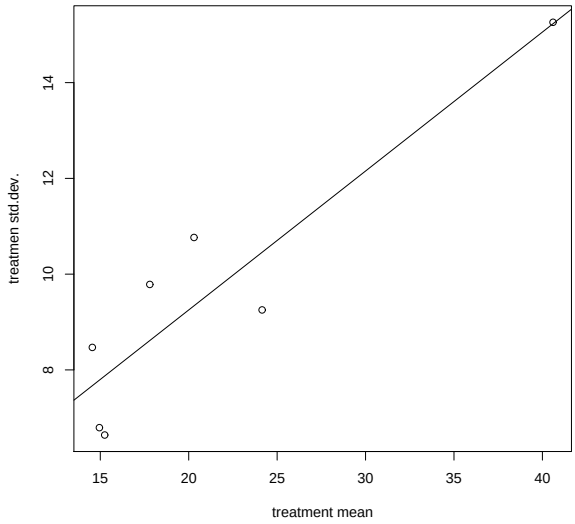

Abbildung 3.2: `treatment`-spezifische Standardabweichungen gegen Mittelwerte.

```
treatment5          -0.51083    0.42004   -1.22   0.2280
treatment6          -0.51083    0.42004   -1.22   0.2280
treatment7           0.48243    0.42004    1.15   0.2547
block2               0.44469    0.42004    1.06   0.2934
block3               0.27763    0.42004    0.66   0.5108
block4              -0.96758    0.42004   -2.30   0.0242 *
block5              -0.44629    0.42004   -1.06   0.2917
block6              -0.22314    0.42004   -0.53   0.5969
block7               0.11333    0.42004    0.27   0.7881
block8              -0.12783    0.42004   -0.30   0.7618
block9              -0.22314    0.42004   -0.53   0.5969
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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(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

Number of Fisher Scoring iterations: 6
```

Der Dispersionsparameter wird nun aus diesem Modell geschätzt und für die weiteren 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
```

Wir wollen nun untersuchen, ob vielleicht auch ein Submodell ausreicht und führen dazu eine *Analysis of Deviance* durch. Dieses Vorgehen entspricht genau der ANOVA beim linearen Modell. Als Teststatistik für eine Modellvergleich müssen wir die (durch denselben Dispersionsparameter) skalierte Differenz der Deviancen (Likelihood-Quotienten Test) verwenden. Dazu halten wir den Wert von $\hat{\phi}$ zuerst fest und berechnen:

```
> 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

> 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
```

Mit diesen Ergebnissen ist es nun möglich, die gesamte *Analysis of Deviance* zu rechnen (unter der Annahme, dass $\phi = \hat{\phi}$ bekannt ist). Während die Interaktion `treatment:block` zu vernachlässigen ist, sind die beiden Haupteffekte `block` und `treatment` hoch signifikant. Wir erhalten im Detail die Ergebnisse in der Tabelle 3.1

Viel einfacher ist es, dafür wiederum die Funktion `anova` zu verwenden. Diese liefert äquivalent dazu die passende Zerlegung des maximalen Modells.

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

Model: Gamma, link: log
```

	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

Tabelle 3.1: Analysis of Deviance Tabelle des Gammamodells mit Log-Link.

```
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
```

Möchte man die Dispersion frei schätzen, dann sollte entsprechend als 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 17.2080 4.124e-12 ***
block           9  16.2892       124     25.319 10.2581 6.647e-10 ***
treatment:block 54   9.2049        70     16.115  0.9661    0.5489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Wir bleiben beim Modell mit den beiden Haupteffekten (ohne Interaktion). Aus dem `summary` Objekt `s.glmmain` erhält man u.a. die Parameterschätzer und deren Standardfehler.

```
> 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
block4       -0.7355     0.1527 -4.8181   0.0000
block5       -0.1531     0.1527 -1.0031   0.3178
block6        0.0574     0.1527  0.3762   0.7074
block7        0.5526     0.1527  3.6199   0.0004
block8        0.4060     0.1527  2.6597   0.0089
block9        0.4536     0.1527  2.9717   0.0036
block10       0.4293     0.1527  2.8123   0.0057
```

Die unter diesem Modell geschätzten `treatment` Erwartungen und ihre (punktweisen) 95% Konfidenzintervalle können für jeden `block` berechnet werden. Wir nehmen `block 1` und machen zuerst eine Vorhersage für sämtliche Stufen von `treatment`. Dies führt zu den geschätzten Erwartungswerten mit Konfidenzlimits in der Abbildung 3.3.

```
> (new.i <- data.frame(treatment=levels(treatment), block=factor(rep(1,7))))
  treatment block
1         1     1
2         2     1
3         3     1
4         4     1
5         5     1
6         6     1
7         7     1
> options(digits=4)
> (i.pred <- predict(i.glmmain, newdata = 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
```

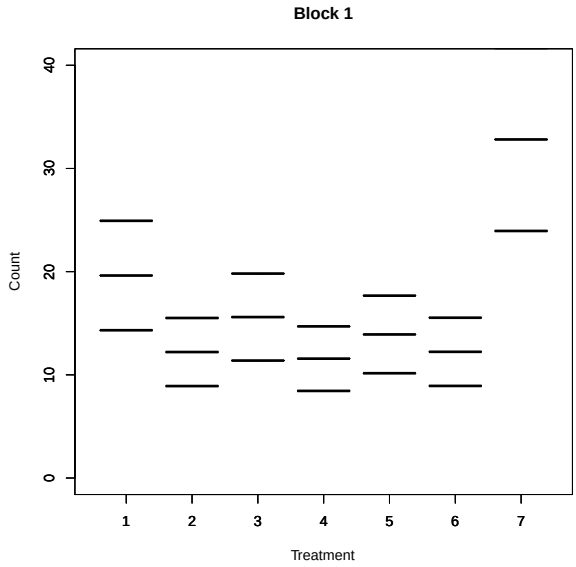

Abbildung 3.3: Geschätzte Erwartungswerte mit 95% Konfidenzlimits für das log-lineare Gammamodell mit Haupteffekten am Beispiel Block 1.

```
$residual.scale
[1] 0.4039
```

```
> 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$treatment, fit,   style="box", ylim=c(0.0,40.0))
> par(new=TRUE); plot(new.i$treatment, lower, style="box", ylim=c(0.0,40.0))
```

Nur wenn `se.fit = TRUE` in `predict` gesetzt ist, bekommt man auch die Standardfehler zu den geschätzten Erwartungen. Unter `residual.scale` wird hier die Wurzel von $\hat{\phi}$ verstanden, d.h.

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

Wie es scheint sind die Unterschiede in den `treatment` Stufen hauptsächlich auf die geringere Effektivität von `treatment` 1 und 7 zurückzuführen (große erwartete Anzahlen). Die mit `treatment` 7 (einziger positiver `treatment` Parameter) behandelten Einheiten weisen die höchste Anzahl von Insekten auf.

Aus den geschätzten Koeffizienten ist ersichtlich, dass sich für alle anderen Blöcke zwar dasselbe Muster ergibt (da keine Wechselwirkungen im Modell), jedoch die Blöcke 4 und 5

im Mittel geringere Anzahlen als Block 1 aufweisen. Alle übrigen Blöcke (positive Blockparameter) ziehen sogar etwas mehr Insekten an als der Block 1 (neutraler Nullparameter, Referenzkategorie).

Kapitel 4

Logistische Regression

Seien dazu $m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)$. Wie bereits gezeigt, gilt dafür

$$\begin{aligned} \text{E}(y_i) &= b'(\theta_i) = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = \mu_i \\ \text{var}(y_i) &= a_i b''(\theta_i) = \frac{1}{m_i}\frac{\exp(\theta_i)}{(1+\exp(\theta_i))^2} = \frac{1}{m_i}\mu_i(1-\mu_i) \end{aligned}$$

mit festem Dispersionsparameter $\phi = 1$. Als kanonische Linkfunktion $g(\mu) = b'^{-1}(\mu) = \theta$ resultiert der **Logitlink**

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

Bemerkung 4.1. Diese Bezeichnung bezieht sich auf die Verteilungsfunktion einer logistisch verteilten Zufallsvariablen mit Dichte

$$f(y|\mu,\tau) = \frac{\exp((y-\mu)/\tau)}{\tau\Big(1+\exp((y-\mu)/\tau)\Big)^2}\,, \qquad \mu\in\mathbb{R},\ \tau>0\,. \tag{4.1}$$

Für eine derart verteilte Zufallsvariable gilt $\text{E}(y) = \mu$ und $\text{var}(y) = \tau^2\pi^2/3$.

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

Prinzipiell kann in dieser Modellklasse wegen $\mu = g^{-1}(\eta)$ und $0 < \mu < 1$ anstelle des inversen Logitlinks auch eine beliebige andere stetige Verteilungsfunktion verwendet werden. So spricht man beispielsweise bei der Wahl $g^{-1}(\eta) = \Phi(\eta)$ von einem **Probitmodell**. Sowohl Logitlink als auch Probitlink sind symmetrische Linkfunktionen.

Manchmal werden aber auch Extremwertverteilungen (Typ 1) betrachtet. Dazu zählt die Maximum–Extremwertverteilung mit (Standard-) Verteilungsfunktion

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

Die ersten beiden Momente dafür sind $\mathrm{E}(y) = \gamma$ (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 Linkfunktion $g(\mu) = -\log(-\log(\mu))$.

Demgegenüber sprechen wir von einer Minimum–Extremwertverteilung, wenn $-y$ einer Maximum-Extremwertverteilung genügt und somit die Verteilungsfunktion folgende Gestalt hat:

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

Hierfür ist $\mathrm{E}(y) = -\gamma$ 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.

Generell erlaubt `R` für `family=binomial` als Spezifikation der Linkfunktion `logit`, `probit`, `cauchit`, (entsprechen den Verteilungsfunktionen einer logistischen, Normal und Cauchy Verteilung) sowie `log` und `cloglog` (complementary log-log).

Diese Linkfunktionen bilden den linearen Prädiktor folgendermaßen auf die Wahrscheinlichkeitsskala ab:

```
> euler.gamma <- 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.gamma + 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)
```

Als Log-Likelihood Funktion und für die (skalierte) Deviance erhält man

$$\begin{aligned}
\ell(\hat{\mu}|y) &= \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i} - m_i \log \frac{1}{1-\hat{\mu}_i} + \log \binom{m_i}{m_i y_i} \right\}\,, \\
D(y, \hat{\mu}) &= 2 \sum_{i=1}^{n} m_i \left\{ (1-y_i) \log \frac{1-y_i}{1-\hat{\mu}_i} + y_i \log \frac{y_i}{\hat{\mu}_i} \right\}\,.
\end{aligned}$$

Ein Spezialfall der logistischen Regression ist die Modellierung binärer Daten $y_i \in \{0, 1\}$ ($m_i = 1$ für alle i). Hier gilt für den relevanten Teil der Log-Likelihood Funktion

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

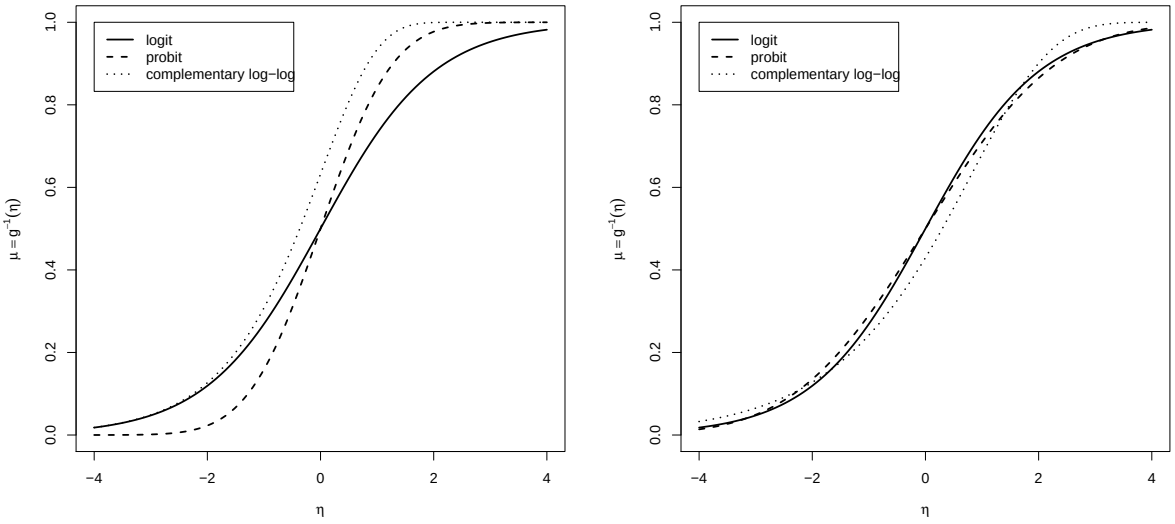

Abbildung 4.1: Responsefunktionen (links) und adjustierte Responsefunktionen (rechts) im binären Regressionsmodell.

und

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

Das Deviance-Inkrement d_i beschreibt also den Anteil einer binären Beobachtung an der Log-Likelihood Funktion der Stichprobe

$$\ell(\mu|y) = \sum_{i=1}^{n} \ell(\mu_i|y_i) = -\frac{1}{2}\sum_{i=1}^{n} d_i\,.$$

4.1 Toleranzverteilungen – Linkfunktionen

Historisch gesehen entwickelten sich die Modelle für binäre Responses aus der Notwendigkeit einer experimentellen Untersuchung, dem biologischen Assay oder kurz **Bioassay**. Typischerweise wurde hierbei die Wirkung verschiedener Konzentrationen einer chemischen Verbindung in Tierversuchen erprobt. Die Anzahl der Tiere, welche darauf angesprochen haben wurde aufgezeichnet und als Realisierung einer binomialen Response betrachtet.

So wird zum Beispiel in einem insektentötenden Versuch ein spezielles Insektizid in mehreren bekannten Konzentrationen auf Insektengruppen (**batches**) von bekannten Umfängen angewendet. Falls einer Insektengruppe eine sehr niedrige Dosis verabreicht wird, wird wahrscheinlich keines dieser Insekten daran versterben. Wird jedoch einer anderen Insektengruppe eine sehr hohe Dosis verabreicht, so werden sehr viele (wenn nicht alle) daran versterben. Ob nun ein spezielles Insekt bei einer gegebenen Dosis verstirbt oder nicht,

hängt von der **Toleranz** dieses Individuums ab. Insekten mit einer geringen Toleranz werden bei einer gewissen Dosis eher sterben als andere mit einer großen Toleranz.

Nun nimmt man an, dass es für die gesamte Insektenpopulation (aus der auch die Versuchstiere stammen) eine Verteilung der Toleranz gibt. Insekten mit einer Toleranz kleiner als d_i werden bei einer verabreichten Dosis d_i daran sterben. Sei U die Zufallsvariable, welche mit der Toleranzverteilung assoziiert ist, und sei u die Toleranz eines speziellen Tieres. Die entsprechende Dichtefunktion sei $f(u)$. Somit ist die Wahrscheinlichkeit eines Insektes, bei einer Dosis d_i daran zu sterben, gegeben durch

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

Falls nun die Toleranz normalverteilt ist mit Erwartungswert μ und Varianz σ^2, so folgt dafür

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

Bezeichnet $\beta_0 = -\mu/\sigma$ und $\beta_1 = 1/\sigma$, dann erhält man

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

also ein **Probitmodell** für die Beziehung zwischen der Sterbewahrscheinlichkeit p_i und der dafür verabreichten Dosis des Insektizids d_i.

Für U logistisch verteilt mit Dichte (4.1) entspricht dies

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

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

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

Die Annahme einer logistischen Toleranzverteilung führt zu einem **logistischen Modell** für p_i.

4.1.1 Beispiel (Venables & Ripley)

Untersucht wird die Wirkung eines Giftes auf einen Wurm, der die Knospen von Tabakpflanzen schädigt. Gruppen zu je 20 Motten beiderlei Geschlechts (`sex`) wurden dazu drei Tage lang verschieden hohen Dosen eines Giftes ausgesetzt und die Anzahl der daran verstorbenen Tiere wurde 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

Da die Dosen Potenzen von 2 sind, verwenden wir als Prädiktorvariable $\log_2$(Dosis) zusätzlich zum Geschlecht.

```
> (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
```

Die Spezifikation binomialer Daten kann in `R` unterschiedlich gemacht werden. Wir werden dazu die zweispaltige Matrix `SF` (success/failure) verwenden, deren erste (zweite) Spalte die Anzahl der Erfolge (Misserfolge) enthält. Das resultierende Modell schätzt dann die Erfolgswahrscheinlichkeit, in unserem Fall also die Wahrscheinlichkeit, dass eine Motte an der verabreichten Dosis verstirbt.

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

Number of Fisher Scoring iterations: 4
```

Eine alternative Art der Kodierung ist die Verwendung eines numerischen Vektors mit Eintragungen der Form s_i/a_i, wobei a_i die Anzahl der Versuche beinhaltet und s_i die Anzahl der Erfolge. Die Werte von a_i selbst müssen in diesem Fall mittels `weights` spezifiziert werden.

```
> 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
```

Das Ergebnis weist auf eine signifikante Steigung (`ldose`) hin. Die erste Stufe von `sex` bezieht sich auf die weiblichen Tiere (alphabetische Reihenfolge der Stufen, `"F"` vor `"M"`) und wird somit durch den Intercept beschrieben. Der Parameter `sexM:ldose` repräsentiert einen (nicht signifikant) größeren Anstieg für männliche Tiere, während der Parameter `sexM` den (nicht signifikant) Unterschied in den Intercepts beschreibt. Die Daten und die Modellschätzung sind in der Abbildung 4.2 dargestellt.

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

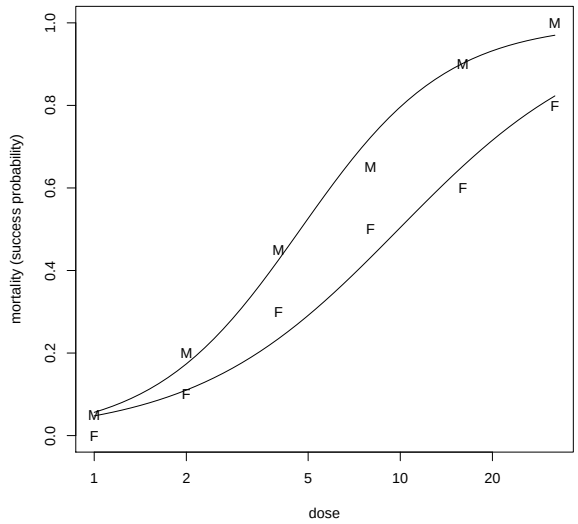

Abbildung 4.2: Beobachtete und geschätzte Ausfallsraten beim Tabakwurm.

Der Parameter `sexM` scheint irrelevant zu sein. Dieser beschreibt jedoch gerade den Unterschied bei der Dosis 1μg ($\log_2(\text{Dosis}) = 0$). Ist man am Unterschied bei 8μg interessiert, erhält man

```
> 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
```

Also doch ein signifikanter Geschlechtsunterschied bei einer Dosis von 8μg. Das Modell selbst scheint ausgezeichnet zu passen (Deviance von 5 bei 8 Freiheitsgraden). Die Analysis of Deviance Tabelle bestätigt dies. Auf eine Interaktion kann verzichtet werden. Auch die Hinzunahme eines quadratischen Terms ist nicht notwendig.

```
> 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

> 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
```

Diese Analyse empfiehlt daher ein Modell mit zwei parallelen Geraden bezüglich der Prädiktor- (logit)-Achse, eine für jedes Geschlecht.

Oft ist man auch an einer Schätzung der Dosis für eine vorgegebene Ausfallswahrscheinlichkeit interessiert. Dazu reparametrisieren wir zuerst das Modell, sodass wir für jedes Geschlecht einen eigenen Intercept haben.

```
> 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

(Dispersion parameter for binomial family taken to be 1)

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

Number of Fisher Scoring iterations: 4
```

Sei nun ξ_p jener Wert von $\log_2$(Dosis) für den die Wahrscheinlichkeit eines Erfolgs gerade p ist. Also beschreibt $2^{\xi_{0.5}}$ die 50% Ausfallsdosis (**50% lethal dose** oder **LD50**). Für diese Dosis hält bei einer allgemeinen Linkfunktion $g(p) = \beta_0 + \beta_1 \xi_p$ gerade

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

Nun hängt die Dosis vom Parameter $\beta = (\beta_0, \beta_1)^\top$ ab, also $\xi_p = \xi_p(\beta)$. Ersetzt man den Parameter durch seinen Schätzer, so liefert dies den Schätzer $\hat{\xi}_p = \xi_p(\hat{\beta})$ wofür wiederum approximativ gilt

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

Da $\hat{\beta}$ erwartungstreu ist für β, folgt approximativ auch $\mathrm{E}(\hat{\xi}_p) = \xi_p$. Darüberhinaus resultiert als approximative Varianz nach der Delta-Methode

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

wobei

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

Die Funktion `dose.p` aus der Bibliothek `MASS` benutzt die obigen Resultate und liefert für weibliche und für männliche Motten das folgende Ergebnis:

```
> require(MASS)
> options(digits=4)
> 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

> 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
```

Man benötigt also eine geschätzte Dosis von $\log_2(\text{Dosis}) = 3.26$, also Dosis $= 9.60$, damit 50% der weiblichen Schädlinge ausfallen, und Dosis $= 4.69$ für 50% der männlichen.

Alternativ kann für diesen Bioassay auch ein Probitmodell gerechnet werden. Diese Linkfunktion ist im zentralen Bereich sehr ähnlich dem Logitlink und unterscheidet sich nur marginal in den Schwänzen. Man erhält unter diesem Modell 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
```

4.2 Interpretation der Parameter

Hängt der Erwartungswert binärer Daten nur von einem erklärenden zweistufigen Faktor x ab, so können die Parameter sehr einfach interpretiert werden. Dazu betrachtet man die folgende Kontingenztafel mit den Zellwahrscheinlichkeiten:

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

Im Falle $x = 1$ bezeichnet $\pi_1/(1 - \pi_1)$ die **Quote, Chance (odds)** für das Eintreten von $y = 1$ zu $y = 0$. Die Log-Transformation dieser Quote bei $x = 1$ ist

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

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

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

Bezeichne nun $\mu(x) = \Pr(y=1|x)$ und $1-\mu(x) = \Pr(y=0|x)$ mit $x \in \{0,1\}$. Falls dafür zusätzlich das Modell

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

hält, so bekommt man für die Zellwahrscheinlichkeiten die folgende Struktur:

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

Für diese Situation erhält man als log-odds ratio

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

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

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

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

4.2.1 Beispiel (Agresti)

An 27 Krebspatienten wird beobachtet, ob es nach einer Injektionsbehandlung zum Abklingen des Karzinoms kommt. Die wichtigste erklärende Variable `LI`, der Labelling Index, beschreibt die Zellteilungsaktivität nach dieser Behandlung (Anteil der Zellen, welche "gepolt" wurden). Dabei ergaben sich $n = 14$ unterschiedliche `LI` Werte. Die Response-Variablen $m_i y_i$ beschreiben die Anzahl erfolgreicher Rückbildungen bei m_i Patienten mit jeweiliger Zellaktivität $\texttt{LI}_i$. Die Datensituation ist wie folgt:

$\texttt{LI}_i$	m_i	$m_i y_i$	$\texttt{LI}_i$	m_i	$m_i y_i$	$\texttt{LI}_i$	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			

Es wird angenommen, dass sich die Patienten in einer $\texttt{LI}_i$ Gruppe homogen verhalten, d.h. wir betrachten nicht 27 einzelne Bernoulli Variablen sondern deren 14 gruppenspezifischen Summen (absolute Häufigkeiten)

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

Als MLE erhält man dafür

```
> 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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
AIC: 24.29

Number of Fisher Scoring iterations: 4
```

Nimmt LI um eine Einheit zu, so wird die geschätzte Quote für eine Rückbildung mit $\exp(0.145) = 1.156$ multipliziert. Somit nimmt durch Erhöhung des Index um Eins die Quote einer Rückbildung um 15.6% zu. Die unter dem Modell geschätzte Wahrscheinlichkeit für eine Rückbildung ist gerade dann 1/2, falls $\hat{\eta} = 0$, also für einen Index von $\texttt{LI} = -\hat{\beta}_0/\hat{\beta}_1 = 26.07$.

Das Mittel aller 27 LI-Werte ist $\sum_i \texttt{LI}_i m_i / \sum_i m_i = 20.07$, wofür sich als linearer Prädiktor $\hat{\beta}_0 + \hat{\beta}_1 20.07 = -0.8691$ ergibt, was einer Rückbildungswahrscheinlichkeit von 29.54% entspricht. Beobachtet konnten 9 Erfolge aus 27 Patienten werden, also 33.33%.

Für die logistische Regressionskurve gilt $\partial\mu(x)/\partial x = \beta_1\mu(x)(1-\mu(x))$. Der stärkste Anstieg ist daher an der Stelle $\mu(x) = 1/2$, also im Index $\texttt{LI} = 26.07$, und beträgt dort $\hat{\beta}_1/4 = 0.0362$.

Es stellt sich auch die Frage, ob die Rückbildung signifikant vom LI-Wert abhängt. Der entsprechende p-Wert beim Wald-Test von 1.46% scheint dies klar zu bestätigen.

Für das iid Zufallsstichprobenmodell resultiert als (NULL) Deviance 23.96 bei $df = 13$ Freiheitsgraden. Die Deviance-Differenz 8.30 bei einem Verlust von einem Freiheitsgrad

entspricht einem $\chi^2_{1;1-\alpha}$-Quantil mit $\alpha = 0.004$. Dieses Ergebnis ist noch deutlicher als das Resultat des Wald-Tests. In beiden Fällen wird die signifikante (positive) Assoziation zwischen dem Index `LI` und der Rückbildungswahrscheinlichkeit bestätigt.

```
> 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
```

Modelliert man die Resultate der einzelnen Patienten als Bernoullivariablen, so liefert dieses Vorgehen wiederum dieselben Parameterschätzer, aber andere Werte für die Deviancen und deren Freiheitsgrade. Die Deviancedifferenz ist jedoch dieselbe wie zuvor bei binomialen Responses.

```
> index <- rep.int(li, times=total)
> backB <- 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(backB ~ 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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom
AIC: 30.07

Number of Fisher Scoring iterations: 4

> 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
```

Man bemerke, dass auch durch diesen Aufruf wiederum die Wahrscheinlichkeit für $y = 1$ (Rückbildung) modelliert wird. Da bei Bernoullivariablen deren Gewichte alle Eins sind, ist deren explizite Angabe durch `weights` nicht notwendig.

4.3 Logit-Modelle

In diesem Spezialfall der logistischen Regression bei binären Beobachtungen sind nur Faktoren und keine Variablen im linearen Prädiktor enthalten.

4.3.1 Beispiel

Untersuchungen von 313 Frauen nach einer Operation des Zervixkarzinoms an der Universitätsfrauenklinik in Graz ergaben 123 Rezidive. Von welchen Risikogrößen hängt der Rezidiveintritt ab? Die Anzahl der befallenen Lymphknotenstationen (`LK`) und der Befall der Zervix-Grenzzone (`GZ`) wurden zum Zeitpunkt der Operation festgestellt und stellen potentielle Risikogrößen dar. In der folgenden Tabelle sind die Rezidivfälle (y_{ij}/m_{ij}) als 3×4 Kontingenztafel angegeben.

	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

Falls ein lineares Verhalten beider Risikofaktoren gewährleistet ist, kann `LK` beispielsweise mit $0, 1, 2, 3$ und `GZ` mit $0, 1, 2$ kodiert werden und wir betrachten das logistische Modell für die Daten in diesen $n = 12$ Zellen, d.h.

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

```
> 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))
 [1] 0 1 2 3 0 1 2 3 0 1 2 3
> (GZ <- rep(0:2, each=4))
 [1] 0 0 0 0 1 1 1 1 2 2 2 2
> SF <- cbind(nonrez=total-rez, rez)
> summary(rez.glm <- glm(SF ~ LK + GZ, family = binomial))

Coefficients:
            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 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 103.493  on 11  degrees of freedom
Residual deviance:  13.332  on  9  degrees of freedom
AIC: 51.21

Number of Fisher Scoring iterations: 4
```

Kann man Linearität in den beiden Prädiktoren nicht gewährleisten, so ist es besser zwei Faktoren L und G zu verwenden. Sei dazu $m_{ij}y_{ij} \sim \text{Binomial}(m_{ij}, \mu_{ij})$, wobei μ_{ij} die Wahrscheinlichkeit rezidivfrei zu sein bei LK-Status i ($i = 1, 2, 3, 4$) und GZ-Befall j ($j = 1, 2, 3$) beschreibt. Wir unterscheiden 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.

```
> 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")
Analysis of Deviance Table

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 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aus der Analysis of Deviance folgt, dass die Interaktion zwischen L und G nicht wirklich relevant zu sein scheint, und das minimal notwendige Modell somit das mit den beiden Haupteffekten ist, also $\mathtt{L} + \mathtt{G}$. Dafür resultieren als ML Schätzungen

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

Vergleicht man dieses Ergebnis mit den Schätzungen der beiden Steigungen zu den linearen Einflussgrößen LK und GZ von früher, so ist zu erkennen, dass 1.069 als negativer Slope zu LK in etwa vergleichbar ist mit 1.287, $2 * 1.069 = 2.138$ die geschätzte L2 Stufe überschätzt und $3 * 1.069 = 3.207$ die geschätzte L3 Stufe unterschätzt. Diese Erkenntnis sollte dazu motivieren, stets keine künstlich linearisierten Prädiktoren sondern Faktoren mit frei schätzbaren Stufen in das Modell aufzunehmen.

Das geschätzte Modell für den linearen Prädiktor mit den beiden Haupteffekten L und G ist somit

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

Alle Parameter zu höheren Faktorstufen beider Effekte sind negativ und nehmen monoton ab. Dies bedeutet, dass die Wahrscheinlichkeit rezidivfrei zu sein auch entsprechend abnimmt je mehr Lymphknoten befallen sind und je intensiver der Grenzzonenbefall ausgeprägt ist.

Die durch dieses Modell geschätzten Wahrscheinlichkeiten für Rezidivfreiheit sind

```
> 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
```

Im Vergleich dazu reproduziert das volle Modell gerade die beobachteten relativen 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
```

Wir wollen noch grafisch sämtliche geschätzten Modelle gegenüberstellen.

```
> 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")
```

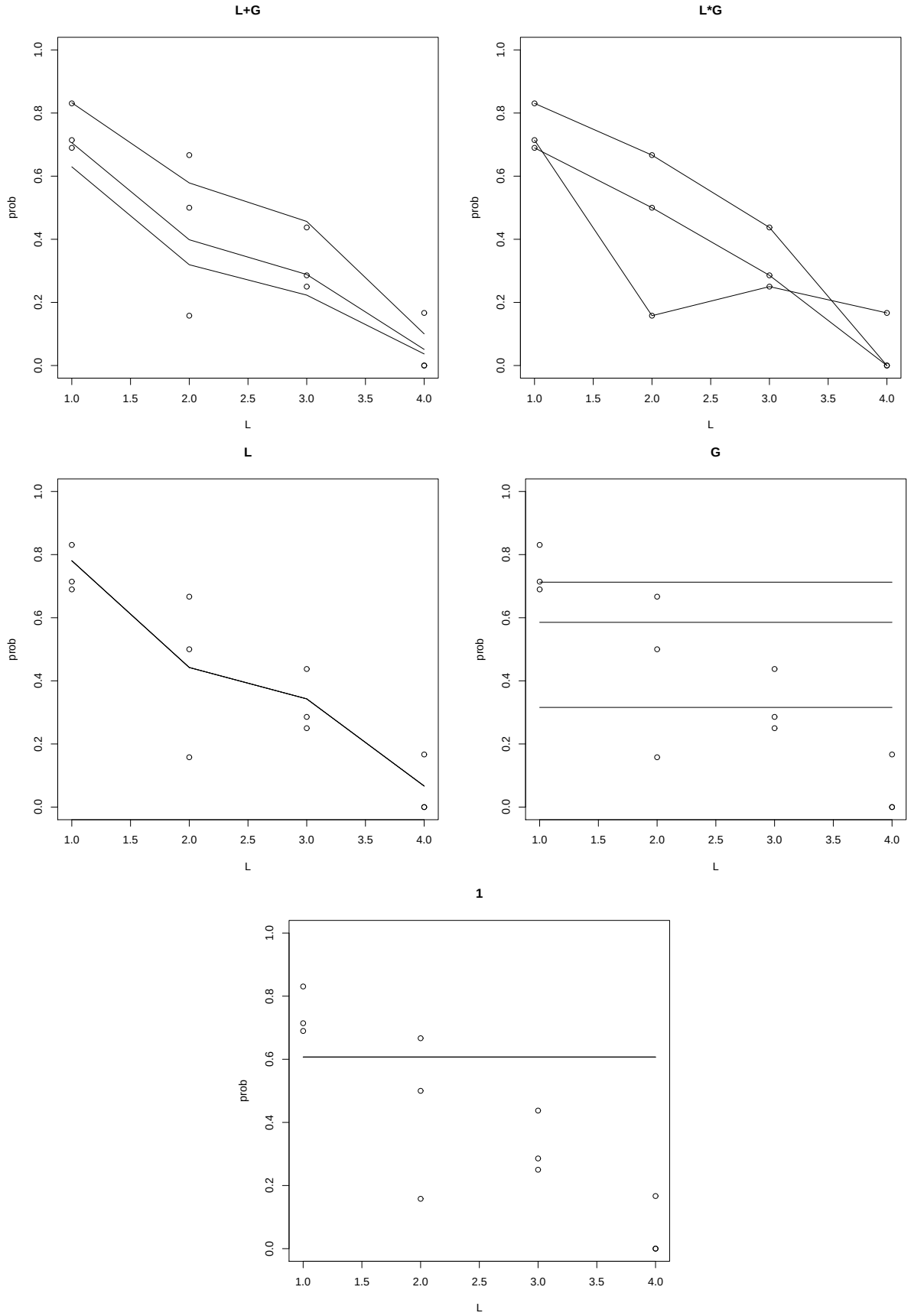

Abbildung 4.3: Geschätzte Wahrscheinlichkeiten für Rezidivfreiheit unter den betrachteten Logit-Modellen.

4.4 Überdispersion

Unter **Überdispersion** versteht man eine Situation, in der die tatsächliche Varianz der Responses die nominale Varianz unter dem betrachteten Modell übersteigt. Sehr viel seltener ist der Sachverhalt einer **Unterdispersion**. In unserem Fall modellieren wir absolute Häufigkeiten und vergleichen daher die Responsevarianz mit der binomialen Varianz.

Wie erkennt man Überdispersion?

Ist das logistische Modell korrekt, dann resultiert als asymptotische Verteilung der Deviance die χ^2_{n-p}-Verteilung. Somit kann $D(y, \hat{\mu}) > n - p = \mathrm{E}(\chi^2_{n-p})$ ein Hinweis auf Überdispersion sein.

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

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

Zuerst sollte man diese Gründe mittels explorativer Datenanalyse und diagnostischer Verfahren ausschließen.

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

4.4.1 Generelle Überlegungen

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

Lemma 4.1. Seien π und σ^2 reelle Zahlen. Es existiert genau dann eine auf $[0, 1]$ verteilte Zufallsvariable P mit vorgegebenen Momenten $\mathrm{E}(P) = \pi$ und $\mathrm{var}(P) = \sigma^2$, 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. Dasselbe gilt natürlich auch für eine standardisierte binomialverteilte Zufallsvariable y/m mit $\mathrm{E}(y/m) = \pi$. Mit dem obigen Lemma folgt dafür sofort

$0 \leq \mathrm{var}(y/m) \leq \pi(1-\pi)$. Für die korrespondierende (nicht-standardisierte) binomialverteilte Zufallsvariable y mit $\mathrm{E}(y) = m\pi$ gilt für alle ganzzahligen $0 \leq y \leq m$ die Abschätzung $\mathrm{E}(y^2) \geq \mathrm{E}(y)$ und somit $\mathrm{var}(y) \geq \mathrm{E}(y) - \mathrm{E}^2(y) = \mathrm{E}(y)(1-\mathrm{E}(y))$. Zusammen ergibt dies

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

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

$$\max(0, \mu(1-\mu)) \leq \mathrm{var}(y) \leq \mu(m-\mu) \tag{4.2}$$

als zulässiger Varianzbereich für y. Verwenden wir die für Dispersionsmodelle (relativ zum Binomialmodell) gebräuchliche Annahme $\mathrm{var}(y) = \phi m\pi(1-\pi)$, 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 $\mathrm{var}(y_i) = \phi m_i\pi_i(1-\pi_i)$ (bei gemeinsamen Dispersionsparameter ϕ) folgt somit als Restriktion

$$\max_{i=1,\ldots,n}\left(0, \frac{1-m_i\pi_i}{1-\pi_i}\right) \leq \phi \leq \min_{i=1,\ldots,n}(m_i)\,. \tag{4.3}$$

Das hat nun zur Folge, dass ϕ kleiner gleich allen Gruppenumfängen m_i sein muss. Wird in zumindest einer Experimentierumgebung lediglich eine Beobachtung gemacht, dann folgt für das Verteilungsmodell der gesamten Stichprobe, dass $\phi = 1$ sein muss. Dies wiederum bedeutet, dass in diesem Fall weder Über- noch Unterdispersion bezüglich des vorliegenden Modells betrachtet werden darf. Die folgende Überlegung liefert eine plausible Motivation für die Notwendigkeit eines Dispersionsparameters im Verteilungsmodell.

Dazu nimmt man an, dass in der i-ten Experimentierumgebung m_i binäre Bernoulli-Variablen $y_{i1}, \ldots, y_{im_i}$ beobachtet werden. Weiters wird angenommen, dass diese Bernoulli-Variablen denselben Erwartungswert $\mathrm{E}(y_{ij}) = \pi_i$ haben und somit $\mathrm{var}(y_{ij}) = \pi_i(1-\pi_i)$ gilt. Interessant ist nun zu untersuchen, wie eine **Korrelationsstruktur** von Paaren dieser Bernoulli-Variablen wirkt. Es ist oft naheliegend, in einer Gruppe für alle möglichen Paare denselben Korrelationskoeffizienten ρ anzunehmen mit

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

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

$$\begin{aligned}
\mathrm{E}(y_i) &= m_i\pi_i = \mu_i \\
\mathrm{var}(y_i) &= \sum_{j=1}^{m_i}\mathrm{var}(y_{ij}) + \sum_{j\neq k}^{m_i}\mathrm{cov}(y_{ij}, y_{ik}) \\
&= m_i\pi_i(1-\pi_i) + m_i(m_i-1)\pi_i(1-\pi_i)\rho \\
&= m_i\pi_i(1-\pi_i)\big(1 + (m_i-1)\rho\big) = \phi_i m_i\pi_i(1-\pi_i)\,.
\end{aligned} \tag{4.4}$$

Das wiederum bedeutet, dass sich die Varianz der Summe korrelierter Bernoullis verglichen mit der Varianz einer Binomial-Verteilung um den Faktor $\phi_i = \big(1 + (m_i - 1)\rho\big)$ unterscheidet. Positive Werte von ρ ergeben daher Überdispersion, während unterdispersionierte Daten von negativen ρ Werten stammen. Da jedoch $\phi \geq 0$ halten muss, ergibt sich als natürliche untere Schranke für den Korrelationskoeffizienten $\rho \geq -1/(m_i - 1)$. Ist $m_i = 1$ dann gibt es keine Dispersion, da hierfür $\phi_i = 1$ folgt.

4.4.2 Beta-Binomiale Varianz

Wir wissen, dass die Summe von homogenen Bernoulli-Variablen einer Binomialverteilung genügt. Verzichtet man jedoch auf die Homogenitätsforderung an die unabhängigen Bernoulli-Variablen, so erhalten wir folgende Aussage:

Lemma 4.2. 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 $s_\pi^2 = \frac{1}{m-1}\sum_j (\pi_j - \overline{\pi})^2$

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

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

Dies zeigt ganz offensichtlich, dass die Stichprobenvarianz von y verglichen mit der binomialen Varianz um den Term $(m-1)s_\pi^2$ kleiner ist. Intuitiv würde man jedoch bei einem Verstoß gegen die Homogenitätsannahme der y_j mit einem Ansteigen der Varianz von y rechnen. In der Praxis ist jedoch meist nur bekannt, dass eine Variabilität in den π_j's besteht, wobei die exakten Werte $p_1, \ldots, p_m$ dieser Parameter unbekannt sind.

Lemma 4.3. Seien $P_1, \ldots, P_m$ stetige, auf $[0,1]$ unabhängig verteilte Zufallsvariablen mit Dichte f_{P_j} und sei $\mathrm{E}(P_j) = \pi$. Sei weiters $y_j|(P_j = p_j) \overset{ind}{\sim}$ Binomial$(1, p_j)$. Für die Randverteilungen der y_j folgt dann, unabhängig von der exakten Verteilung der P_j, dass $y_j \overset{ind}{\sim}$ Binomial$(1, \pi)$, womit wiederum $y = \sum_{j=1}^m y_j \sim$ Binomial(m, π) resultiert.

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

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

Die gleiche Überlegung führt zu $P(y_j = 0) = 1 - \pi$. Somit gilt $y_j \overset{ind}{\sim}$ Binomial$(1, \pi)$. □

Unter den Voraussetzungen in Lemma 4.3 erhält man also trotz der Einbeziehung einer Variabilität in den P_j's wegen identischer Erwartungswerte wieder die binomiale Varianz. Nehmen wir nun eine spezielle stetige Verteilung für die latenten Variablen an, so folgt im Falle der Beta-Verteilung die folgende Aussage.

Lemma 4.4. Seien $P_1, \ldots, P_n$ unabhängige (nicht notwendigerweise identisch) Beta-verteilte Zufallsvariablen mit Dichtefunktionen

$$f_{P_i}(p) = \frac{1}{B(a_i, b_i)} p^{a_i - 1}(1-p)^{b_i - 1} \qquad \text{mit} \quad 0 < p < 1\,,\ a_i, b_i > 0\,,\ i = 1, \ldots, n\,.$$

Somit gilt dafür

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

Weiters seien $y_i | (P_i = p_i) \overset{ind}{\sim} \text{Binomial}(m_i, p_i)$. Für die Randverteilung von y_i resultiert nun die Beta-Binomial- (Negativ-Hypergeometrische)- Verteilung mit Wahrscheinlichkeitsfunktion

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

und

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

Bemerkung 4.2. Wir berechnen die marginale Wahrscheinlichkeitsfunktion als

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

Für die Berechnung der marginalen Momente kann man die bekannten Zusammenhänge zu den konditionalen Momente verwenden:

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

Somit resultiert

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

sowie

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

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

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

Reparametrisiert man die Wahrscheinlichkeitsfunktion der Beta-Binomial-Verteilung mittels $\gamma_i = 1/(a_i+b_i) > 0$ und $\pi_i = a_i/(a_i+b_i)$, so folgt dafür $a_i = \pi_i/\gamma_i$ und $b_i = (1-\pi_i)/\gamma_i$. Damit kann die Wahrscheinlichkeitsfunktion geschrieben werden als

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

(Die letzte Identität resultiert durch Multiplikation von Nenner und Zähler mit $\gamma_i^{m_i}$.) Als Grenzverteilung für $\gamma_i = 1/(a_i+b_i) \to 0$ erhalten wir somit die Binomial(m_i, π_i)-Verteilung

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

Bei unterschiedlichen Erwartungswerten der P_i zeigt sich also, dass im Falle $m_i > 1$ die Varianz der y_i gegenüber der binomialen Varianz um einen Faktor $\tau_i^2(m_i - 1)$ größer ist. Fordern wir anstelle der Beta-Verteilung nur eine bestimmte Relation zwischen Erwartungswert und Varianz, so führt dies zu

Lemma 4.5. $P_1, \ldots, P_n$ seien stetige, auf $[0,1]$ unabhängig verteilte Zufallsvariablen mit $\mathrm{E}(P_i) = \pi_i$ und $\mathrm{var}(P_i) = \phi^2 \pi_i (1-\pi_i)$, $0 < \phi^2 \leq 1$. 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

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

4.4.3 Beispiel: Klinischer Versuch

An 22 Kliniken wird die Wirkung eines neuen Medikaments mit einer Standardtherapie verglichen. Dazu wird in jeder Klinik jeweils eine Patientengruppe mit einem der beiden Mitteln behandelt. Die Behandlung wird als erfolgreich eingestuft, wenn eine unerwünschte Nebenwirkung selten auftritt.

```
> 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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 129.051  on 43  degrees of freedom
Residual deviance:  95.317  on 42  degrees of freedom
AIC: 161.62

Number of Fisher Scoring iterations: 5
```

Der `treatment` Effekt ist hochsignifikant und hat ein negatives Vorzeichen, was auf die deutlich bessere Wirkung des neuen Medikaments hindeutet. Andererseits ist der Wert der Deviance mit 95.317 um einiges größer als der dazugehörende Freiheitsgrad 42.

Würde man auch den Klinikeffekt in das Modell aufnehmen, so würde dies am Verhalten der Deviance verglichen mit dem Freiheitsgrad nicht viel ändern. Die folgende Tabelle liefert eine Übersicht über diese Modelle 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)

Wir interpretieren dieses Verhalten als Hinweis auf Überdispersion und betrachten ein Modell, das eine Varianzstruktur der Form $\text{var}(y_i) = \phi m_i \pi_i (1 - \pi_i)$ erlaubt. Dieser Quasi-

Likelihood Ansatz kann in R sehr 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 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.763606)

    Null deviance: 129.051  on 43  degrees of freedom
Residual deviance:  95.317  on 42  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Natürlich liefert dieses Modell dieselbe (unskalierte) Deviance und es werden die gleichen Schätzer generiert, jedoch sind deren Standardfehler um den Faktor $\sqrt{2.76} = 1.66$ größer. Dadurch reduziert sich zwar der p-Wert zum `treatment` Effekt, aber noch immer ist dieser hochsignifikant. Hierbei wird der in das Modell eingebrachte Dispersionsparameter ϕ mittels der mittleren Pearson-Statistik geschätzt, also durch

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

Ein logistisches Modell unter der beta-binomialen Varianzstruktur kann nicht mit `glm()` geschätzt werden. Dies liegt daran, dass diese Verteilung kein Mitglied der einparametrigen linearen Exponentialfamilie ist.

Eine ausgesprochen mächtige Bibliothek ist `gamlss` (Generalized Additive Models for Location Scale and Shape). In dieser Bibliothek sind Modelle für über 70 diskrete, stetige und gemischte Verteilungstypen implementiert, darunter auch die Beta-Binomial-Verteilung mittels `family = BB`. Hierbei wird eine Linkfunktion für den Erwartungswert $g_\mu(\mu) = \eta$, sowie eine weitere Linkfunktion für den Dispersionsparameter $g_\sigma(\phi) = \eta^*$ angeboten. Bei der Standardeinstellung wird der Logit-Link für μ und der Log-Link für ϕ verwendet.

```
> 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

----------------------------------------------------------------------
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
```

Der signifikante `treatment` Effekt ist hierbei etwas schwächer aber größenmäßig gut vergleichbar mit dem unter dem Binomial-Modell. Als Schätzung für ϕ erhalten wir 0.078 mit Standardfehler 0.0198. Der p-Wert von 0.00029 bewertet einen Test der Hypothese $H_0 : \phi = 0$, die wir daher auch sogleich verwerfen können. Somit liegen entweder variierende Wahrscheinlichkeiten vor, oder die einzelnen Bernoullivariablen, die das Verhalten der Patienten von ein und derselben Klinik beschreiben, sind signifikant korreliert.

Alternativ kann auch die Bibliothek `aod` (Analysis of Overdispersed Data) für Schätzung der Parameter in einem Beta-Binomial-Modell eingesetzt werden.

```
> library(aod)
> betabin(cbind(responses, size-responses) ~ treatment, ~1, data=clinics)
Beta-binomial model
-------------------
betabin(formula = cbind(responses, size - responses) ~ treatment,
    random = ~1, data = clinics)

Convergence was obtained after 80 iterations.

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
```

Die 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.

Kapitel 5

Poisson Regression

5.1 Poisson Loglineare Modelle für Anzahlen

Bei binomialverteilten Responses wird die Anzahl der Versuche vor der Durchführung der Experimente fixiert. Danach wird erst beobachtet, in wie vielen Fällen davon ein Ereignis eingetreten ist. Man spricht bei dieser Art Daten von der Modellierung relativer oder absoluter Häufigkeiten. Wird jedoch die Anzahl der Versuche vor den Experimenten nicht fixiert, so spricht man von Zählvariablen oder Anzahlen y_i. Standardmäßig wird dafür die Poissonverteilung angenommen, für die die Erwartung der Varianz entspricht, $\mathrm{E}(y_i) = \mu_i = \mathrm{var}(y_i)$, und als kanonischer Link der Loglink verwendet, also

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

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

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

Falls das Modell auch einen Intercept enthält, so reduziert sich die Deviance auf

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

Der Devianceanteil der i-ten Beobachtung verschwindet für $y_i = 0$ unabhängig von $\hat{\mu}_i$.

5.1.1 Beispiel: Modellierung von Anzahlen

Die folgenden Daten stammen aus einer Studie über die Aufbewahrbarkeit von Mikroorganismen im tiefgekühlten Zustand(-70^oC). Die Beobachtungen beschreiben die Bakterienkonzentrationen (Anzahl auf einer konstanten Fläche) am Anfang und nach 1, 2, 6, und 12 Monaten.

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

Ein passendes Modell soll gefunden werden, mit dem die Konzentrationsänderung in Abhängigkeit von der Zeit beschrieben werden kann, um Aussagen darüber zu ermöglichen, wie lange Bakterien durchschnittlich aufbewahrt werden können. Es besteht die Vermutung, dass die durchschnittliche Anzahl jeweils proportional zu $1/\texttt{time}^{\gamma}$ ist.

```
> 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))
```

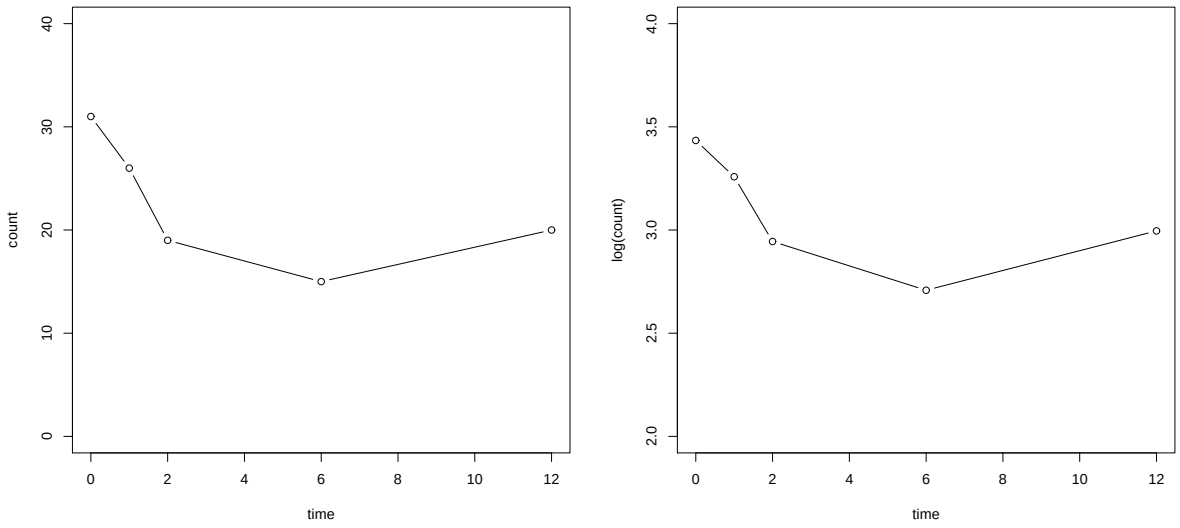

Abbildung 5.1: Bakterienanzahlen und deren Logarithmus in Abhängigkeit von der Zeit.

Eigentlich hätte man in den Daten, dargestellt in Abbildung 5.1 links, eine exponentielle Abnahme erwartet, jedoch wird dies nicht von den Beobachtungen unterstützt, da der letzte Wert sogar größer ist als die beiden Werte zuvor. Natürlich könnte die wahre Konzentration sehr wohl einer exponentiellen Kurve folgen, und die Beobachtungen weichen davon nur wegen des Messfehlers ab. Falls dies der Fall ist, müsste der Logarithmus der Konzentrationen eine lineare Beziehung zur Zeit aufweisen (vgl. Abbildung 5.1 rechts).

Die einfachste Art zu testen, ob die erkennbare Krümmung mehr als zufällig ist, besteht in der Modellierung mit einem zusätzlichen quadratischen Zeitterm. Dazu nehmen wir vorerst an, dass die Anzahlen normalverteilt sind und einem linearen Modell mit Prädiktoren `time` und $\texttt{time}^2$ genügen.

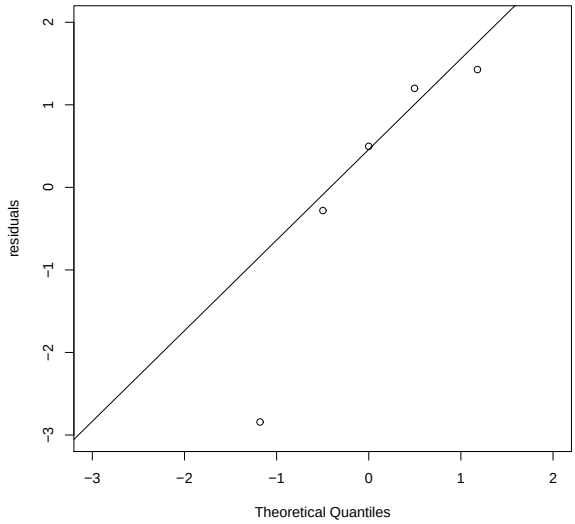

Abbildung 5.2: Normal Q-Q-Plot der Residuen aus dem linearen Modell.

```
> 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))
```

Der quadratische Term scheint im Modell notwendig (p-Wert 0.058). Zuvor sollte noch die getroffene Verteilungsannahme mittels Normal Q-Q-Plot der Residuen $r_i = y_i - \hat{\mu}_i$ geprüft werden. Der Punktverlauf in der Abbildung 5.2 weicht von einer Geraden ab, so dass die Annahme einer Normalverteilung nicht wirklich gerechtfertigt zu sein scheint.

Daher versuchen wir jetzt ein Poisson-Modell. Für gewöhnlich werden hierbei die Poisson-Erwartungen linear auf einer Log-Skala modelliert. Dies bedeutet, dass für den Fall einer exponentiellen Abnahme die Erwartungen auf der Log-Skala modelliert werden, während die Verteilung um diese Erwartungen als Poissonvariablen auf der Originalskala betrachtet werden. Der rechte Teil der Abbildung 5.1 zeigt noch immer eine beachtliche Krümmung,

weshalb auch hier wiederum der quadratische Zeitterm in das Modell aufgenommen wird.

```
> 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 *
---
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: 0.2793  on 2  degrees of freedom
AIC: 30.849

Number of Fisher Scoring iterations: 3

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

Die Deviance (0.2793) und die Pearson-Statistik X^2 (0.2745) sollten unter dem korrekten Modell etwa ihrem Freiheitsgrad $n-p=2$ entsprechen und stellen einen Test auf Güte der Anpassung dar. Kleine Werte zeugen von einer recht guten Anpassung, während große auf eine schlechte Anpassung hinweisen. Da hier beide Werte sehr klein sind, spricht nichts gegen die Poisson-Annahme ($\text{var}(y_i) = \mu_i$).

```
> 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"))
```

Auch der Residuenplot im linken Teil der Abbildung 5.3 zeigt dies. Falls die Varianz der Erwartung entspricht, ist ein Schätzer für die Varianz der Residuen unter dem wahren Modell gerade der geschätzte Erwartungswert. Daher sollten die Pearson-Residuen r_i etwa Erwartung Null und Varianz Eins aufweisen. Da dieser Residuenplot relativ ($n = 5$) unauffällig ist, erscheint die Poisson-Annahme auch 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 im rechten Teil der Abbildung 5.3 ziemlich gut an die 5 Beobachtungen anpassen.

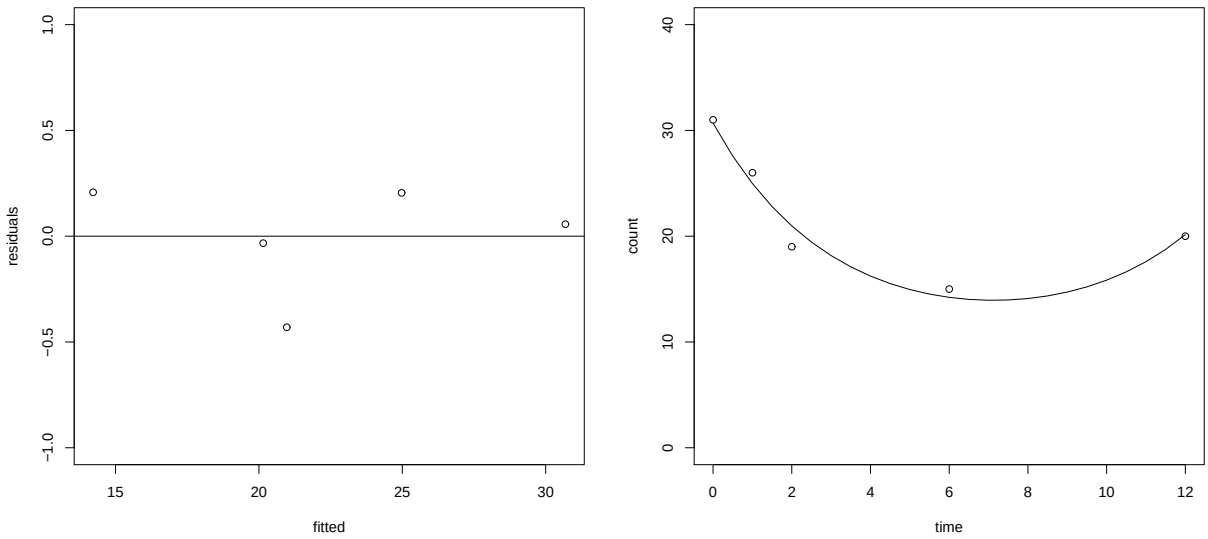

Abbildung 5.3: Residuenplot (links) und geschätztes Poissonmodell (rechts).

Obwohl Messfehler auch zu einem Zuwachs der Anzahl führen können, sollte dies aber in der Realität unmöglich sein. Wir testen daher die Notwendigkeit des quadratischen Zeitterms mittels eines Modellvergleichs.

```
> 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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Die Analysis of Deviance weist zwar auf die Notwendigkeit des quadratischen Effektes (p-Wert= 0.039) zusätzlich zum linearen Zeiteffekt hin, wir wissen jedoch, dass ein quadratisches Modell nicht viel Sinn macht. Möglicherweise erreicht man ein realistischeres und einfacheres Modell durch eine logarithmische Transformation der Zeitachse.

Falls die Zeit multiplikativ wirkt, sollte sich das Modell auf log(`time`) als Prädiktor beziehen. Jedoch ist dafür die Anfangszeit, log(0), problematisch. Wir betrachten daher die Transformation log(`time` + c) mit unbekanntem Shift $0 < c$. Eine Möglichkeit c zu bestimmen, ist die Deviance in Abhängigkeit von c zu minimieren. Das Ergebnis dieses Vorgehens ist in der Abbildung 5.4 dargestellt. Der optimale Wert liegt für das Modell 1+log(`time`+c) um $c = 0.105$ und log(`time`+0.105) wird ab jetzt als Prädiktor verwendet.

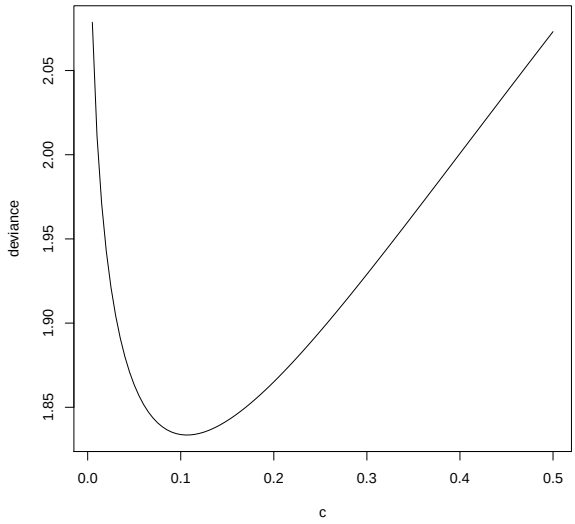

Abbildung 5.4: Abhängigkeit der Deviance von c.

```
> 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
> time.c <- time + 0.105
> summary(mo.P2 <- 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

Number of Fisher Scoring iterations: 4
```

Zunächst ist es wiederum ratsam ein Modell zu betrachten, das auch einen quadratischen Zeitterm beinhaltet, um damit zu prüfen, ob noch immer eine verbleibende Krümmung vorhanden ist.

```
> mo.P3 <- glm(count ~ log(time.c)+I(log(time.c)^2), family=poisson)
> anova(mo.P2, mo.P3, 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
```

Der quadratische Zeitterm ist nicht mehr signifikant. Wegen der geringen Deviancedifferenz von 0.041 verzichten wir auf den quadratischen Term im Modell. Es scheint, dass mit der transformierten Zeitachse ein linearer Trend im Prädiktor ausreicht.

Interessant ist es auch, approximative (zweiseitige) punktweise Konfidenzintervalle für die Modellkurve $\mu_0 = \exp(\eta_0)$ zu betrachten. Eine Möglichkeit besteht darin, $\hat{\eta}_0 = x_0^\top \hat{\beta}$ zusammen mit $\widehat{s.e.}(\hat{\eta}_0)$ dafür zu verwenden. Das transformierte 95% Intervall ist somit

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

was auch in der Abbildung 5.5 dargestellt ist.

Alternativ liefert die Delta Methode wegen der linearen Approximation

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

als approximative Varianz, bzw. Standardfehler

$$\begin{aligned} \text{var}(\log \hat{\mu}) &\approx \text{var}(\hat{\mu})\frac{1}{\mu^2} \\ \widehat{\text{var}}(\hat{\mu}) &\approx \hat{\mu}^2 \text{var}(\hat{\eta}) \\ \widehat{s.e.}(\hat{\mu}_0) &\approx \hat{\mu}_0 \widehat{s.e.}(\hat{\eta}_0)\,. \end{aligned}$$

Somit resultiert mit diesem Ansatz das symmetrische Intervall basierend auf einer Normalverteilungsannahme für $\hat{\mu}_0$

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

Natürlich können beide Ideen auch recht unterschiedliche Intervallsgrenzen liefern.

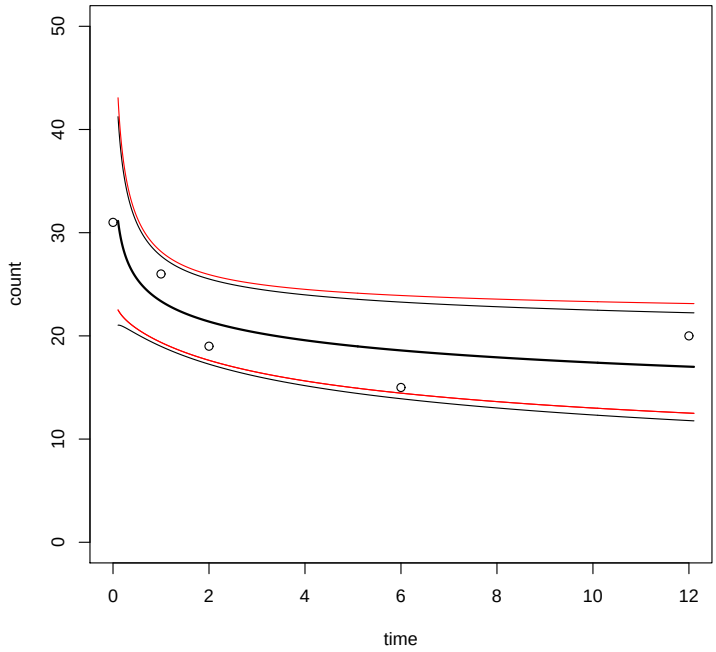

Abbildung 5.5: Punktweise 95% Konfidenzintervalle KIV (rot) und KIV_Δ (schwarz) unter dem Poissonmodell mit Prädiktor $\log(\texttt{time} + 0.105)$.

```
> # pointwise (1-alpha) confidence intervalls for \mu:
> # either by the Delta-Method
> time.c.new <- data.frame(time.c = seq(0,12,.005) + 0.105)
> r.pred <- predict(mo.P2, newdata=time.c.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", ylim=c(0,50), xlab="time", ylab="count")
> lines(time.c.new[,1], upper)
> lines(time.c.new[,1], fit)
> lines(time.c.new[,1], lower)

> # or using the prediction of type="link"
> l.pred <- predict(mo.P2, newdata=time.c.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)
```

5.2 Zweidimensionale Kontingenztafeln

Über die Modellierung der Erwartungswerte von Anzahlen hinaus, können loglineare Modelle auch verwendet werden, um Beziehungen zwischen Variablen zu beschreiben wie (stochastische) Unabhängigkeit, konditionale Unabhängigkeit oder eben Abhängigkeit. Im Speziellen wird hierbei keiner dieser Faktoren als Response definiert. Man spricht hierbei eher nur von den **Klassifikatoren**.

Beispiel: Lebensraum von Eidechsen Gezählt wurde, wieviele Eidechsen welchen Aufenthaltsort auf einer Sitzstange (engl. perch) gewählt haben. Diese Sitzstange ist charakterisiert durch jeweils zweistufige Faktoren, ihre Höhe (`height`, ≥ 4.75, < 4.75) und ihren Durchmesser (`diameter`, ≤ 4.0, > 4.0). Folgende Anzahlen konnten beobachtet werden:

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

Es stellt sich hierbei die Frage, ob die `diameter` und `height` Klassifikationen unabhängig sind. Die Assoziation kann mittels **odds-ratios** gemessen werden. Bei Unabhängigkeit wäre das odds-ratio gerade Eins (siehe nächster Abschnitt). Wir erhalten jedoch als Schätzer den Wert

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

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

Wir wollen nun ein loglineares Modell zur Modellierung von 2×2 Tabellen einführen und betrachten dazu ganz allgemein 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.

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

Wir sind jetzt an zwei Modellen interessiert:

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

5.2.1 Unabhängigkeitsmodell

Nehmen wir an, dass für alle beobachteten Paare (a_i, b_i), $i = 1, \ldots, n$, die Wahrscheinlichkeit in Zelle (k, l) zu fallen gleich π_{kl} ist. Somit gilt für den Erwartungswert der Gesamtanzahl von Beobachtungen y_{kl} in Zelle (k, l)

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

Falls stochastische Unabhängigkeit vorliegt, d.h. falls für alle Wahrscheinlichkeiten gilt

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

dann liefert dafür das loglineare Modell gerade

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

Der Logarithmus des Erwartungswertes für die Zelle (k, l) ist daher eine additive Funktion des k-ten Zeileneffekts und des l-ten Spaltenffekts. Dieses Modell ist somit äquivalent mit dem Modell

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B\,, \qquad k, l \in \{1, 2\}\,. \tag{5.1}$$

Zu überlegen verbleibt noch, wie diese Parameter genau zu definieren sind, und wieviele davon überhaupt identifizierbar sind. Wir definieren dazu beispielsweise (bei Verwendung der Kontrast Parametrisierung)

$$\begin{aligned}
\lambda_k^A &= \log \pi_k^A - \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^A \\
\lambda_l^B &= \log \pi_l^B - \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^B \\
\lambda &= \log n + \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^A + \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^B\,.
\end{aligned}$$

Mit dieser gewählten Parametrisierung (Abweichungen von den Mitteln) gelten somit als Bedingungen für die Identifizierbarkeit der Parameter

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

Daher ist außer λ nur noch ein Zeilen- und ein Spaltenparameter identifizierbar. Die beiden anderen Parameter unterscheiden sich nur im Vorzeichen, d.h. $\lambda_2^A = -\lambda_1^A$ und $\lambda_2^B = -\lambda_1^B$. Modell (5.1) wird loglineares **Unabhängigkeitsmodell** genannt. Wir erhalten unter dieser Parametrisierung folgende linearen Prädiktoren

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

Viel einfacher ist es jedoch, bei der Parametrisierung anstelle von Kontrasten mit einer Referenzzelle zu arbeiten. Man zeichnet hierbei eine beliebige Zelle als Referenz aus und betrachtet Parameter, welche die Abweichungen zu dieser beschreiben. Bildet beispielsweise die Zelle $(1,1)$ die Referenz, dann liefert diese Idee

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

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	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

Wir erinnern uns, dass der MLE einer Wahrscheinlichkeit (Binomialverteilung) gerade der entsprechenden relativen Häufigkeit entspricht. In unserem Fall von 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

$$\hat\mu_{kl} = n\hat\pi_{kl} = y_{\bullet\bullet}\frac{y_{k\bullet}}{y_{\bullet\bullet}}\frac{y_{\bullet l}}{y_{\bullet\bullet}} = \frac{1}{y_{\bullet\bullet}} y_{k\bullet}y_{\bullet l}\,.$$

Als MLE unserer Parameter liefert dies wiederum sofort

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

$$\log\hat\mu_{21} = \hat\lambda + \hat\lambda_2^A = \log\frac{y_{2\bullet}y_{\bullet 1}}{y_{\bullet\bullet}} \Rightarrow \hat\lambda_2^A = \log\frac{y_{2\bullet}y_{\bullet 1}}{y_{\bullet\bullet}} - \log\frac{y_{1\bullet}y_{\bullet 1}}{y_{\bullet\bullet}} = \log\frac{y_{2\bullet}}{y_{1\bullet}}$$

$$\log\hat\mu_{12} = \hat\lambda + \hat\lambda_2^B = \log\frac{y_{1\bullet}y_{\bullet 2}}{y_{\bullet\bullet}} \Rightarrow \hat\lambda_2^B = \log\frac{y_{1\bullet}y_{\bullet 2}}{y_{\bullet\bullet}} - \log\frac{y_{1\bullet}y_{\bullet 1}}{y_{\bullet\bullet}} = \log\frac{y_{\bullet 2}}{y_{\bullet 1}}$$

Bemerke, dass durch diese Schätzung auch die Summe der geschätzten Erwartungen die Summe der Beobachtungen reproduziert. Die folgt sofort aus

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

also

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

Man beachte auch, dass unter dieser Parametrisierung für das log-odds-ratio gilt

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

Dies hat zur Folge, dass in dieser Parametrisierung ein odds-ratio von $\psi = 1$ mit der Unabhängigkeit äquivalent ist. Dieses Ergebnis hält unabhängig von der Wahl der Referenzzelle.

5.2.2 Saturiertes (volles) Modell

Falls keine Unabhängigkeit angenommen werden kann, definieren wir das Modell

$$\log\mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}\,, \qquad k,l \in \{1,2\}\,. \tag{5.2}$$

In diesem Prädiktor beschreiben die sogenannten Interaktionsparameter λ_{kl}^{AB} gerade die Abweichungen vom Unabhängigkeitsmodell (5.1).

Will man mit Kontrasten arbeiten, dann ist es am einfachsten mit den linearen Prädiktoren $\eta_{kl} = \log\mu_{kl}$ die Parameter zu definieren. Seien dazu die mittleren (Zeilen-, Spalten-, overall-) Prädiktoren gleich

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

Wir definieren damit den Zeileneffekt λ_k^A, Spalteneffekt λ_l^B und Interaktionseffekt (Wechselwirkung) λ_{kl}^{AB} als entsprechende Abweichung vom mittleren Prädiktor, d.h.

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

Hierbei bezeichnen $\{\lambda_k^A\}$ und $\{\lambda_l^B\}$ eine Abweichung vom generellen Prädiktormittel λ. Die Wechselwirkung $\{\lambda_{kl}^{AB}\}$ ist der um den Zeilen- wie auch Spalteneffekt bereinigte Zelleneffekt. Da hier sämtliche Parameter mittelwertsbereinigte Effekte darstellen, folgt sofort als Identifizierbarkeitsbedingung

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

Deshalb gibt es wiederum nur einen unabhängigen Zeilenparameter bzw. einen unabhängigen Spaltenparameter. Falls $\lambda_k^A > 0$, ist die durchschnittliche (log) erwartete Häufigkeit für Zellen der k-ten Zeile größer als die durchschnittliche (log) erwartete Häufigkeit über die gesamte Tabelle.

Weiters gilt für die Interaktion

$$\begin{aligned}\sum_{k=1}^{2} \lambda_{kl}^{AB} &= \sum_{k=1}^{2} \eta_{kl} - \sum_{k=1}^{2} \eta_{k\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet}\\ &= 2\eta_{\bullet l} - 2\eta_{\bullet\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet} = 0 = \sum_{l=1}^{2} \lambda_{kl}^{AB}\,.\end{aligned}$$

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 einen freien Interaktionsparameter.

Das Unabhängigkeitsmodell (5.1) ist ein Spezialfall des vollen Modells (5.2) mit $\lambda_{kl}^{AB} = 0$ für alle (k, l). Die zusätzlichen Parameter λ_{kl}^{AB} nennt man auch **Assoziationsparameter**, welche die Abweichungen von der Unabhängigkeit zwischen A und B beschreiben. Die totale Anzahl freier Parameter ist 3 beim Unabhängigkeitsmodell, bzw. 4 beim Abhängigkeitsmodell.

In `R` arbeitet man standardmäßig mit der `treatment` Kontrastierung, d.h. mit der Referenzzelle $(1, 1)$. Möchte man mit der obigen Parametrisierung Modelle interpretieren, so ist dazu folgender Aufruf notwendig

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

Die Default-Einstellung von `R` ist jedoch

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

Wiederum ist es einfacher, mit einer 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 $\lambda_1^A = \lambda_1^B = 0$. Darüberhinaus sind sämtliche 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	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_1^B + \lambda_{22}^{AB}$	

Als MLE ergibt dies

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

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

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

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

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

5.2.3 Beispiel: Lebensraum von Eidechsen

Um auch in R die korrekte Zelle $(1,1)$ als Referenz zu erhalten, geben wir die Daten folgendermaßen ein und erhalten als MLEs:

```
> count <- c(61, 41, 73, 70)
> (hei <-   factor(c(">4.75", ">4.75", "<4.75", "<4.75"))) # auxiliary variable
[1] >4.75 >4.75 <4.75 <4.75
Levels: <4.75 >4.75
> (height <- relevel(hei, ref = ">4.75"))            # correct order of levels
[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"))

> 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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

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

Die Deviance ist Null bei Null Freiheitsgraden. Dieses Modell reproduziert exakt die Daten und das geschätzte odds-ratio ist wie bereits erwähnt

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

Unter dem Unabhängigkeitsmodell erhalten wir die Schätzer

```
> 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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10.9036  on 3  degrees of freedom
Residual deviance:  1.8477  on 1  degrees of freedom
AIC: 31.574
```

Das odds-ratio ist jetzt Null und die 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.

5.2.4 Mehrstufige Faktoren

Die Ergebnisse bei zweistufigen Klassifikatoren können unmittelbar auf den Fall mehrstufiger klassifizierender Faktoren verallgemeinert werden. Sei dafür nun A ein K-stufiger und B ein L-stufiger Faktor. Wir betrachten damit das **Unabhängigkeitsmodell**

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

Wollen wir wieder die Zelle $(1,1)$ als Referenz verwenden, so definieren wir entsprechend

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

und es gelten dieselben Identifizierbarkeitsbedingungen wie schon zuvor, d.h.

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

Somit sind im Unabhängigkeitsmodell gerade $1+(K-1)+(L-1)$ Parameter frei schätzbar und wir erhalten 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	$\lambda+\lambda_K^A$	$\lambda+\lambda_K^A+\lambda_2^B$	$\cdots$	$\lambda+\lambda_K^A+\lambda_l^B$	$\cdots$	$\lambda+\lambda_K^A+\lambda_L^B$

Die MLE sind jetzt für $k = 1,\ldots,K$ und $l = 1,\ldots,L$

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

$$\log \hat\mu_{k1} = \hat\lambda + \hat\lambda_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} \Rightarrow \hat\lambda_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{k\bullet}}{y_{1\bullet}}$$

$$\log \hat\mu_{1l} = \hat\lambda + \hat\lambda_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} \Rightarrow \hat\lambda_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{\bullet l}}{y_{\bullet 1}}$$

Das **saturierte Modell** für eine $K \times L$ Tabelle ist nun gegeben als

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

Mit der Referenzzelle $(1,1)$ erhalten wir für alle $k = 1, \ldots, K$ und $l = 1, \ldots, L$

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

wofür $\lambda_1^A = \lambda_1^B = 0$ gilt. Wiederum sind alle Interaktionen in der ersten Zeile und in der ersten Spalte Null. Dies hat zur Folge, dass im saturierten Modell genau $1 + (K-1) + (L-1) + (K-1)(L-1) = K \times L$ Parameter frei schätzbar sind und wir erhalten damit die folgenden 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	$\lambda + \lambda_K^A$	$\lambda + \lambda_K^A + \lambda_2^B + \lambda_{K2}^{AB}$	$\cdots$	$\lambda + \lambda_K^A + \lambda_l^B + \lambda_{Kl}^{AB}$	$\cdots$	$\lambda + \lambda_K^A + \lambda_L^B + \lambda_{KL}^{AB}$

Das saturierte Modell hat somit gerade um $(K-1)(L-1)$ freie Parameter mehr als das Unabhängigkeitsmodell.

Beispiel: Zervixkarzinom Wir wollen nun untersuchen, ob bei den Daten zum Zervixkarinom aus Kapitel 4 stochastische Unabhängigkeit zwischen den beiden Prädiktoren Grenzzonenbefall (GZ) und Anzahl befallener Lymphknotenstationen (LK) besteht. 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.

```
> 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
```

Das saturierte loglineare Modell weist darauf hin, dass die sechs Interaktionsparameter nicht Null sind und somit die Unabhängigkeitshypothese verworfen werden kann.
Eine alternative Vorgehensweise ist die Betrachtung der Pearson-Statistik unter dem Unabhängigkeitsmodell, d.h.

$$X^2 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

mit linearem Prädiktor $\log \mu_{ij} = \lambda + \lambda_i^G + \lambda_j^K$. Diese Statistik realisiert in

```
> ind <- glm(total ~ G+L, family=poisson)
> r <- residuals(ind, type="pearson")
> sum(r^2)
[1] 43.83645
```

was gerade der χ^2-Teststatistik bei der Kontingenztafelanalyse entspricht. Am einfachsten erhält man diese Statistik durch Aufruf von

```
> (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- und binomialverteilte Responsevariablen scheinen also im Falle einer $K \times 2$ Tabelle einige Gemeinsamkeiten aufzuweisen. Bezeichnen wir nun die Tabelleneintragungen mit N_{kl}, $k = 1, \ldots, K$ und $l \in \{1, 2\}$ (k bezeichnet beispielsweise die unterschiedlichen Experimentierumgebungen und l die beiden Responsekategorien “Erfolg/Misserfolg”). Für unabhängige Anzahlen $N_{kl} \overset{ind}{\sim} \text{Poisson}(\mu_{kl})$ liefert das Konditionieren auf die Gesamtanzahl der beiden Responses

$$N_{k\bullet} = N_{k1} + N_{k2}\,, \qquad k = 1, \ldots, K\,,$$

als konditionale Verteilung für jede Experimentierumgebung

$$N_{k1}|N_{k\bullet} \sim \text{Binomial}(N_{k\bullet}, \pi_k)\,, \qquad k = 1, \ldots, K\,,$$

mit

$$\pi_k = \frac{\mu_{k1}}{\mu_{k1} + \mu_{k2}} .$$

Dieser Aspekt wird im Abschnitt 5.4 noch im Detail besprochen werden. Wir diskutieren jedoch dieses Verhalten trotzdem bereits hier für den Fall einer $K \times 2$ 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 $n_{k\bullet} = n_{k1} + n_{k2}$. Man bemerke, dass die Anzahl N_{k1} hierbei die Anzahl der Erfolge beschreibt und dass $N_{k\bullet}$ die Gesamtanzahl durchgeführter Versuche im k-ten Umfeld zählt. Somit ergibt sich bei gegebener (fixierter) Gesamtanzahl an Versuchen

$$\begin{aligned}
\Pr(N_{k1} = n_{k1} | N_{k\bullet} = n_{k\bullet}) &= \frac{\Pr(N_{k1} = n_{k1}, N_{k2} = n_{k2})}{\Pr(N_{k\bullet} = n_{k\bullet})} \\
&= \frac{\frac{\mu_{k1}^{n_{k1}}}{n_{k1}!} e^{-\mu_{k1}} \frac{\mu_{k2}^{n_{k2}}}{n_{k2}!} e^{-\mu_{k2}}}{\frac{(\mu_{k1} + \mu_{k2})^{n_{k1}+n_{k2}}}{(n_{k1} + n_{k2})!} e^{-(\mu_{k1} + \mu_{k2})}} \\
&= \frac{(n_{k1} + n_{k2})!}{n_{k1}! n_{k2}!} \frac{\mu_{k1}^{n_{k1}}}{(\mu_{k1} + \mu_{k2})^{n_{k1}}} \frac{\mu_{k2}^{n_{k2}}}{(\mu_{k1} + \mu_{k2})^{n_{k2}}} \\
&= \binom{n_{k1} + n_{k2}}{n_{k1}} \left(\frac{\mu_{k1}}{\mu_{k1} + \mu_{k2}} \right)^{n_{k1}} \left(\frac{\mu_{k2}}{\mu_{k1} + \mu_{k2}} \right)^{n_{k2}} \\
&= \binom{n_{k\bullet}}{n_{k1}} \pi_k^{n_{k1}} (1 - \pi_k)^{n_{k2}} .
\end{aligned}$$

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.

Beispiel: Rezidivbildung beim Zervixkarzinom Wir untersuchen, ob der Lymphknotenbefall einen Einfluss auf die 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 Modells 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)
> 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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  3.0017e+02  on 7  degrees of freedom
Residual deviance: -4.4409e-15  on 0  degrees of freedom
AIC: 55.771

> 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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 419.46  on 7  degrees of freedom
Residual deviance: 337.34  on 4  degrees of freedom
AIC: 345.34
```

Die Spezifikation `R*L+L` für den Prädiktor beim loglinearen Modell besteht aus zwei Teilen. Der erste Teil `R*L` bezieht sich auf die Interaktion zwischen dem binomialen `R` mit dem

Prädiktor unter dem logistischen Ansatz (L). Der zweite Teil (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.

5.3 Dreidimensionale Kontingenztafeln

Die folgende Datensituation motiviert weitere Überlegungen bezüglich möglicher Unabhängigkeiten.

Beispiel: Diabetesstudie Es liegt eine Zufallsstichprobe von Diabetespatienten vor, die bezüglich

- Familien-Vorgeschichte betreffs Diabetes (vorhanden/nicht-vorhanden)
- Insulinabhängigkeit (ja/nein)
- Alter zum Zeitpunkt des Ausbruchs (< 45 / ≥ 45)

klassifiziert sind. 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

Wir sind an der Abhängigkeitsstruktur dieser drei Klassifikatoren interessiert und betrachten dazu verschiedene loglineare Modelle. Im Folgenden sind die 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")
> 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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.2516e+02  on 7  degrees of freedom
Residual deviance: 1.3323e-15  on 0  degrees of freedom
AIC: 47.626
```

Wir suchen nun dafür ein passendes Modell und betrachten dazu die Werte der 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

Das saturierte Modell `A*I*H` hat natürlich eine Deviance von Null bei keinen Freiheitsgraden. Die Dreifachinteraktion scheint darin aber irrelevant zu sein (Verschlechterung der Deviance um nur 0.04), was zum Modell `A*I+I*H+A*H` führt. Weiteres Weglassen der Interaktion `A:I` verschlechtert das Modell jedoch signifikant (Deviance Differenz von 49.08). Jedoch kann man darin auf die Interaktion `I:H` verzichten (Deviance Differenz von nur 0.01), was `A*I+A*H` ergibt. Jetzt kann man nur noch auf die Interaktion `A:H` verzichten (Deviance Differenz von 1.90). Weitere Vereinfachungen würden dieses Modell `A*I+H` signifikant verschlechtern.

Die einzige im Modell verbliebene Interaktion ist `age:insulin`. Somit 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 daher die Tabelle über `history` zusammenlegen und erhalten folgende Häufigkeiten

	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. 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 Familien-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.

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

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}\,.$$

Das log-odds-ratio zwischen A und B gegeben $C = 1$ ergibt sich unter diesem Modell als

$$\begin{aligned} \log \frac{\mu_{11(1)}\mu_{22(1)}}{\mu_{12(1)}\mu_{21(1)}} &= \big(\lambda\big) + \big(\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}\big) - \big(\lambda + \lambda_2^B\big) - \big(\lambda + \lambda_2^A\big) \\ &= \lambda_{22}^{AB}\,, \end{aligned}$$

weil in dieser Parametrisierung sämtliche Parameter Null sind, falls $i = 1$, $j = 1$ oder $k = 1$ ist. Konditionieren wir auf die zweite Stufe von C, so resultiert dafür

$$\begin{aligned} \log \frac{\mu_{11(2)}\mu_{22(2)}}{\mu_{12(2)}\mu_{21(2)}} &= \big(\lambda + \lambda_2^C\big) + \big(\lambda + \lambda_2^A + \lambda_2^B + \lambda_2^C + \lambda_{22}^{AB} + \lambda_{22}^{BC} + \lambda_{22}^{AC} + \lambda_{222}^{ABC}\big) \\ &\quad -\big(\lambda + \lambda_2^B + \lambda_2^C + \lambda_{22}^{BC}\big) - \big(\lambda + \lambda_2^A + \lambda_2^C + \lambda_{22}^{AC}\big) \\ &= \lambda_{22}^{AB} + \lambda_{222}^{ABC}\,. \end{aligned}$$

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
```

Für das schließlich gefundene adäquate Modell erhalten wir die 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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    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
```

Wie können derartige Modelle im Allgemeinen interpretiert werden? Angenommen, es liegen drei klassifizierende Faktoren A, B und C vor, dann gibt es die folgenden Klassen von Modellen:

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

$$(A \perp C)|B$$

   (a) die geschätzten odds-ratios für A, B sind gleich auf sämtlichen Stufen von C
   (b) die geschätzten odds-ratios für B, C sind gleich auf sämtlichen Stufen von A
   (c) 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).

5.4 Loglineare Multinomiale Response Modelle

Wir konzentrieren uns jetzt auf den Zusammenhang zwischen einem loglinearen Modell für Anzahlen und einem multinomialen (binomialen) Response Modell für Häufigkeiten. Diese Beziehung rührt von der Tatsache her, dass die Multinomialverteilung aus einer Serie unabhängiger Poissonvariablen generiert werden kann, wenn deren Summe fixiert wird.

5.4.1 Die Multinomialverteilung

Die Objekte in einer Population weisen eines von K Merkmalen auf, beispielsweise die Haarfarbe oder eine Todesursache. Aus dieser Population zieht man eine Zufallsstichprobe und ist an der Anzahl Y_k aller Beobachtungen mit Ausprägung $k = 1, \ldots, K$ interessiert. Nehmen wir an, dass es sich bei den Anzahlen um unabhängige Poissonvariablen handelt mit Erwartungswerten $\mu_1, \ldots, \mu_K$, also

$$\Pr(Y_k = y_k) = \frac{\mu_k^{y_k}}{y_k!} \exp(-\mu_k)\,, \qquad k = 1, \ldots, K\,.$$

Die Summe dieser Anzahlen, $Y_\bullet = \sum_{k=1}^K Y_k$, ist naturgemäß wiederum Poissonverteilt mit Erwartung $\mu_\bullet = \sum_{k=1}^K \mu_k$. Als bedingte Wahrscheinlichkeit der $(Y_1, \ldots, Y_K)$ gegeben deren Summe $Y_\bullet$ resultiert

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_K = y_K | Y_\bullet) &= \frac{\prod_{k=1}^K \frac{\mu_k^{y_k}}{y_k!} \exp(-\mu_k)}{\frac{\mu_\bullet^{y_\bullet}}{y_\bullet!} \exp(-\mu_\bullet)} = \frac{y_\bullet! \prod_{k=1}^K \mu_k^{y_k}}{\mu_\bullet^{y_\bullet} \prod_{k=1}^K y_k!} \\
&= \frac{y_\bullet!}{\prod_{k=1}^K y_k!} \prod_{k=1}^K \left(\frac{\mu_k}{\mu_\bullet}\right)^{y_k} = \frac{y_\bullet!}{\prod_{k=1}^K y_k!} \prod_{k=1}^K \pi_k^{y_k}\,. \qquad (5.3)
\end{aligned}$$

Somit entspricht dies einer Multinomialverteilung auf K Zellen mit Wahrscheinlichkeiten beschrieben durch die relativen 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. Natürlich ist durch diese Konstruktion auch sichergestellt, dass $\sum_k \pi_k = 1$ gilt.

Gegeben sei nun eine Stichprobe $y_1, \ldots, y_n$ von n nicht identisch verteilten, jedoch unabhängigen Multinomialvektoren $y_i = (y_{i1}, \ldots, y_{iK})$, und sei die Zeilensumme $y_{i\bullet} = \sum_k y_{ik}$ fest für jedes i (dies beschreibt gerade den "Stichprobenumfang" für den i-ten Multinomialvektor). Zu jedem y_i korrespondiert ein Vektor $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ für $i = 1, \ldots, n$.

Ein 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}	$\cdots$	y_{1k}	$\cdots$	y_{1K}	$y_{1\bullet}$
2	x_2	y_{21}	$\cdots$	y_{2k}	$\cdots$	y_{2K}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_{i1}	$\cdots$	y_{ik}	$\cdots$	y_{iK}	$y_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_{n1}	$\cdots$	y_{nk}	$\cdots$	y_{nK}	$y_{n\bullet}$

Der für die Schätzung von π_i relevante Beitrag der i-ten Beobachtung an der Log-Likelihood der Stichproben ist laut (5.3) (nach Weglassen aller von Parametern unabhängigen Termen) gerade

$$\ell(\pi_i|y_i) = \sum_{k=1}^{K} y_{ik} \log \pi_{ik} \,, \qquad \text{mit} \quad \sum_{k=1}^{K} \pi_{ik} = 1 \,.$$

Wegen der Unabhängigkeit der n Vektoren y_i resultiert als Log-Likelihood der Stichprobe

$$\ell(\pi|y) = \sum_{i=1}^{n} \ell(\pi_i|y_i) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log \pi_{ik} \,.$$

Für die Maximierung von $\ell(\pi|y)$ folgt unter den n Bedingungen $\sum_k \pi_{ik} = 1$, $i = 1, \ldots, n$, als Scorefunktion

$$\frac{\partial}{\partial \pi_{ik}} \Big(\ell(\pi|y) - \sum_{i=1}^{n} \lambda_i \Big(\sum_{k=1}^{K} \pi_{ik} - 1 \Big) \Big) = \frac{y_{ik}}{\pi_{ik}} - \lambda_i \,.$$

Nullsetzen liefert also $\hat{\pi}_{ik} = y_{ik}/\hat{\lambda}_i$. Um die Werte $\hat{\lambda}_i$ zu bestimmen summiert man die geschätzten Wahrscheinlichkeiten für ein i über alle K Kategorien auf. Dies führt zu $1 = \sum_k \hat{\pi}_{ik} = \sum_k y_{ik}/\hat{\lambda}_i = y_{i\bullet}/\hat{\lambda}_i$, also $\hat{\lambda}_i = y_{i\bullet}$. Die Scorefunktion ist somit

$$\frac{\partial \ell(\pi|y)}{\partial \pi_{ik}} = \frac{y_{ik} - y_{i\bullet}\pi_{ik}}{\pi_{ik}} \,.$$

In logistischen Modellen hatte man für binomialverteilte Responses (mit gerade 2 Kategorien) den Logarithmus des einen odds-ratios betrachtet und diesen linear modelliert, also

$$\text{logit}(\pi_i) = \log \frac{\pi_{i1}}{1 - \pi_{i1}} = \log \frac{\pi_{i1}}{\pi_{i2}} = \eta_i = x_i^\top \beta \,.$$

Bei den multinomialen Logitmodellen stehen zum Vergleich K Kategorien zur Verfügung. Für gewöhnlich wird jedoch die erste Kategorie als Referenz verwendet. Dies führt zum Modell der Form

$$\log \frac{\pi_{ik}}{\pi_{i1}} = \log \frac{y_{i\bullet}\pi_{ik}}{y_{i\bullet}\pi_{i1}} = \log \frac{\mu_{ik}}{\mu_{i1}} = \eta_{ik} = x_i^\top \beta_k \,, \qquad k = 2, \ldots, K$$

mit $\eta_{i1} = 0$ für $i = 1, \ldots, n$. Man bemerke, dass hier der Parametervektor β_k kategorienspezifisch ist um den Effekt von x_i auf die Wahrscheinlichkeit π_{ik} beschreiben zu können. Die Verwendung der ersten Kategorie ist keinerlei Restriktion, da sofort

$$\log \frac{\pi_{ik}}{\pi_{i1}} - \log \frac{\pi_{ij}}{\pi_{i1}} = \log \frac{\pi_{ik}}{\pi_{ij}} = \eta_{ik} - \eta_{ij}\,, \qquad k = 2, \ldots, K$$

folgt.
Unter diesem Modell ergibt die Restriktion $\sum_k \pi_{ik} = 1$ sofort die entsprechenden Ausdrücke für die Wahrscheinlichkeiten, d.h.

$$\begin{aligned} \log \frac{\pi_{ik}}{\pi_{i1}} &= \eta_{ik} \\ \pi_{ik} &= \pi_{i1} e^{\eta_{ik}} \\ 1 = \sum_{k=1}^{K} \pi_{ik} &= \pi_{i1} \sum_{k=1}^{K} e^{\eta_{ik}} \end{aligned} \tag{5.4}$$

liefert

$$\pi_{i1} = \frac{1}{\sum_{k=1}^{K} e^{\eta_{ik}}}$$

und (5.4) ergibt damit weiters

$$\pi_{ik} = \frac{e^{\eta_{ik}}}{\sum_{k'=1}^{K} e^{\eta_{ik'}}}\,, \qquad k = 2, \ldots, K\,.$$

Für $K = 2$ liegen binomiale Daten vor und das multinomiale Modell reduziert sich auf ein binomiales GLM mit logit-Linkfunktion. Falls jedoch $K > 2$ gilt, dann passt das multinomiale Logitmodel nicht mehr in den GLM Rahmen. Es wird sich aber herausstellen, dass wir die Parameter im Modell mittels eines Poisson GLM mit log-Link schätzen können. Der Schlüssel liegt in der Spezifikation des linearen Prädiktors.

5.4.2 Vergleich von Poisson-Erwartungen

Um herauszufinden, wie derartige multinomiale Modelle einfach geschätzt werden können, betrachten wir vorerst unabhängige $Y_k \overset{ind}{\sim} \text{Poisson}(\mu_k)$ Variablen mit

$$\log \mu_k = \phi + x_k \beta\,, \qquad k = 1, \ldots, K,$$

wobei x_k gegebene Konstanten sind und β einen unbekannten Parameter (-vektor ohne Intercept) bezeichnet. Der relevante Teil der Poisson-Log-Likelihood Funktion

$$\ell(\mu|y) = \sum_{k=1}^{K} \Big(y_k \log \mu_k - \mu_k \Big)$$

entspricht hier

$$\ell(\phi, \beta|y) = \sum_{k=1}^{K} \Big(y_k(\phi + x_k\beta) - \exp(\phi + x_k\beta) \Big) .$$

Um zu sehen, wie dies zu einem multinomialen Modell transformiert, verwenden wir

$$\mu_\bullet = \sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} \exp(\phi + x_k\beta) = \exp(\phi) \sum_{k=1}^{K} \exp(x_k\beta) ,$$

also

$$\log \mu_\bullet = \phi + \log \Big(\sum_{k=1}^{K} \exp(x_k\beta) \Big) , \qquad \text{bzw.} \quad \phi = \log \mu_\bullet - \log \Big(\sum_{k=1}^{K} \exp(x_k\beta) \Big) .$$

Bezüglich $(\mu_\bullet, \beta)$ erhält man

$$\begin{aligned} \ell(\mu_\bullet, \beta|y) &= \sum_{k=1}^{K} \left\{ y_k \Big(\log \mu_\bullet - \log \Big(\sum_{k'=1}^{K} \exp(x_{k'}\beta) \Big) + x_k\beta \Big) \right\} - \mu_\bullet \\ &= \log \mu_\bullet \sum_{k=1}^{K} y_k - \sum_{k=1}^{K} y_k \log \Big(\sum_{k'=1}^{K} \exp(x_{k'}\beta) \Big) + \sum_{k=1}^{K} y_k x_k \beta - \mu_\bullet , \end{aligned}$$

was mittels $y_\bullet = \sum_k y_k$ geschrieben werden kann als

$$\begin{aligned} \ell(\mu_\bullet, \beta|y) &= \Big\{ y_\bullet \log \mu_\bullet - \mu_\bullet \Big\} + \Big\{ \sum_{k=1}^{K} y_k x_k \beta - y_\bullet \log \Big(\sum_{k=1}^{K} \exp(x_k\beta) \Big) \Big\} \\ &= \ell(\mu_\bullet|y_\bullet) + \ell(\beta, y|y_\bullet) . \end{aligned}$$

Die Log-Likelihood zerfällt in zwei Komponenten. Die Komponente $\ell(\mu_\bullet|y_\bullet)$ entspricht der Log-Likelihood zu $y_\bullet \sim \text{Poisson}(\mu_\bullet)$. Um die zweite Komponente zu untersuchen, verwenden wir die Parametrisierung

$$\pi_k = \frac{\mu_k}{\mu_\bullet} = \frac{\exp(\phi + x_k\beta)}{\sum_{k'} \exp(\phi + x_{k'}\beta)} = \frac{\exp(x_k\beta)}{\sum_{k'} \exp(x_{k'}\beta)} .$$

Wegen

$$\log \pi_k = x_k\beta - \log \Big(\sum_{k'} \exp(x_{k'}\beta) \Big) ,$$

ist der zweite Term in der obigen Log-Likelihood gleich

$$\ell(\beta, y|y_\bullet) = \sum_{k=1}^{K} y_k \log \pi_k ,$$

was einer multinomialen Log-Likelihood entspricht, also der konditionalen Poisson-Log-Likelihood von y gegeben die Gesamtanzahl $y_\bullet$.

Bemerkung 5.1. Die wichtigste Erkenntnis hierbei ist, dass der erste Term $\ell(\mu_\bullet|y_\bullet)$ nicht von β abhängt und der zweite Term $\ell(\beta, y|y_\bullet)$ nicht von $\mu_\bullet$. Somit ist die gesamte Information über den uns interessierenden Parameter β im zweiten Ausdruck enthalten. Dies hat zur Folge, dass der MLE von β und dessen asymptotische Varianz dieselben sind ungeachtet, ob wir dafür die volle Log-Likelihood $\ell(\mu_\bullet, \beta|y)$ (die zum loglinearen Poissonmodell gehört) nehmen, oder nur die Log-Likelihood $\ell(\beta, y|y_\bullet)$ für das multinomiale Logitmodell. Wir müssen einzig in passender Weise einen Nuisance Parameter $\mu_\bullet$ in das Modell aufnehmen.

5.4.3 Multinomiale Responsemodelle

Mit diesen Ergebnissen kann man zeigen, dass bestimmte loglineare Modelle äquivalent zu multinomialen Modellen sind. Dazu ordnet man wie bereits in Abschnitt 5.4.1 die Beobachtungen y_{ik}, mit Kovariablensituation $i = 1, \ldots, n$ und Kategorie $k = 1, \ldots, K$, den Zellen einer zweidimensionalen Kontingenztafel zu und betrachtet das Modell

$$\log \mu_{ik} = \phi_i + x_i^\top \beta_k \,, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K \tag{5.5}$$

mit $\mu_{ik} = \mathrm{E}(y_{ik})$ und erklärenden Prädiktorvektoren x_i (mit $p-1$ Elementen). Man bemerke, dass in diesen Vektoren kein Intercept enthalten ist und dass sie beobachtungsspezifisch und nicht kategoriespezifisch sind. Natürlich interessieren uns ausschließlich die Parameter β_k während die ϕ_i die Rolle von Nebenparameter ("nuisance parameters") innehaben. In diesem Modell wächst die Dimension des Parameterraums, diese ist $n+(p-1)$, mit $n \to \infty$ für festes p. Daher ist auch für $n \to \infty$ nicht garantiert, dass der Maximum-Likelihood Schätzer effizient oder konsistent ist.

Es wird nun gezeigt, dass die konditionale Likelihood nur von $\beta = (\beta_1, \ldots, \beta_K)^\top$ und nicht von $\phi = (\phi_1, \ldots, \phi_n)^\top$ abhängt. Analog wie zuvor ist für Poissonverteilte Responses unter Modell (5.5) die Log-Likelihood der Stichprobe gegeben durch

$$\begin{aligned} \ell(\phi, \beta|y) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \Big(y_{ik}(\phi_i + x_i^\top \beta_k) - \exp(\phi_i + x_i^\top \beta_k) \Big) \\ &= \sum_{i=1}^{n} \phi_i y_{i\bullet} + \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} x_i^\top \beta_k - \sum_{i=1}^{n} \sum_{k=1}^{K} \exp(\phi_i + x_i^\top \beta_k) \,. \end{aligned}$$

Wir fixieren die i-te Responsesumme $y_{i\bullet}$ und betrachten die Transformation

$$\mu_{i\bullet} = \sum_{k=1}^{K} \mu_{ik} = \exp(\phi_i) \sum_{k=1}^{K} \exp(x_i^\top \beta_k) \,, \qquad \text{also} \quad \phi_i = \log \mu_{i\bullet} - \log \Big(\sum_{k=1}^{K} \exp(x_i^\top \beta_k) \Big) \,.$$

Die Log-Likelihood als Funktion in $(\mu_\bullet, \beta)$ ist somit

$$\begin{aligned} \ell(\mu_\bullet, \beta|y) &= \sum_{i=1}^{n} \Big\{ y_{i\bullet} \log \mu_{i\bullet} - \mu_{i\bullet} \Big\} + \sum_{i=1}^{n} \Big\{ \sum_{k=1}^{K} y_{ik} x_i^\top \beta_k - y_{i\bullet} \log \Big(\sum_{k'=1}^{K} \exp(x_i^\top \beta_{k'}) \Big) \Big\} \\ &= \ell(\mu_\bullet|y_\bullet) + \ell(\beta, y|y_\bullet) \,. \end{aligned}$$

Der erste Ausdruck ist wieder die Log-Likelihood zu den n Poissonverteilten Summen $y_{i\bullet} \sim \text{Poisson}(\mu_{i\bullet})$. Der zweite Ausdruck ist die konditionale Log-Likelihood zu den $y_i|y_{i\bullet}$ und hängt nur von β und nicht von ϕ ab. Die ganze Information über β ist im zweiten Ausdruck. Daher ist klar, dass $\hat{\beta}$ und $\text{var}(\hat{\beta})$ basierend auf die Maximierung von $\ell(\beta, y|y_\bullet)$ identisch den entsprechenden Resultaten unter Verwendung der vollen Log-Likelihood ist. Somit ist das loglineare Modell (5.5) äquivalent mit dem multinomialen Modell

$$\pi_{ik} = \frac{\mu_{ik}}{\mu_{i\bullet}} = \frac{\exp(x_i^\top \beta_k)}{\sum_{k'=1}^{K} \exp(x_i^\top \beta_{k'})}\,. \tag{5.6}$$

Beispiel: Rezidiv beim Zervixkarzinom Es werden wiederum die Häufigkeiten von Rezidiven analysiert. Hatten wir dafür in Abschnitt 4.3.1 unabhängige binomialverteilte Responses mit dem Logit-Link verwendet, so werden nun Poissonvariablen mit dem Log-Link das Modell bilden. Wir betrachten somit eine Tabelle mit $n = 24$ Zeilen (hier die Zellen), wobei in jeder Zelle die dazugehörende binomiale Information steht.

	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

Um die gleichen Ergebnisse wie beim logistischen Modell mit Prädiktor $\texttt{L} + \texttt{G}$ zu erhalten, muss die Summe der beiden Poisson-Beobachtungen in den von `L*G` aufgespannten Zellen (in jeder Zelle befindet sich jeweils eine $\texttt{R} = 0$ und eine $\texttt{R} = 1$ Häufigkeit) fixiert werden. Das erreicht man durch Aufnahme der Interaktion `L*G` in den Prädiktor (ergibt die Nebenparameter ϕ). Das eigentliche Modell für den Erwartungswert geht als Interaktion mit `R` in den Prädiktor ein (rezidivspezifische Parameter), also betrachtet man

$$\log \mu = \texttt{R} * (\texttt{L} + \texttt{G}) + \texttt{L} * \texttt{G}\,.$$

```
> 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

> 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
L2:G1       -0.360864   0.502329  -0.718 0.472522
L3:G1       -0.705828   0.592756  -1.191 0.233749
L1:G2        1.766372   0.441979   3.997 6.43e-05 ***
L2:G2        1.456786   0.503794   2.892 0.003833 **
L3:G2        1.375382   0.546435   2.517 0.011836 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 434.130  on 23  degrees of freedom
Residual deviance:  10.798  on  6  degrees of freedom
AIC: 134.75
```

Der Term `L*G` vergibt jeder Zelle einen Parameter (ϕ), während die erklärenden Faktoren des Logit-Modells in den Wechselwirkungen mit `R` (hier `L+G`) zu finden sind. Das Modell beinhaltet daher $p = 18$ Parameter.

Man 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. Natürlich haben sich Deviance und Freiheitsgrade auch nicht geändert.

```
> coefficients(rez.LG)
(Intercept)          L1          L2          L3          G1          G2
  1.6041433  -1.2872803  -1.7785486  -3.7978510  -0.7285015  -1.0735344
> deviance(rez.LG)
[1] 10.79795
```

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 $\sum_{r=1}^{2} y_{ir}$. Das Modell $\text{logit}(\mu) = 1$ entspricht beispielsweise der Annahme, dass die Rezidivwahrscheinlichkeit in jeder Zelle gleich ist (Zufallsstichprobe multinomialverteilter Anzahlen). Spezifiziert wird es als

$$\log \mu = \mathtt{R} + \mathtt{L} * \mathtt{G}\,.$$

Man erhält ein Modell mit $p = 13$ Parametern. Wiederum fixiert `L*G` nur die Ränder und `R` $\equiv$ `R*1` steht nur in "Wechselwirkung" mit dem Intercept. Unter diesem Modell ist die 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
```

Eine sehr einfache Alternative zur Verwendung von `glm` und der Spezifikation eines loglinearen Poissonmodell mit allen Interaktionen stellt die Funktion `multinom` in der Bibliothek `nnet` dar. Hiermit können multinomiale Daten direkt modelliert werden.

```
> library(nnet)
> MN.direct <- multinom(R ~ L+G, weights=count)
# weights:  7 (6 variable)
initial  value 216.955068
iter  10 value 163.393129
final  value 163.380752
converged
> summary(MN.direct)
Call:
multinom(formula = R ~ L + G, weights = count)
```

```
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
```

Hierbei vergleicht die Deviance das betrachtete Modell mit einem Modell, das jede der 313 einzelnen Beobachtungen korrekt vorhersagt, und nicht mit dem saturierten Modell zu den 12 Eintragungen der 3×4 Tabelle. Dieser Vergleich resultiert aus dem Aufruf

```
> summary(MN.direct.sat <- multinom(R ~ L*G, weights=count))

Coefficients:
               Values Std. Err.
(Intercept)   1.59020    0.2394
L1           -0.89700    0.5212
L2           -1.84149    0.5579
L3          -14.26143  156.5144
G1           -0.79166    0.3713
G2           -0.67376    0.6382
L1:G1         0.09846    0.8280
L2:G1         0.12664    1.0449
L3:G1        -0.61093  532.3983
L1:G2        -1.69349    1.0087
L2:G2        -0.17364    1.0516
L3:G2        11.73544  156.5174

Residual Deviance: 316
AIC: 340

> anova(MN.direct, MN.direct.sat) # resulting deviance difference
  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
```

Kapitel 6

Modelle mit zufälligen Effekten

6.1 Zufällige Prädiktoren

Bis jetzt basierten die betrachteten GLMs auf lineare Prädiktoren der Form

$$\eta_i = g(\mu_i) = x_i^\top \beta \,.$$

Die erklärenden Größen zur i-ten Responsevariablen sind hierbei zum Vektor x_i zusammengefasst und β stellt den unbekannten Parametervektor dar, der geschätzt werden muss. Wir werden jetzt eine Klasse von Modellen diskutieren, die im linearen Prädiktor zusätzlich zu diesen festen Effekten noch einen zufälligen Effekt verwendet. Dies hat in den folgenden beiden Situationen eine sehr klare Motivation.

Sind beispielsweise einige relevante erklärende Variablen $u_i = (u_{i1}, \ldots, u_{ip'})^\top$ gar nicht beobachtet, kann durch Hinzunahme eines **zufälligen** (skalarwertigen) Effektes $z_i = u_i^\top \gamma$ die ansonsten resultierende Überdispersion (Datenvariabilität ist größer als die Modellvariabilität) berücksichtigt werden. Diese Überlegung führt zu zufälligen Prädiktoren der Form

$$\eta_i = g(\mu_i) = x_i^\top \beta + z_i \,.$$

Andererseits könnten n unabhängige Gruppen von Beobachtungen $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ vorliegen, wofür innerhalb jeder Gruppe Abhängigkeit unter den y_{ij} besteht. Nimmt man einen zufälligen Effekt für sämtliche Responses einer Gruppe in deren Prädiktor auf, d.h.

$$\eta_{ij} = g(\mu_{ij}) = x_{ij}^\top \beta + z_i \,, \qquad j = 1, \ldots, n_i \,,$$

so wird dadurch die Korrelation in den Responsevariablen ein und derselben Gruppe (marginal) berücksichtigt. Da sich alle y_{ij} auf denselben Zufallseffekt z_i beziehen, gewinnt man automatisch ein Modell für deren Abhängigkeit.

Die nichtbeobachtbaren (**latenten**) z_i seien generell eine Zufallsstichprobe. Im Gegensatz zu früher verwenden wir jetzt zufällige Prädiktoren für den konditionalen Erwartungswert $\mu_i = \mathrm{E}(y_i|z_i)$ (bzw. $\mu_{ij} = \mathrm{E}(y_{ij}|z_i)$). Somit stammt die konditionale Verteilung der Response, gegeben die zufälligen Effekte, aus der Exponentialfamilie. Für die Berechnung des

MLEs soll jedoch die marginale Likelihood-Funktion maximiert werden. Da dies jedoch nur selten analytisch durchgeführt werden kann, wird oft als Alternative die sogenannte **EM-Schätzung** verwendet.

6.2 EM-Schätzer

Für den von Dempster, Laird & Rubin (1977) entwickelten EM-Algorithmus wird angenommen, dass die Daten aus einem beobachtbaren Teil y und einem nicht beobachtbaren Teil z zusammengesetzt sind. Die gemeinsame Dichte einer vollständigen Beobachtung (y, z) sei $f(y, z|\theta)$, wobei alle unbekannten Parameter im System zum Vektor $\theta \in \Theta$ zusammengefasst sind (hierbei bezeichnet Θ den Parameterraum). Somit gilt allgemein für die marginale Dichte $f(y|\theta)$ der Response y

$$\ell(\theta|y) = \log f(y|\theta) = \log \int f(y, z|\theta) dz \,. \tag{6.1}$$

Um den MLE $\hat{\theta}$ zu bestimmen, wird diese marginale Log-Likelihood Funktion $\ell(\theta|y)$ maximiert. Dabei stößt man aber in der Praxis häufig auf Probleme mit dem Integral (6.1). Mit der bedingten Dichte von $z|y$, gegeben durch

$$f(z|y;\theta) = \frac{f(y, z|\theta)}{f(y|\theta)} \,,$$

lässt sich (6.1) schreiben als $\ell(\theta|y) = \log f(y|\theta) = \log f(y, z|\theta) - \log f(z|y;\theta)$. Der MLE $\hat{\theta}$ maximiert gerade $\ell(\theta|y)$. Es gilt nun

$$\log f(y, z|\theta) = \log f(z|y;\theta) + \ell(\theta|y) \,.$$

Da per Annahme der Teil z nicht beobachtet werden kann, ersetzen wir diese fehlende Information durch ihren konditionalen Erwartungswert, gegeben das was beobachtet vorliegt (also gegeben y). Für dessen Berechnung verwenden wir einen beliebigen zulässigen Parameterwert $\theta_0 \in \Theta$, d.h. bezüglich $f(z|y, \theta_0)$. Der konditionale Erwartungswert von $\ell(\theta|y)$, gegeben die beobachteten Daten y, ist aber wiederum $\ell(\theta|y)$, womit gilt

$$\begin{aligned}
\mathrm{E}\Big(\log f(y, z|\theta) \Big| y, \theta_0 \Big) &= \mathrm{E}\Big(\log f(z|y, \theta) \Big| y, \theta_0 \Big) + \mathrm{E}\Big(\ell(\theta|y) \Big| y, \theta_0 \Big) \\
\int \log f(y, z|\theta) f(z|y, \theta_0) dz &= \int \log f(z|y, \theta) f(z|y, \theta_0) dz + \int \ell(\theta|y) f(z|y, \theta_0) dz \\
Q(\theta|\theta_0) &= H(\theta|\theta_0) + \ell(\theta|y) \,.
\end{aligned} \tag{6.2}$$

Die Maximierung von $\ell(\theta|y)$ in θ ist daher äquivalent mit der Maximierung der Differenz $Q(\theta|\theta_0) - H(\theta|\theta_0)$. Bemerke, dass (6.2) für beliebige Parameterwerte $\theta \in \Theta$ hält, also auch für $\theta = \theta_0$ wofür wir

$$Q(\theta_0|\theta_0) = H(\theta_0|\theta_0) + \ell(\theta_0|y)$$

erhalten. Somit resultiert als Differenz zu (6.2) für die marginale Log-Likelihood Funktion

$$\ell(\theta|y) - \ell(\theta_0|y) = Q(\theta|\theta_0) - Q(\theta_0|\theta_0) - \Big[H(\theta|\theta_0) - H(\theta_0|\theta_0)\Big]\,. \tag{6.3}$$

Allgemein liefert die **Jensen** Ungleichung für eine konkave Funktion, wie $g(x) = \log(x)$, die Abschätzung $\mathrm{E}(g(X)) \le g(\mathrm{E}(X))$. Damit folgt für beliebiges $\theta \in \Theta$

$$\begin{aligned} H(\theta|\theta_0) - H(\theta_0|\theta_0) &= \int \log \frac{f(z|y,\theta)}{f(z|y,\theta_0)} f(z|y,\theta_0)dz \\ &= \mathrm{E}\Big(\log \frac{f(z|y,\theta)}{f(z|y,\theta_0)}\Big|y,\theta_0\Big) \\ &\le \log \mathrm{E}\Big(\frac{f(z|y,\theta)}{f(z|y,\theta_0)}\Big|y,\theta_0\Big) \\ &= \log \int \frac{f(z|y,\theta)}{f(z|y,\theta_0)} f(z|y,\theta_0)dz \\ &= \log \int f(z|y,\theta)dz = \log(1) = 0\,, \end{aligned}$$

also

$$H(\theta|\theta_0) - H(\theta_0|\theta_0) \le 0\,. \tag{6.4}$$

Dies hat wiederum zur Folge, dass wir (6.3) schreiben können als

$$\ell(\theta|y) - \ell(\theta_0|y) \ge Q(\theta|\theta_0) - Q(\theta_0|\theta_0)\,.$$

Sei θ' jener Wert von θ, der für ein gegebenes (festes) θ_0 die Funktion $Q(\theta|\theta_0)$ maximiert. Somit gilt

$$Q(\theta'|\theta_0) - Q(\theta|\theta_0) \ge 0 \qquad \text{und daher auch} \qquad Q(\theta'|\theta_0) - Q(\theta_0|\theta_0) \ge 0\,.$$

Dies zeigt aber, dass durch die Maximierung von $Q(\theta|\theta_0)$ die Log-Likelihood nach diesem (EM-) Schritt zumindest nicht verkleinert wird, denn das letzte Ergebnis impliziert

$$\ell(\theta'|y) - \ell(\theta_0|y) \ge 0\,.$$

Stationarität: Weiters resultiert durch Differenzieren von (6.2) die Identität

$$\frac{\partial}{\partial\theta} Q(\theta|\theta_0) = \frac{\partial}{\partial\theta} H(\theta|\theta_0) + \frac{\partial}{\partial\theta}\ell(\theta|y)\,.$$

Nun gilt aber wegen (6.4) gerade $H(\theta|\theta_0) \le H(\theta_0|\theta_0)$ für alle $\theta \in \Theta$. Unter dieser Extremalbedingung folgt, dass

$$\frac{\partial}{\partial\theta} H(\theta|\theta_0)|_{\theta=\theta_0} = 0$$

hält was bedeutet, dass die Funktion $H(\theta|\theta_0)$ stationär ist in $\theta = \theta_0$. Ist somit $Q(\theta|\theta_0)$ stationär in $\theta = \theta_0$, dann ist dies dort auch $\ell(\theta|y)$.

Der EM-Algorithmus ist zweistufig. Im **E-Schritt** wird der bedingte Erwartungswert $Q(\theta|\theta_0)$ für gegebenes θ_0 berechnet. Danach wird im **M-Schritt** diese Funktion $Q(\theta|\theta_0)$ bezüglich θ maximiert. Sei das Ergebnis dieser Maximierung θ', so wird damit wiederum ein E-Schritt mit aktualisierten $\theta_0 = \theta'$ durchgeführt. Diese Iteration wiederholt man bis zur Konvergenz im marginalen Likelihood Schätzer $\hat{\theta}$.

Self-consistency des EM-Algorithmus: Falls der MLE $\hat{\theta}$ ein globales Maximum von $\ell(\theta|y)$ darstellt, so muss dieser auch

$$Q(\hat{\theta}|\hat{\theta}) \geq Q(\theta|\hat{\theta})$$

genügen. Ansonsten würde es ja einen Parameterwert θ^* geben mit der Eigenschaft

$$Q(\hat{\theta}|\hat{\theta}) < Q(\theta^*|\hat{\theta})\,,$$

was wiederum

$$\ell(\theta^*|y) > \ell(\hat{\theta}|y)$$

impliziert und somit einen Widerspruch darstellt zur Annahme, dass $\hat{\theta}$ das globale Maximum von $\ell(\theta|y)$ ist.

Anstelle des Integrals in (6.1) muss also beim EM-Algorithmus das Integral $Q(\theta|\theta_0)$ in (6.2) iterativ berechnet und maximiert werden. Wie das folgende Beispiel zeigt, ist diese Berechnung in einigen Anwendungen möglich.

6.2.1 Beispiel: Endliche diskrete Mischungen

Sei $y_1, \ldots, y_n$ eine Stichprobe aus einer diskreten Mischung von K Dichtekomponenten mit gruppenspezifischen 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 den Anteilen $\pi_1, \ldots, \pi_K$ mit $\sum_k \pi_k = 1$. Die marginale Dichte des i-ten Elements y_i kann daher geschrieben werden als

$$f(y_i|\theta, \phi, \pi) = \sum_{k=1}^{K} \pi_k f(y_i|\theta_k, \phi)\,.$$

Somit ist die Likelihood Funktion

$$\begin{aligned} L(\theta, \phi, \pi|y) &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i|\theta_k, \phi) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_{ik} \qquad \text{mit} \quad f_{ik} = f(y_i|\theta_k, \phi)\,. \end{aligned} \tag{6.5}$$

Hier sind θ, ϕ und π unbekannt, also sind $K + 1 + (K - 1) = 2K$ Parameter zu schätzen. Aus der Log-Likelihood Funktion

$$\ell(\theta, \phi, \pi|y) = \log L(\theta, \phi, \pi|y) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_{ik} \right)$$

erhält man beispielsweise bezüglich ϕ

$$\begin{aligned}\frac{\partial}{\partial\phi}\ell(\theta,\phi,\pi|y) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\pi_k f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}}\frac{\partial\log f_{ik}}{\partial\phi}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\frac{\partial\log f_{ik}}{\partial\phi} \qquad \text{mit} \quad w_{ik}=\frac{\pi_k f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}}\,, \end{aligned} \tag{6.6}$$

wobei die Gewichte w_{ik} von allen Parametern abhängen, also $w = w(\theta,\phi,\pi)$ gilt. Diese EM-Score-Funktion ist eine gewichtete Summe von (Likelihood-) Score-Funktionen, in der die Summanden aus jeder einzelnen Komponentendichte der Mischung stammen.

Die Gewichte w_{ik} können auch aus der Sicht eines formalen Bayes-Modells sehr gut interpretiert werden. Da per Definition

$$\Pr(k|y_i) = \frac{\Pr(k)\Pr(y_i|k)}{\sum_{l=1}^{K}\Pr(l)\Pr(y_i|l)} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}} = w_{ik} \tag{6.7}$$

gilt, haben diese Gewichte folgende Bedeutung: Wähle die Komponente k zufällig mit Wahrscheinlichkeit π_k. Ziehe nun 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}.

Die w_{ik} sind auch in den Score-Funktionen bezüglich θ und π wesentlich. So ist

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\ell(\theta,\phi,\pi|y) &= \sum_{i=1}^{n}\frac{\pi_k\partial f_{ik}/\partial\theta_k}{\sum_{l=1}^{K}\pi_l f_{il}}\\ &= \sum_{i=1}^{n} w_{ik}\frac{\partial\log f_{ik}}{\partial\theta_k}\,. \end{aligned} \tag{6.8}$$

Für den Schätzer von π benötigt man die Randbedingung $\sum_k \pi_k = 1$, was unter Verwendung eines Lagrange Multiplikators zu

$$\frac{\partial}{\partial\pi_k}\left(\ell(\theta,\phi,\pi|y) - \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)\right) = \sum_{i=1}^{n}\frac{f_{ik}}{\sum_{l=1}^{K}\pi_l f_{il}} - \lambda = \sum_{i=1}^{n} w_{ik}\frac{1}{\pi_k} - \lambda$$

führt. Setzt man diesen Term Null, so resultiert $\pi_k = \sum_i w_{ik}/\lambda$. Summiert man über alle Komponenten k, ergibt sich wegen (6.7) als Lösung $1 = \sum_k \pi_k = \sum_i\sum_k w_{ik}/\lambda = n/\lambda$, also $\lambda = n$. Damit folgt als Score für die Anteilswerte

$$\frac{1}{\pi_k}\sum_{i=1}^{n} w_{ik} - n\,. \tag{6.9}$$

Der **MLE** ist definiert als die simultane Nullstelle der Gleichungssysteme (6.6), (6.8) und (6.9). Dessen Berechnung kann in diesem Fall daher ziemlich aufwendig sein, da sowohl

die marginalen Dichtefunktionen f_{ik} als auch die Gewichte w_{ik} meist recht komplexe nichtlineare Funktionen in allen Parametern sind.

Für die Anwendung des EM-Algorithmus (Berechnung der **EM-Schätzer**) wird jede Beobachtung y_i wegen ihrer unbekannten Gruppenzugehörigkeit als unvollständig betrachtet. Zu jedem y_i definiert man daher zusätzlich einen nichtbeobachtbaren Indikatorvektor $z_i = (z_{i1}, \ldots, z_{iK})$ mit

$$z_{ik} = \begin{cases} 1 & \text{falls } y_i \text{ aus Gruppe } k \\ 0 & \text{sonst.} \end{cases}$$

Die gemeinsame Dichte von (y, z) lautet damit

$$f(y, z|\theta, \phi, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i|\theta_k, \phi)^{z_{ik}} \pi_k^{z_{ik}} ,$$

also

$$\log f(y, z|\theta, \phi, \pi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) .$$

Im E-Schritt berechnen wir für Werte θ_0, ϕ_0, π_0 die Funktion $Q(\theta, \phi, \pi|\theta_0, \phi_0, \pi_0)$, also

$$\mathrm{E}\Big(\log f(y, z|\theta, \phi, \pi) \Big| y, \theta_0, \phi_0, \pi_0 \Big) = \mathrm{E}\left(\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) \Big| y_i, \theta_0, \phi_0, \pi_0 \right) .$$

Durch das Konditionieren hat man im Argument dieses Erwartungswertes außer den z_{ik} nur feste Größen. Somit folgt wegen $\mathrm{E}(z_{ik}|y_i, \theta_0, \phi_0, \pi_0) = \Pr(z_{ik} = 1|y_i, \theta_0, \phi_0, \pi_0) = w_{ik}$, wobei die w_{ik} in θ_0, ϕ_0 und π_0 ausgewertet sind,

$$Q(\theta, \phi, \pi|\theta_0, \phi_0, \pi_0) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Big(\log f(y_i|\theta_k, \phi) + \log \pi_k \Big) . \tag{6.10}$$

Dies wird im **E**-Schritt berechnet, was wiederum im wesentlichen auf der Berechnung der Größen w_{ik} beruht. Die Parameter θ, ϕ und π stellen jene Größen dar, die im folgenden **M**-Schritt für (im Unterschied zur Berechnung der MLEs) **feste Gewichte** w_{ik} maximiert werden. Dazu müssen aber gerade die Nullstellen der gewichteten Score Funktionen (6.6), (6.8), und (6.9) berechnet werden.

Mischung von Normalverteilungen

Liegt eine Mischung von K Normalverteilungen mit unterschiedlichen Erwartungen μ_k und konstanter Varianz σ^2 vor, so ist

$$\log f_{ik} = -\frac{1}{2} \log \Big(2\pi\sigma^2 \Big) - \frac{(y_i - \mu_k)^2}{2\sigma^2}$$

und es folgt

$$\frac{\partial \log f_{ik}}{\partial \mu_k} = \frac{y_i - \mu_k}{\sigma^2}\,, \qquad \frac{\partial \log f_{ik}}{\partial \sigma^2} = -\frac{1}{2}\Big(\frac{1}{\sigma^2} - \frac{(y_i-\mu_k)^2}{\sigma^4}\Big)\,.$$

Für geeignete Werte μ_k, σ^2 und π_k berechnet man im **E**-Schritt die Gewichte w_{ik}. Im **M**-Schritt wird mit diesen w_{ik} die Maximierung durchgeführt. Aus (6.8) folgt

$$\sum_{i=1}^{n} w_{ik}\frac{y_i-\mu_k}{\sigma^2} = 0 \qquad \Longrightarrow \qquad \hat{\mu}_k = \frac{\sum_{i=1}^n w_{ik} y_i}{\sum_{i=1}^n w_{ik}}\,.$$

Mit (6.6) ergibt sich

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Big(-\frac{1}{2\sigma^2} + \frac{(y_i-\mu_k)^2}{2\sigma^4}\Big) = 0 \qquad \Longrightarrow \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(y_i-\hat{\mu}_k)^2\,.$$

Schließlich wird noch (6.9) benötigt, d.h.

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} w_{ik}\,.$$

Mit diesen $\hat{\mu}_k$, $\hat{\sigma}^2$ und $\hat{\pi}_k$ berechnet man im **E**-Schritt aktualisierte w_{ik} und geht damit zum **M**-Schritt zurück.

6.3 Überdispersionsmodelle

Für einen zufälliger Effekt z mit Dichte $f(z)$ sei die bedingte Dichte einer Response y gleich $f(y|z,\theta)$. Somit folgt als gemeinsame Dichte $f(y,z|\theta) = f(y|z,\theta)f(z|\theta)$. Als marginale Dichte ergibt sich

$$f(y|\theta) = \int f(y,z|\theta)dz = \int f(y|z,\theta)f(z|\theta)dz\,. \tag{6.11}$$

Die bedingte Dichte von $z|y$ ist hier gegeben durch $f(z|y,\theta) = f(y|z,\theta)f(z|\theta)/f(y|\theta)$. Als zu maximierende Funktion resultiert daher für die gesamte Stichprobe von unabhängigen Responsevariablen

$$\begin{aligned} Q(\theta|\theta_0) &= \sum_{i=1}^{n}\int \log f(y_i,z_i|\theta) f(z_i|y_i,\theta_0)dz_i \\ &= \sum_{i=1}^{n}\int \log f(y_i,z_i|\theta)\frac{f(y_i|z_i,\theta_0)}{f(y_i|\theta_0)} f(z_i|\theta_0)dz_i \\ &= \sum_{i=1}^{n}\frac{1}{f(y_i|\theta_0)}\int \log f(y_i,z_i|\theta) f(y_i|z_i,\theta_0) f(z_i|\theta_0)dz_i\,. \end{aligned} \tag{6.12}$$

Dabei sind sowohl $f(y_i|\theta_0)$ (siehe (6.11)) als auch das Integral in (6.12) selbst sehr unangenehm und nur selten analytisch geschlossen darstellbar. Zwei Ansätze sollen nun diskutiert werden. Im ersten nehmen wir an, dass z aus einer Normalverteilung stammt und daraufhin werden beide Integrale durch eine K-Punkt **Gauss-Quadratur** approximiert. Falls diese Verteilungsannahme nicht getroffen werden kann, wird im zweiten Ansatz $f(z)$ durch den **nicht-parametrischen Maximum-Likelihood, NPML** Schätzer $\hat{f}(z)$ ersetzt und die dadurch resultierende Zielfunktion maximiert.

6.3.1 Normalverteilte zufällige Effekte

Für $f(z) = \phi(z)$ liefert die Gauss-Quadratur als Approximation

$$f(y|\theta_0) = \int f(y|z,\theta_0)\phi(z)dz \approx \sum_{k=1}^{K} f(y|z_k,\theta_0)\pi_k$$

mit "bekannten" Massen π_k auf den festen (bekannten) Massepunkten z_k. Wendet man dieses Approximationsverfahren auch auf das zweite Integral an, so resultiert

$$\begin{aligned} Q(\theta|\theta_0) &\approx \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \log f(y_i, z_k|\theta) f(y_i|z_k,\theta_0)\pi_k}{\sum_{j=1}^{K} f(y_i|z_j,\theta_0)\pi_j} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log f(y_i, z_k|\theta) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Big(\log f(y_i|z_k,\theta) + \log \pi_k\Big) \end{aligned} \tag{6.13}$$

mit den Gewichten

$$w_{ik} = \frac{\pi_k f(y_i|z_k,\theta_0)}{\sum_{j=1}^{K} f(y_i|z_j,\theta_0)\pi_j}\,, \tag{6.14}$$

die in $\theta = \theta_0$ ausgewertet sind und daher für die folgende Maximierung feste Größen darstellen. Die Approximation (6.13) ist identisch der Funktion Q in (6.10) und die Gewichte (6.14) entsprechen den Größen (6.7). Daher ist die Schätzung der Parameter im Modell mit zufälligen Effekten äquivalent der Schätzung in diskreten Mischmodellen, wobei hier aber die Massen π_k bekannte Größen darstellen und bedingte Dichten gemischt werden. Schreibt man den linearen Prädiktor dieses Modells als

$$\eta_i = x_i^\top \beta + \sigma_z z_i\,, \qquad \text{mit} \quad z_i \overset{iid}{\sim} \text{Normal}(0,1)\,, \tag{6.15}$$

so wird bei der Maximierung von (6.13) bezüglich β gerade die mit (6.14) gewichtete Likelihood einer Stichprobe mit nK Elementen maximiert. Diese erweiterte Struktur erhält man, wenn jeder Beobachtung y_i gerade K lineare Prädiktoren der Form

$$\eta_{ik} = x_i^\top \beta + \sigma_z z_k$$

mit bekannten Massestellen z_k, jedoch unbekannten Parametern β und σ_z zugeordnet werden. Dies bedeutet, dass jede Beobachtung K-fach betrachtet wird und deren lineare Prädiktoren jeweils mit z_k erweitert werden. Der Schätzer zu dieser neuen Spalte stellt somit den Schätzer von σ_z dar.

Für jede Iteration der EM-Schätzung ist daher ein Programm notwendig, mit dem eine gewichtete ML-Schätzung eines GLMs gerechnet werden kann. Als Daten verwendet man dazu die erweiterte 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

6.3.2 Zufällige Effekte aus unbekannter Verteilung

Falls über $f(z)$ keine Annahmen getroffen werden können, so sollte $f(z)$ nichtparametrisch geschätzt werden. Der Schätzer $\hat{f}(z)$ ist dann eine Wahrscheinlichkeitsfunktion, definiert auf K geschätzten Massepunkte z_k mit Massen π_k. Entsprechendes Vorgehen wie bei der Gauss-Quadratur liefert (6.13) als Zielfunktion mit Gewichten (6.14), wobei jetzt jedoch zusätzlich auch die z_k und π_k unbekannt sind. Daher wird hier auch noch die Schätzung (6.9) für π_k verwendet. Für den hierbei betrachteten linearen Prädiktor

$$\eta_i = x_i^\top \beta + z_i\,, \qquad \text{mit} \quad z_i \overset{iid}{\sim} F(z)\,, \tag{6.16}$$

folgt

$$\eta_{ik} = x_i^\top \beta + z_k\,,$$

wobei die K Stellen z_k unbekannt sind. Mittels eines K-stufigen Faktors kann der Prädiktor umgeschrieben werden zu

$$\eta_{ik} = x_i^\top \beta + z_2 \cdot 0 + \dots + z_{k-1} \cdot 0 + z_k \cdot 1 + z_{k+1} \cdot 0 + \dots + z_K \cdot 0$$

und die unbekannten $z_2, \dots, z_K$ stellen gerade die Parameter zu den $K-1$ Dummy-Variablen $(0,1)$ dar. Da x einen Intercept enthält, stellt dieser den Parameter zu z_1 dar (Referenzklasse). Der dazu notwendige, erweiterte Datensatz hat wiederum nK Elemente. Für einen M-Schritt wird eine gewichtete ML-Schätzung für das GLM bezüglich der

erweiterten Struktur

y	w	β				z			
y_1	w_{11}	1	x_{11}	...	$x_{1,p-1}$	0	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n1}	1	x_{n1}	...	$x_{n,p-1}$	0	0	...	0
y_1	w_{12}	1	x_{11}	...	$x_{1,p-1}$	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n2}	1	x_{n1}	...	$x_{n,p-1}$	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_1	w_{1K}	1	x_{11}	...	$x_{1,p-1}$	0	0	...	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{nK}	1	x_{n1}	...	$x_{n,p-1}$	0	0	...	1

durchgeführt. Im folgenden E-Schritt werden nur die Gewichte w_{ik} entsprechend aktualisiert. Als nichtparametrischen Maximum-Likelihood Schätzer (NPMLE) für die Verteilung 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 $\hat{z}$ mit geschätzten Wahrscheinlichkeitsmassen $\hat{\pi}$.

6.3.3 Prädiktionen bei der NPML Schätzung

Als Schätzung für das **konditionale Modell** $\mu_{ik} = \mathrm{E}(y_i|z_k)$ erhält man für jede Komponente $k = 1, \ldots, K$

$$g(\hat{\mu}_{ik}) = \hat{\eta}_{ik} = x_i^\top \hat{\beta} + \hat{z}_k\,, \qquad \text{also} \quad \hat{\mu}_{ik} = g^{-1}(\hat{\eta}_{ik})\,.$$

Diese konditionalen Erwartungswertmodelle stellen parallele Regressionsebenen im (η, x)-Raum dar.

Interessant ist auch eine Schätzung des **marginalen Modells**, also von $\mathrm{E}(y_i)$. Die alternative Bezeichnung *population average* wird in der Literatur nicht konsistent genutzt. Da die marginale Verteilung der y_i die gemischte Exponentialfamilie ist, folgt

$$\begin{aligned}
\mathrm{E}(y_i) &= \int y_i \int f(y_i|z_i) f(z_i) dz_i dy_i \\
&\approx \int y_i \sum_{k=1}^{K} f(y_i|z_k)\pi_k dy_i = \sum_{k=1}^{K} \pi_k \int y_i f(y_i|z_k) dy_i = \sum_{k=1}^{K} \pi_k \mathrm{E}(y_i|z_k)
\end{aligned}$$

und es resultiert der Schätzer

$$\widehat{\mathrm{E}}(y_i) = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_{ik}\,.$$

Der **posterior Mean** eines Zufallseffekts z_i, gegeben die Response y_i, ist gleich

$$\begin{aligned} \mathrm{E}(z_i|y_i) &= \int z_i f(z_i|y_i) dz_i \\ &= \int z_i \, \frac{f(y_i|z_i) f(z_i)}{\int f(y_i|z_i) f(z_i) dz_i} \, dz_i \\ &\approx \sum_{k=1}^{K} z_k \, \frac{f(y_i|z_k)\pi_k}{\sum_{l=1}^{K} f(y_i|z_l)\pi_l} \, . \end{aligned}$$

Der **empirische 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. 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

$$\begin{aligned} \widetilde{\eta}_i &= x_i^\top \hat{\beta} + \widetilde{\mathrm{E}}(z_i|y_i) = \sum_{k=1}^{K} \hat{w}_{ik} \Big(x_i^\top \hat{\beta} + \hat{z}_k \Big) \\ &= \sum_{k=1}^{K} \hat{w}_{ik} \hat{\eta}_{ik} \end{aligned}$$

und damit folgt wiederum

$$\widetilde{\mu}_i = g^{-1}(\widetilde{\eta}_i) \, .$$

6.3.4 Beispiel: Matched Pairs

In einer Studie des Landeshygienikers für die Steiermark wurden alle zwei Wochen über ein volles Jahr u.a. bei zwei benachbarten Grazer Wohnsiedlungen (`site=6` bzw. `site=7`) die Bakterienkonzentrationen in der Außenluft gemessen. Die beiden Siedlungen unterscheiden sich hauptsächlich bezüglich des Vorhandenseins einer Kompostieranlage bei `site=7`. Es ist bekannt, dass die Anzahl kolonienbildender luftgetragener Mikroorganismen (cfu: colonies forming units) vom vorherrschenden Wetter abhängig ist. Zu diesen 26×2 Beobachtungen `bac` gehören deshalb auch die Temperatur `temp` und die Luftfeuchtigkeit `humi`.

Die Erfassung von Bioaerosolen erfolgt mit aktiven Probenahmesystemen (Luftkeimsammler) und anschließender Aufarbeitung und Auswertung der Proben im Labor.
Als Messgerät für diese Bioaerosole wurde ein sechsstufiger Andersen Kaskadenimpaktor verwendet, dessen 6 Stufen unterschiedlich große Keime filtern. Dabei sind besonders die kleinen Mikroorganismen (`stage > 3`) für den Menschen gefährlich, da diese bis zur Lunge vordringen können. Es ist zu untersuchen, ob es einen relevanten Kompostieranlageneffekt gibt.

Die Erstanalyse dieser Daten ergibt für den 28. November einen extrem großen cfu Wert von 36 auf `stage = 6` (die Gesamtanzahl von 50 bei dieser Probe deutet auf eine Kontaminierung hin), weshalb beide Tripel von der folgenden Analyse ausgeschlossen werden. Außerdem muss die Struktur der Daten entsprechend modifiziert werden, so dass jede einzelne Response in einer eigenen Zeile steht und denselben Namen hat.

```
> bacteria <- read.table("bacteria.dat", header=TRUE)
> bac <- bacteria[(bacteria$site > 5), ] # only obs. from site 6 & 7
> b.total <- bac$b4 + bac$b5 + bac$b6    # total cfu's on stages 4 to 6
> date.crit <- bac$date[b.total > 20]    # date is critical, if total cfu > 20
> bac <-  bac[(bac$date != date.crit), ] # without the extreme pair of obs.

> var.sel <- c("date", "site", "humi", "temp") # relevant columns
> bac <- rbind(cbind(bac[ ,var.sel], stage = 4, cfu = bac$b4),
+              cbind(bac[ ,var.sel], stage = 5, cfu = bac$b5),
+              cbind(bac[ ,var.sel], stage = 6, cfu = bac$b6))

> bac$date  <- factor(bac$date)
> bac$site  <- factor(bac$site)
> bac$stage <- factor(bac$stage)
> attach(bac)
```

Nach diesen notwendigen Datenmanipulationen können wir uns auf die Modellfindung für die Response `cfu` in Abhängigkeit von `site` und `stage` sowie den beiden Wetterwerten `temp` und `humi` konzentrieren.

Da es sich hierbei um Anzahlen (cfu's) handelt, wird zuerst ein loglineares Poissonmodell betrachtet. Wir sind vor allem am Verhalten an den beiden Orten auf den drei untersten Stufen interessiert und wollen auf das jeweils vorherrschende Wetter standardisieren (dieses ist zwar an den beiden Orten ähnlich, jedoch über das gesamte Jahr recht unterschiedlich). Wir zeigen zuerst, dass die Information der Luftfeuchtigkeit zusätzlich zur Temperatur keine Relevanz hat.

```
> 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) + stage:site
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       142     258.83
2       140     255.09  2    3.746   0.1537
```

```
> 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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 284.23  on 149  degrees of freedom
Residual deviance: 258.83  on 142  degrees of freedom
AIC: 575.79
```

Nur der lineare und quadratische Effekt der Temperatur scheinen relevant. Bedenklich ist die Überdispersion (Deviance 258.83 bei df=142). Dies könnte im Fehlen von weiteren Prädiktorvariablen begründet sein, z.B. Wind oder Luftdruck. Ein Modell mit zufälligen Effekten sollte dies vielleicht zumindest teilweise zu kompensieren vermögen.

Wir beginnen mit der Hinzunahme eines zufälligen Intercepts und nehmen dafür vorerst an, dass dieser aus einer normalverteilten Population stammt. Eine Möglichkeit, derartige Modelle zu handhaben besteht in der Verwendung des Pakets `npmlreg` von J. Einbeck, R. Darnell und J. Hinde (auch die Bibliothek `statmod` von G. Smyth et al. wird dazu benötigt).

Die Funktion `alldist` erlaubt die Modellierung mit zufälligen Effekten (`random = ~1` spezifiziert hierbei, dass gerade der Intercept zufällig sein soll). Um bei der Approximation durch die Gauß-Hermite Quadratur den Fehler so gering wie möglich zu halten, definieren wir eine enorm große Anzahl von Quadraturpunkten (`k=100`). Bemerke, dass die Option `data=` zwingend erforderlich ist.

```
> library(npmlreg)
> gq <- alldist(cfu ~ stage*site + temp + I(temp^2), data=bac, random = ~1,
+               family=poisson, random.distribution = "gq", k=100)
1 ..2 ..3 ..4 ..5 ..6 ..
EM algorithm met convergence criteria at iteration #  6
Disparity trend plotted.

> 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
```

Unter **Disparität** versteht man die negative doppelte Log-Likelihood, $-2 \log L(\hat{\theta}|y)$. Die Entwicklung dieses Maßes über die Iterationen hinweg bis zur Konvergenz zeigt der Disparity Plot. Gestartet wird die Iteration mit dem Modell ohne zufälligen Effekten. Bereits nach einer Iteration liegt der Wert der Disparität in der Nähe des Wertes bei Konvergenz.

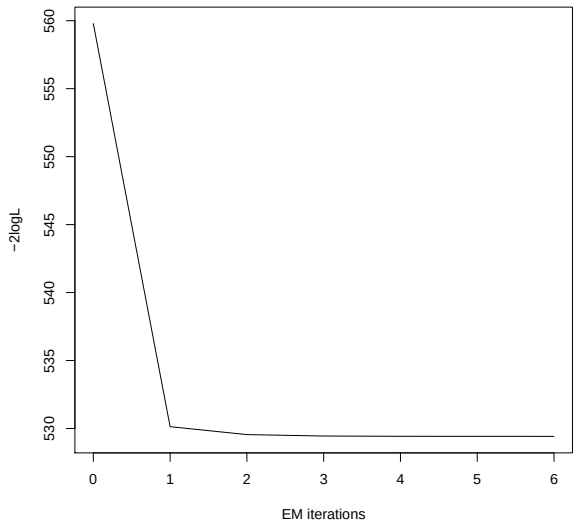

Abbildung 6.1: Disparity Plot zur Schätzung des Random Intercept Models mit Gauß-Hermite Quadratur.

Die geschätzte Standardabweichung des zufälligen Intercepts ist $\hat{\sigma}_z = 0.59$. Sämtliche Standardfehler und t-Werte stammen nur aus der letzten Iteration beim EM-Algorithmus (gewichtetes GLM bezüglich der erweiterten Datenstruktur) und sind nicht direkt interpretierbar.

Alternativ wollen wir auch eine nichtparametrische Maximum Likelihood Schätzung durchführen und beginnen mit 1 Massepunkt. Die Ergebnisse sind natürlich dieselben wie beim loglinearen Poissonmodell zuvor. Wir erhöhen sukzessive die Anzahl an Massepunkten und erhalten bei 2 Massepunkten Konvergenz nach 56 Iterationen.

```
> 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 ..15 ..16 ..17 ..
EM algorithm met convergence criteria at iteration #  56
Disparity trend plotted.
EM Trajectories plotted

> 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
```

Zusätzlich zum bereits bekannten Disparity Plot werden jetzt auch noch die EM Trajektorien grafisch dargestellt. Die Disparität erreicht nach etwas mehr als 20 Iterationen ihre etwaige endgültige Größenordnung. Die Verläufe der Schätzungen beider Massepunkte über die EM-Iterationen hinweg zeigen große Stabilität. Die Punkte am rechten Rand stellen Residuen (fixed part residuals) auf der Skala der linearen Prädiktoren dar, d.h.

$$r_i = g(y_i) - x_i^\top \hat{\beta}\,.$$

Dadurch wird die Lage dieser Massepunkte zu den Daten bewertet.

Es resultiert eine nur etwas geringere Disparität als zuvor bei normalverteilten Effekten. Die beiden Massepunkte nehmen die Rolle von Intercepts ein und sind geschätzt

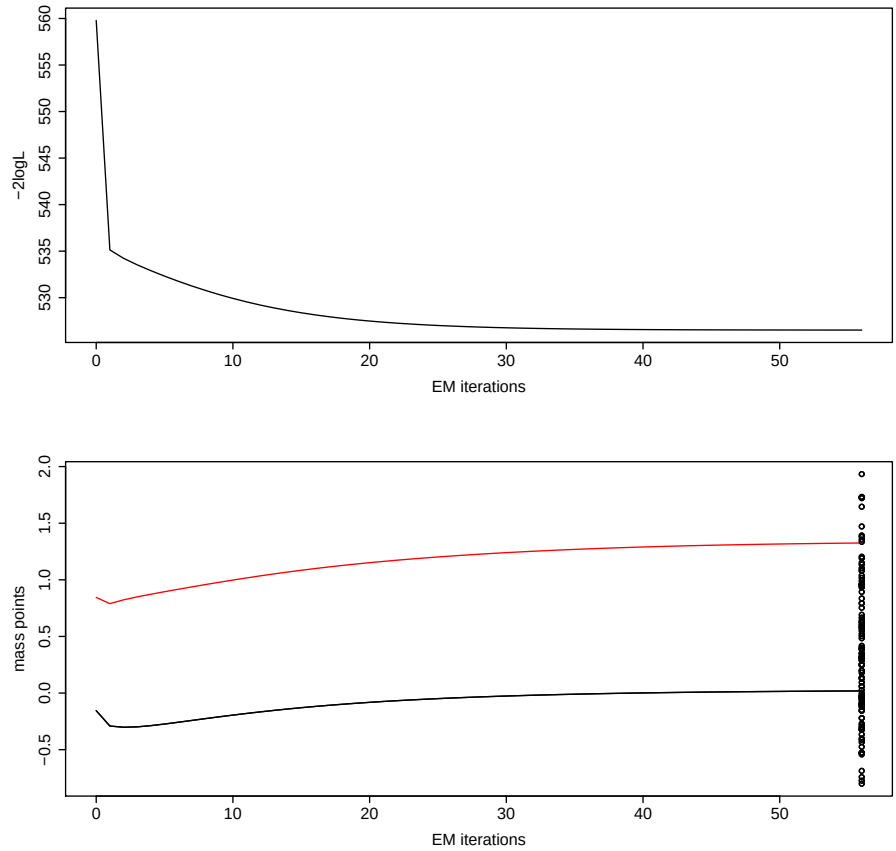

Abbildung 6.2: Disparity Plot zur Schätzung des Random Intercept Models mit NPMLE Technik bei 2 Massepunkten.

$(0.02, 1.33)$ mit Massen $(0.87, 0.13)$, was auf Abweichungen von der Normalverteilungsannahme hindeutet. Diese nichtparametrische Schätzung ergibt eine zweipunktige (diskrete) Effektverteilung mit Momenten

```
> (e <- sum(np2$masses * np2$mass.p))
[1] 0.189
> sqrt(sum(np2$masses * (np2$mass.p - e)^2))
[1] 0.439
```

Es verbleibt noch zu erwähnen, dass nur die beiden Parameter zur Temperatur (0.0585, -0.0021) mit den Ergebnissen bei der Quadratur (0.0617, -0.0022) direkt vergleichbar sind.

Wir erhöhen die Anzahl von Massepunkten und prüfen die resultierenden Deviancen.

```
> np3 <- alldist(cfu ~ stage*site + temp + I(temp^2), data=bac, random = ~1,
                 family=poisson, random.distribution = "np", k=3)
```

```
> np4 <- alldist(cfu ~ stage*site + temp + I(temp^2), data=bac, random = ~1,
                 family=poisson, random.distribution = "np", k=4)
> np5 <- alldist(cfu ~ stage*site + temp + I(temp^2), data=bac, random = ~1,
                 family=poisson, random.distribution = "np", k=5)

> np3$deviance
[1] 225.583
> np4$deviance
[1] 225.609
> np5$deviance
[1] 225.52
```

Diese bleiben ziemlich konstant, wenn man die Anzahl an Punkten erhöht. Man bemerke jedoch, dass die Anzahl an Freiheitsgraden pro zusätzlichen Massepunkt um 2 (1 Massepunkt und 1 Masse) reduziert wird. Somit bleiben wir beim Modell `np2`, welches nur 2 Massepunkte erlaubt. Interessant ist jedoch zu sehen, was bei der Erhöhung dieser Anzahl genauer passiert. Dies ist in der nächsten Abbildung für die Wahl $K = 5$ dargestellt.

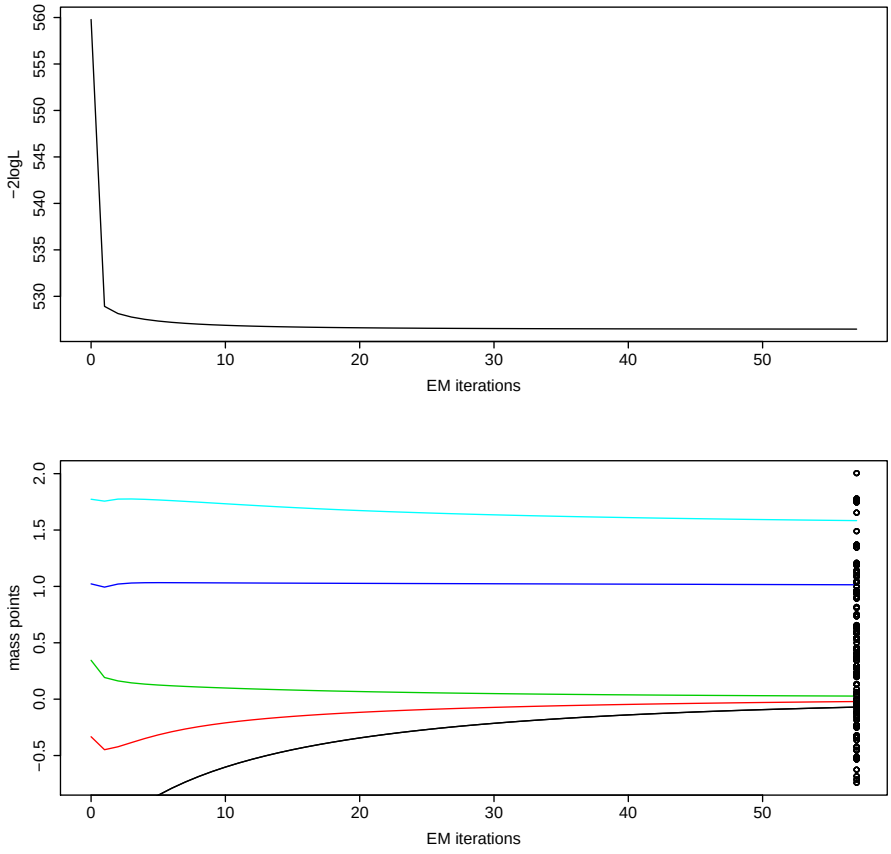

Abbildung 6.3: Grafischer Output für das nichtparametrisch geschätzte Modell mit $K = 5$.

Deutlich ist zu sehen, dass sich die ersten 3 Massepunkte über die Iterationen hinweg

stark annähern und eigentlich nur einen Punkt (der etwa den Wert Null hat) darstellen.

```
> 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
```

Auch die Summe deren Massen ist mit 0.84 etwa ähnlich groß wie die erste Masse 0.87 im Modell `np2` mit 2 Punkten. Die beiden anderen Punkte sind nur etwas kleiner und etwas größer als der zweite Punkt (1.3) im Modell `np2`.

Die geschätzten marginalen Erwartungswerte erhält man direkt aus dem Objekt `np2`.

```
> t <- -10:30
> newobs <- expand.grid(list(stage=levels(stage), site=levels(site), temp=t))
> emm <- predict(np2, newdata=newobs, type="response") # estim. marginal means
> plot(newobs[ ,"temp"], emm, xlab="temperature",
+      ylab="estimated marginal means",
+      col=newobs[ ,"site"], pch=as.numeric(newobs[ ,"stage"]))
```

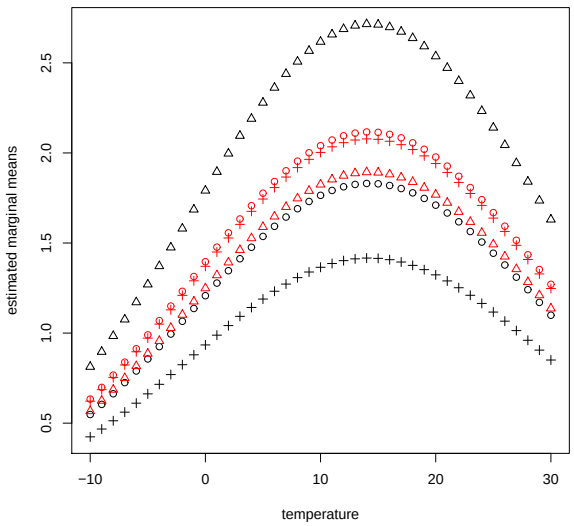

Abbildung 6.4: Marginale Modelle in Abhängigkeit von der Temperatur, der Siedlung mit(rot) sowie ohne(schwarz) Kompostieranlage, und den Stufen 4 (○), 5 (△), 6 (+) des Keimsammlers.

Die Lage der geschätzten marginalen Modelle in der Abbildung 6.4 entspricht gerade den geschätzten Interaktionsparametern. Die drei Kurven für die Siedlung ohne Kompostieranlage `site=6` sind schwarz, jene für die Siedlung mit Kompostieranlage `site=7` rot eingezeichnet. Als Plotsymbol wird ein Kreis ($\circ$) für `stage=4`, ein Dreieck ($\triangle$) für `stage=5` und ein Pluszeichen ($+$) für `stage=6` verwendet. Da keine Interaktion mit der Temperatur im Modell enthalten ist, schneiden sich diese Kurven auch nicht. Deutlich ist die Notwendigkeit der Interaktion zwischen den Stufen des Keimsammlers und dem Ort erkennbar. So ist `stage=5` die am höchsten belastete Stufe bei den Messungen in der Siedlung ohne Kompostieranlage und es sind gerade die Stufen `stage=4` und `stage=6` bei der Siedlung mit einer Kompostieranlage. Dies deutet stark darauf hin, dass sich die Partikelgrößen an diesen beiden Orten unterscheiden. Die für den Menschen gefährlichen kleinen (`stage=6`) scheinen in der Siedlung ohne Kompostieranlage viel seltener vorzukommen.

Zu den diagnostischen Aspekten zählt ein Plot, in dem die Residuen gegen die posterior Gewichte aufgetragen sind. Die Residuen sind hierbei definiert als Differenzen zwischen den beobachteten Responses und deren empirischen Bayesprädiktoren. Man sieht, dass nur sehr wenige Responses mit großen Residuen auch größere posteriori Wahrscheinlichkeiten für die zweite Komponente haben ($w_{i2} > 0.5$). Für die Mehrzahl gilt jedoch $w_{i1} > w_{i2}$, also eher eine Zugehörigkeit zur ersten Komponente.

```
> plot(np2$residuals, np2$post.prob[ , 2], xlab="residuals",
+      ylab="posterior probability for component 2")
```

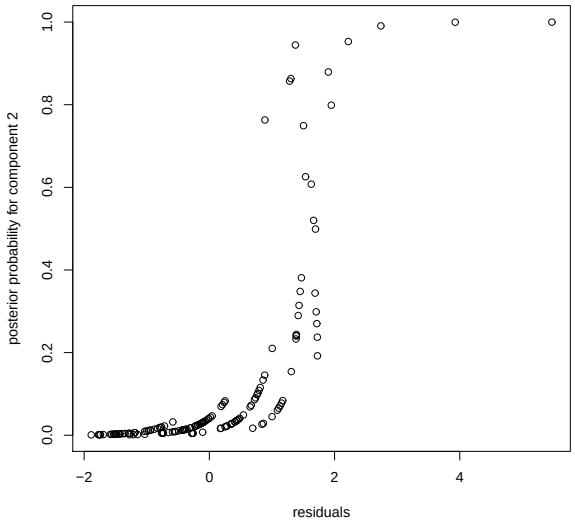

Abbildung 6.5: Residuen gegen die geschätzten posteriori Wahrscheinlichkeiten für die zweite Komponente.

Etwas anders sieht diese Abbildung aus, wenn man sie durch Aufruf von plot erzeugt. Hierbei sind die Residuen sogenannte fixed part residuals ($y_i - x_i^\top \hat{\beta}$), also ohne Einbeziehen des Intercepttterms).

```
> plot(np2) # default plot.opt=15 (for details see manual)
```

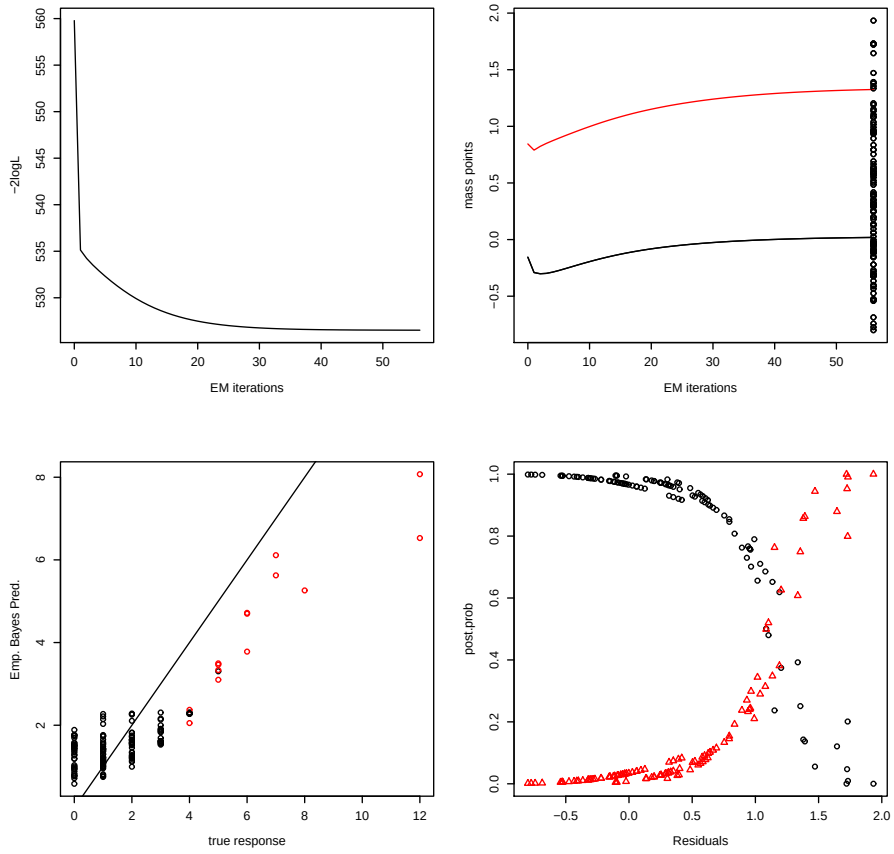

Abbildung 6.6: Ergebnis von Aufruf `plot()` ohne Spezifikation von `plot.opt`.

```
> eBz <- np2$post.prob %*% np2$mass.points # equals np2$post.int
> plot(cfu, eBz)
> mean(eBz)
[1] 0.1886
```

In der Abbildung 6.7 ist zu erkennen, dass die empirischen Bayes Schätzer der z_i (posterior intercepts) von den y_i abhängen. Diese Werte sind dann groß, wenn auch die cfu-Werte hoch sind. Ihr globales Mittel ist 0.1886, und entspricht etwa dem Intercept beim Modell mit Gauss-Quadratur (0.1714) oder dem geschätzten Erwartungswert des zufälligen Effektes beim NPMLE Ansatz (0.1886).

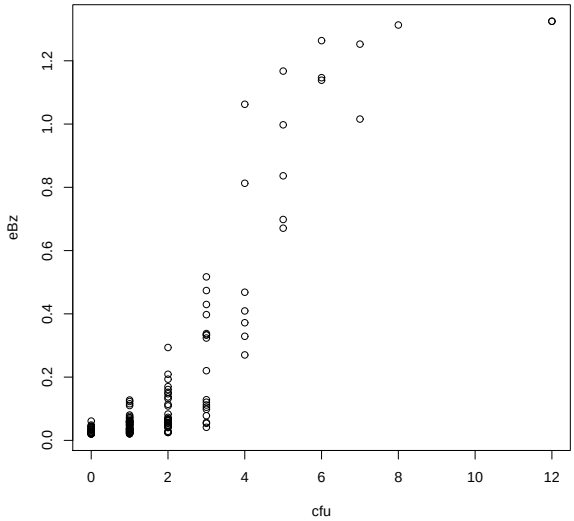

Abbildung 6.7: Posteriori Means der Zufallseffekte z_i gegen die Responses y_i.

Anhang A

Gauß-Hermite-Quadratur

Allgemein ist die Gauß-Quadratur eine Methode, um Integrale der Form

$$\int_a^b f(z)\omega(z)dz$$

durch endliche Summen zu approximieren, wobei $a < b$ und ω eine positive Gewichtsfunktion ist.

Im Speziellen werden wir später an der sogenannten Gauß-Hermite-Quadratur interessiert sein. Dabei ist das Intervall, über das integriert wird, gleich $(-\infty, \infty)$ und die Gewichtsfunktion $\omega(z) = e^{-z^2}$,

$$\int_{-\infty}^{\infty} f(z)\exp(-z^2)dz \approx \sum_{k=1}^{K} \pi_k f(z_k)\,. \tag{A.1}$$

Für die Approximation ist also die Berechnung von k Massepunkten z_k und zugehörigen Gewichten π_k notwendig. Dabei sind die Massepunkte die Nullstellen des Hermitepolynoms K-ten Grades.

Zur Erinnerung lassen sich die Hermite-Polynome $H_k(z)$ in einfacher Weise sukzessive aus den Ableitungen der Funktion

$$\omega(z) = e^{-z^2}$$

erzeugen:

$$\begin{aligned}
\omega'(z) &= -2ze^{-z^2} \\
\omega''(z) &= (4z^2 - 2)e^{-z^2} \\
&\vdots \\
\omega^{(k)}(z) &= (-1)^k H_k(z)e^{-z^2}\,.
\end{aligned}$$

Prinzipiell gibt es zwei unterschiedliche Definitionen der Hermitepolynome: Einmal bezüglich der Gewichtsfunktion $\omega(z) = e^{-z^2}$ (physikalische Hermitepolynome, H_n) und einmal bezüglich der Gewichtsfunktion $\tilde{\omega}(z) = e^{-z^2/2}$ (probabilistische Hermitepolynome, H_{e_n}).

Da in weiterer Folge über die Normalverteilungsdichte integriert wird, sind in unserem Fall insbesondere die probabilistischen Hermitepolynome von Interesse. Allerdings wird bei den üblichen R-Funktionen — `gauss.quad()` aus Smyth (2013) bzw. `gqz()` aus Einbeck et al. (2012) — mit den physikalischen Hermitepolynomen gerechnet. Dies ist jedoch sofort in den Griff zu bekommen, da ein einfacher Zusammenhang zwischen den beiden Polynomtypen (Abramowitz und Stegun, 1964) besteht,

$$\begin{aligned} H_{e_n}(z) &= 2^{-\frac{n}{2}} H_n\left(\frac{z}{\sqrt{2}}\right), \\ H_n(z) &= 2^{\frac{n}{2}} H_{e_n}\left(\sqrt{2}z\right). \end{aligned}$$

Damit ergibt sich zusammen mit (A.1) und der Definition der Dichte der Standardnormalverteilung $\phi(\cdot)$

$$\int_{-\infty}^{\infty} f(z)\phi(z)dz = \int_{-\infty}^{\infty} f(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \approx \sum_{k=1}^{K}\frac{\sqrt{2}\pi_k}{\sqrt{2\pi}} f\left(\sqrt{2}z_k\right).$$

und weiter die Transformation für die neuen Massepunkte z'_k und Gewichte π'_k

$$\tilde{z}_k = \sqrt{2}z_k \qquad \text{und} \quad \tilde{\pi}_k = \frac{\pi_k}{\sqrt{\pi}}.$$

Die Funktion `gqz()` berechnet Massepunkte und Gewichte für Integrale, die mit der Normalverteilungsdichte gewichtet werden. Wie gerade beschrieben, werden zunächst die Massepunkte und Gewichte aus der Gauß-Hermite-Quadratur benötigt. Im Paket `statmod` von Gordon Smyth steht für deren Berechnung die Funktion `gauss.quad()` zur Verfügung. Genau diese Funktion wird auch intern beim Aufruf von `gqz()` benutzt. Um die gewünschten Massepunkte und Gewichte zu erhalten, muss bei `gauss.quad()` als Argument `kind = hermite` übergeben werden. Die Funktion `gqz()` adaptiert im Weiteren die Massepunkte und Gewichte derart, dass sie direkt auf die konkrete Situation angewandt werden können. Die Funktion `gqz()` hat zwei Argumente: `numnodes` und `minweight`. Mittles `numnodes` wird die gewünschte Anzahl K an Massepunkten festgelegt. Über `minweight` wird ein minimales Gewicht zu jedem Massepunkt verlangt. Massepunkte, deren Gewicht kleiner als `minweight` sind, werden ignoriert. Die Default-Einstellungen der Funktion `gqz()` sind

```
> gqz(numnodes=20, minweight=0.000001).
```

Als Output liefert die Funktion eine Liste, welche die Masspunkte z_k und die zugehörigen Gewichte π_k enthält. Das folgende Beispiel demonstriert die Funktionsweise von `gqz()`.

```
> gqz(numnodes=6,minweight=1e-3)
  location   weight
1   3.3243 0.002556
2   1.8892 0.088616
```

```
3   0.6167 0.408828
4  -0.6167 0.408828
5  -1.8892 0.088616
6  -3.3243 0.002556

> gqz(numnodes=6,minweight=1e-2)
  location  weight
2   1.8892 0.08862
3   0.6167 0.40883
4  -0.6167 0.40883
5  -1.8892 0.08862
```

Eine Übersicht über die Lage der Massepunkte in Abhängigkeit der Anzahl an Massepunkten gibt die nächste Abbildung. Man sieht deutlich, dass die Punkte symmetrisch um Null angeordnet sind und dass mit zunehmender Distanz zum Nullpunkt die Distanz zum Nachbarpunkt leicht zunimmt. Man sieht außerdem, dass Punkte mit zu geringem Gewicht gar nicht betrachtet werden und somit die Gesamtanzahl von Punkten ziemlich beschränkt bleibt.

```
> maxK <- 20
> plot(c(-6,6), c(2,maxK), type="n", xlab="quadrature points", ylab="K")
> for(K in 2:maxK){
+  z <- gqz(K)[, 1]
+  points(z, rep(K,length(z)))
+  lines(c(-6,6), c(K,K), lty=3)
+ }
> abline(v=0, lty=2)
```

Als numerisches Beispiel soll zuerst für $X \sim \text{Normal}(0,1)$ der Erwartungswert $\text{E}(X^4) = 3 = \int x^4 \phi(x) dx \approx \sum_{k=1}^{K} z_k^4 \pi_k$ approximiert werden. Für Polynome, deren Grad maximal $2K-1$ ist, ist der Approximationsfehler der Gauß-Hermite Quadratur Null.

```
> z  <- gqz(2)[, 1];  pi <- gqz(2)[, 2];  sum(z^4*pi)
[1] 1
> z  <- gqz(3)[, 1];  pi <- gqz(3)[, 2];  sum(z^4*pi)
[1] 3
> z  <- gqz(6)[, 1];  pi <- gqz(6)[, 2];  sum(z^4*pi)
[1] 3
```

Im zweiten Beispiel soll $\int \exp(y)\phi(y)dy \approx \sum_{k=1}^{K} \exp(z_k)\pi_k$ betrachtet werden. Dies ist gerade dann relevant, wenn $Y \sim \text{Normal}(\mu, \sigma^2)$ und wir am Erwartungswert von $\exp(Y)$ interessiert sind. Wie wir bereits wissen ist für eine Lognormalverteilung der Erwartungswert $\text{E}(\exp(Y)) = \exp(\mu + (\sigma^2/2))$. Falls noch $Y \sim \text{Normal}(0,1)$ gilt, dann folgt daher $\text{E}(\exp(Y)) = \exp(1/2)$.

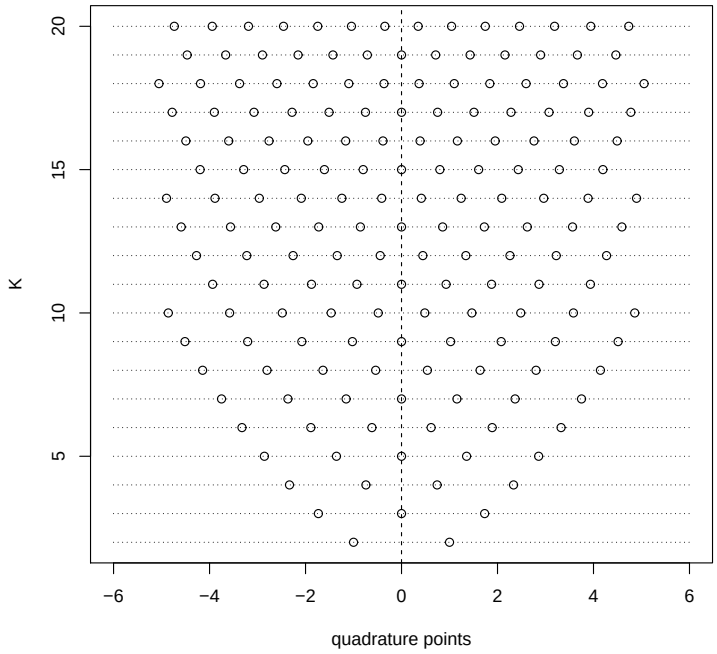

Abbildung A.1: Lage der Quadratur Punkte für verschiedene Werte von K.

```
> z  <- gqz(2)[, 1];  pi <- gqz(2)[, 2];  sum(exp(z)*pi)
[1] 1.54308063482
> z  <- gqz(3)[, 1];  pi <- gqz(3)[, 2];  sum(exp(z)*pi)
[1] 1.63819248006
> z  <- gqz(6)[, 1];  pi <- gqz(6)[, 2];  sum(exp(z)*pi)
[1] 1.64871936647
> exp(1/2)
[1] 1.6487212707
```