Christine TERWUL, Bakk. techn.

Statistische Modelle für Anzahlen
Praktische Beispiele aus dem Bereich der Mikrobiologie

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

Masterstudium Technische Mathematik: Operations Research und Statistik

eingereicht an der

Technischen Universität Graz

Betreuer:
Ao. Univ.-Prof. Dipl.-Ing. Dr. Herwig FRIEDL

Institut für Statistik

Graz, Mai 2014

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

Zusammenfassung

Die Beschreibung von Zählvariablen bzw. Anzahlen durch statistische Modelle findet in der Praxis eine Vielzahl von Anwendungsbereichen, setzt aber für eine korrekte Modellfindung ein fundiertes Wissen über die theoretischen Mittel und Fallstricke voraus. Diese Arbeit beschreibt ausführlich den theoretischen Hintergrund der *generalisierten linearen Modelle* (kurz: *GLM*), welche von großer Bedeutung für die Analyse und Modellierung von Anzahlen sind. Besonderes Augenmerk wird auf das Phänomen der Überdispersion (Variabilität in den Daten übersteigt den Erwartungswert) gelegt, welches ein häufig auftretendes Problem darstellt. Aus diesem Gesichtspunkt heraus führt die theoretische Reise vom GLM mit zugrunde liegender Exponentialfamilie über die Quasi-Likelihood- bis hin zur Extended Quasi-Likelihood-Funktion. Die Basis dazu bildet die Arbeit von McCullagh und Nelder (1989), welche die Klasse der GLMs erstmals vorgestellt haben. Den Abschluss bildet die Umsetzung des erworbenen Wissens anhand mikrobiologischer Beispiele im Rahmen von Studien über Schimmelpilzkonzentrationen in der Luft von steirischen Weinkellern und Grazer Wohnungen. Alle Analysen werden hierbei mit der freien Statistiksoftware *R* durchgeführt.

Abstract

The description of counts with help of statistical models has in practice a wide range of application but presumes consolidated knowledge about theoretical methods and drawbacks for reaching a correct model. This master thesis examines in detail the theoretical background of *generalized linear models* (short: *GLM*) which are of particular importance for analysis and modelling of counts. Special attention is paid to so-called overdisperion (variability of data exceeds the mean) which is a common problem with this kind of data. Bearing this phenomenon in mind leads to a theoretical journey from GLM assuming an underlying exponential family to Quasi Likelihood until the Extended Quasi Likelihood function. Literary basis for these concepts is to be found in McCullagh und Nelder (1989) who first presented the class of GLMs. Finally the theoretical knowledge is applied to microbiological data from surveys concerning concentrations of mould spores in the ambient air of wine cellars in Styria and flats in the city center of Graz. All analyses are performed using the free statistics software *R*.

Inhaltsverzeichnis

1 Einleitung

Im Rahmen dieser Arbeit werden Möglichkeiten zur statistischen Modellierung von Zählvariablen, sogenannten Anzahlen oder Counts, mit Hilfe der Regressionsanalyse ausführlich beschrieben. Zählvariablen haben ihren Wertebereich in den nicht negativen ganzen Zahlen und können zum Beispiel die Anzahl von Keimen zu einem bestimmten Messzeitpunkt oder die Menge an Blitzen an einem bestimmten Ort beschreiben. Die Problematik bei der Analyse solcher Daten liegt in der Modellierung ihrer meist unbekannten Varianzstruktur.

Die Grundlagen der Regressionsanalyse werden im zweiten Kapitel vorgestellt. Es behandelt das klassische lineare Modell und mögliche Verfahren zur Parameterschätzung. Von besonderer Wichtigkeit wird sich in den darauffolgenden Kapiteln der Maximum Likelihood-Ansatz zeigen. Klassische lineare Modelle setzen normalverteilte und varianzhomogene Responsevariablen voraus und sind somit zur Modellierung von Anzahlen wenig geeignet.

Dieser Missstand führt über zum nächsten Kapitel, das eine wichtige Verallgemeinerung des klassischen linearen Modells zum Thema hat, nämlich die Klasse der generalisierten linearen Modelle. Bei dieser Modellklasse wird die Voraussetzung der normalverteilten Responses relaxiert und auf die weitaus umfassendere Exponentialfamilie ausgedehnt. Diese Erweiterung erlaubt nicht nur die Annahme erwartungswertabhängiger Varianzfunktionen, sondern lässt die Linkfunktion zwischen zufälliger (Response) und systematischer Komponente (Kovariablen) auch von nichtlinearer Natur sein. Zur Schätzung der Modellparameter nach dem Maximum Likelihood-Prinzip wird ein iterativer gewichteter Kleinste Quadrate Algorithmus herangezogen.

Im vierten Kapitel wird näher auf Modellklassen speziell zur Beschreibung von Zählvariablen eingegangen. Der Standard-Ansatz ist die Annahme einer Poissonverteilung, welcher sich aber aufgrund der vorausgesetzten Equidispersion (Übereinstimmung von Erwartungswert und Varianz) manchmal als Problem herausstellen kann. Übersteigt die Variabilität in den Daten den Erwartungswert, so spricht man von Überdispersion, was bei Nichtberücksichtigung zu unterschätzten Standardfehlern und somit zu falschen Inferenzaussagen führen kann. Diesem Phänomen wird man durch Wahl einer breiter streuenden Verteilungsannahme habhaft. Die Negativ-Binomial-Verteilung stellt eine gute Alternative dar, da ihre Varianzstruktur extra-Poisson Variabilität widerspiegelt. Kann keine vollständige Verteilungsannahme getroffen werden, so ist der Quasi-Likelihood-Ansatz ein wertvolles Werkzeug. Für diese Methode ist die Spezifikation der Erwartungswert-Varianz-Relation ausreichend, ein multiplikativer Dispersionsparameter wird aus den Daten geschätzt. Quasi-Likelihood-Schätzer haben ähnlich nützliche Eigenschaften

wie ihre log-Likelihood-Pendants und lassen sich ebenso durch einen iterativen Algorithmus berechnen. Gibt es zum angenommenen Erwartungswert-Varianz-Verhältnis einen Vertreter aus der Exponentialfamilie mit gleichen Eigenschaften der ersten beiden Momente, so stimmen Quasi-Likelihood- und log-Likelihood-Funktion sogar überein. Eine weitere Lockerung der Modellvoraussetzungen bietet die Extended Quasi-Likelihood-Funktion. Sie erlaubt die Schätzung von Varianzfunktionen aus parametrisierten Familien und vektorwertigen Dispersionsparametern.

Das fünfte Kapitel beinhaltet die praktische Umsetzung der in den vorangegangenen Kapiteln besprochenen Theorie. Es werden mikrobiologische Daten über Schimmelpilzkonzentrationen in Weinkellern und Wohnräumen betrachtet und alle vorgestellten Ansätze miteinander verglichen. Die praktische Ausführung der Modellschätzung geschieht hierbei mit der freien Statistiksoftware *R*. Alle relevanten Funktionen und Bibliotheken werden im Zuge der Ausarbeitung kurz vorgestellt.

2 Das klassische lineare Modell

Das klassische lineare Modell ist ein einfaches Werkzeug den Einfluss von mehreren erklärenden Variablen $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$, sogenannten Kovariablen oder Prädiktoren, auf eine interessierende Variable $\mathbf{y}$, die Response-Variable, zu beschreiben. Wie man schon aus der Bedeutung des Wortes *Statistik* erahnen kann, als *Wissenschaft von der zahlenmäßigen Erfassung, Untersuchung und Auswertung von Massenerscheinungen*[1], spielen lineare Modelle eine wesentliche Rolle. Einen detailierten Einblick hierzu geben Christensen (2011), Fahrmeir, Kneib und Lang (2009) und Casella und Berger (2001).

Im Allgemeinen betrachten wir die Abhängigkeit

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad \text{für } i = 1, \ldots, n$$

oder in Matrixschreibweise $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, wobei

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Hierbei sind y_i Zufallsvariablen und x_{ij} gegebene Beobachtungen (deterministischer Natur) zusammengefasst zur Designmatrix $\mathbf{X}$, wobei $\mathbf{x}_i^T$ $(i = 1, \ldots, n)$ Zeilenvektoren und $\mathbf{x}_j$ $(j = 1, \ldots, p-1)$ Spaltenvektoren darstellen. Der Parameter β_0 wird als Intercept bezeichnet, $\beta_1, \ldots, \beta_{p-1}$ als Slope-Parameter und ϵ_i beschreiben den zufälligen Modellfehler. Weiters wird angenommen, dass

- $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$,
- $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$,
- $\mathrm{rg}(\mathbf{X}) = p$, d.h. die Designmatrix hat vollen Rang,
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Hieraus ergibt sich folgende (sehr restriktive) Eigenschaft des Responsevektors $\mathbf{y}$:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \text{ mit} \qquad \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \text{ und } \mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}.$$

[1] http://www.duden.de/node/649301/revisions/649306/view

2.1 Kleinste Quadrate Schätzung

Ein Ansatz zur Schätzung des Modellparametervektors $\boldsymbol{\beta}$ ist die Kleinste Quadrate Schätzung, welche ohne spezifische Verteilungsannahmen über $\mathbf{y}$ auskommt. Die erstmalige Erwähnung dieses Ansatzes geht auf Gauß und Legendre im frühen 19. Jahrhundert zurück (Stigler, 1981).

Die Methode basiert auf der Minimierung des Modellfehlers $\boldsymbol{\epsilon}$, genauer gesagt $\min_{\boldsymbol{\beta}} \sum_i \epsilon_i^2$, es wird also $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$ über $\boldsymbol{\beta}$ minimiert. Die Minimierung führt zu einer eindeutigen Lösung $\hat{\boldsymbol{\beta}}$, wenn die Designmatrix vollen Rang p hat (für den algebraischen Beweis und weitere Details wird auf Seber und Lee, 2003, verwiesen). Hierzu betrachtet man das System der *Normalgleichungen*

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}, \tag{2.1}$$

welche zur Lösung

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.2}$$

führen.

Zur Nomenklatur: Der Vektor $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbb{E}}(\mathbf{y}) = \hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)^T$ beinhaltet die sogenannten *fitted values* und $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}$ wird als *Residuenvektor* bezeichnet.

Eigenschaften des Schätzers $\hat{\boldsymbol{\beta}}$:

- $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$,
- Die Statistik $SSE(\hat{\boldsymbol{\beta}}) = \mathbf{r}^T\mathbf{r} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$ wird *residual sum of squares* genannt. In skalierter Form ist sie ein erwartungstreuer Schätzer für σ^2, d.h. für $S^2 = \dfrac{SSE(\hat{\boldsymbol{\beta}})}{n-p}$ gilt $\mathbb{E}(S^2) = \sigma^2$.

2.2 Maximum Likelihood Schätzung

Eine weitere Möglichkeit der Parameterschätzung ist die Maximum Likelihood Methode. Hierbei fließen auch die Verteilungseigenschaften der Fehlerterme $\boldsymbol{\epsilon}$ und somit der Response $\mathbf{y}$ ein, was statistische Eigenschaften der Schätzer bestimmen lässt und damit Hypothesentests und Konfidenzintervalle definierbar macht. Die Maximum Likelihood Methode wurde 1922 von R. A. Fisher eingeführt (siehe Fisher, 1997 (digitalisierte Version) bzw. Aldrich, 1997).

Im Fall der Normalverteilung des Responsevektors $\mathbf{y}$ (y_i unabhängig und homoskedastisch) sieht die Likelihood-Funktion wie folgt aus (vgl. SEBER UND LEE, 2003):

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}, \tag{2.3}$$

bzw. mit $l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \log L(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ und Weglassen der Konstanten

$$l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{2.4}$$

Daraus ergeben sich folgende zur Maximierung notwendigen Ableitungen

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2|\mathbf{y})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) \tag{2.5}$$

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2|\mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{2.6}$$

Durch Nullsetzen erhält man dann einerseits für $\boldsymbol{\beta}$ wiederum den Kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ und als Varianzschätzer

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}SSE(\hat{\boldsymbol{\beta}}). \tag{2.7}$$

Eigenschaften des MLE $\hat{\boldsymbol{\beta}}$:

- $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$,
- $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$,
- $\hat{\boldsymbol{\beta}}$ ist unabhängig von $\hat{\sigma}^2$,
- $SSE(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$.

2.3 Box-Cox Transformation

Eine wesentliche Annahme des oben beschriebenen Modells, nämlich die Varianzhomogenität der Responses y_i, ist bei realen Daten nicht immer haltbar. In diesem Fall besteht die Möglichkeit einer varianzstabilisierenden Transformation der Responses nach BOX UND COX (1964), was dann eine Anwendung des klassischen linearen Modells bzw. einer Varianzanalyse wieder möglich macht.

Im Allgemeinen handelt es sich um eine Transformation $y \to y(\lambda)$ der Form

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

für positive Responses $y > 0$ und

$$y(\lambda) = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, & \lambda_1 \neq 0 \\ \log(y + \lambda_2), & \lambda_1 = 0 \end{cases}$$

für Responses $y > -\lambda_2$.

Der optimale Wert des Transformations-Parameters (bzw. Parametervektors) λ wird durch simultane Maximierung der Likelihood-Funktion bzgl. λ und sämtlicher Modellparameter $(\boldsymbol{\beta}, \sigma^2)$ ermittelt.

3 Generalisierte lineare Modelle

Da nicht nur die Annahme der einheitlichen Varianz der Responses in der Praxis schwer zu halten ist, sondern auch die Normalverteilung in den meisten Fällen nicht vorliegt, wird im Folgenden eine umfassendere Klasse der linearen Modelle vorgestellt (vgl. McCullagh und Nelder, 1989): die *generalisierten linearen Modelle (GLM).*

Eingeführt wurde der Begriff des "generalisierten linearen Modells" erstmals von Nelder und Wedderburn (1972). Zur Einleitung wird das klassische lineare Modell etwas umformuliert:

- Zufällige Komponente: Die unabhängig normalverteilte Response $\mathbf{y}$ hat konstante Varianz σ^2 und $\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu}$.
- Systematische Komponente: Die Kovariablen $\mathbf{x}_0, ..., \mathbf{x}_{p-1}$ bilden den linearen Prädiktor $\boldsymbol{\eta}$ mit
$$\boldsymbol{\eta} = \sum_{j=0}^{p-1} \beta_j \mathbf{x}_j.$$
- Der *Link* zwischen zufälliger und systematischer Komponente ist
$$\eta_i = g(\mu_i)$$
bzw.
$$\mathbb{E}(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$
mit $g() = id()$ die identische Abbildung, wobei $g()$ als *Linkfunktion* bezeichnet wird.

Die Klasse der GLM erlaubt nun eine Erweiterung der normalverteilten Responses auf eine beliebige Verteilung aus der Exponentialfamilie und die freie Wahl der Linkfunktion (einzige Voraussetzung: Monotonie und Differenzierbarkeit).

3.1 Die Exponentialfamilie

Definition 3.1. *Die Dichte- oder Wahrscheinlichkeitsfunktion einer Zufallsvariable Y mit*

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{3.1}$$

mit festen Funktionen $a()$, $b()$, $c()$, $a(\phi) > 0$ und bekanntem ϕ gehört zur einparametrigen linearen ***Exponentialfamilie mit kanonischem Parameter*** *θ.*

Mitglieder der Exponentialfamilie sind zum Beispiel die Normalverteilung, die Poissonverteilung, die Gammaverteilung, die Invers-Gaussverteilung und die Binomialverteilung.

3.1.1 Eigenschaften der Exponentialfamilie

- Für die Ableitung der log-Likelihood-Funktion, auch *Scorefunktion* genannt, gelten die folgenden Eigenschaften der ersten beiden Momente (vgl. Casella und Berger, 2001):

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\log f(y|\theta,\phi)\right) = 0, \tag{3.2}$$

$$-\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(y|\theta,\phi)\right) = \mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log f(y|\theta,\phi)\right)^2\right). \tag{3.3}$$

- Weiters erhält man mit den obigen Identitäten

$$\begin{aligned}\mathbb{E}(Y) &= b'(\theta) = \mu,\\ \mathrm{Var}(Y) &= b''(\theta)a(\phi) = V(\mu)a(\phi).\end{aligned} \tag{3.4}$$

  Die Funktion $V(\mu)$ wird *Varianzfunktion* und die Größe ϕ *Dispersionsparameter* genannt. Der Term $a(\phi)$ ist meist von der Form

$$a(\phi) = a \cdot \phi,$$

  wobei ϕ konstant über die Beobachtungen ist und a als Gewicht auch von Beobachtung zu Beobachtung variieren kann (also $a_i(\phi) = a_i \cdot \phi$).

- Der Maximum Likelihood Schätzer für $\boldsymbol{\mu}$ bzw. $\boldsymbol{\beta}$ maximiert die log-Likelihood Funktion der Stichprobe (vgl. MYERS, MONTGOMERY, VINING UND ROBINSON, 2010):

$$l(\boldsymbol{\beta}|\mathbf{y}) = \log L(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}. \tag{3.5}$$

Für $\eta_i = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \theta_i$, dem sogenannten *kanonischen Link*, gilt somit

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} &= \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \right) \mathbf{x}_i \\ &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i)\mathbf{x}_i. \end{aligned} \tag{3.6}$$

Den MLE für $\boldsymbol{\beta}$ erhält man nun durch Lösen der Scoregleichungen. Falls die $a_i(\phi)$ konstant sind, vereinfacht sich dies zu

$$\sum_{i=1}^{n} (y_i - \mu_i)\mathbf{x}_i = \mathbf{0}. \tag{3.7}$$

Falls $g()$ nicht die kanonische Linkfunktion ist, gilt für die Scorefunktion:

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \tag{3.8}$$

und weiters mit $\dfrac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i$

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} \frac{\partial \theta_i}{\partial \eta_i} \mathbf{x}_i. \tag{3.9}$$

Mit $\dfrac{\partial \theta_i}{\partial \eta_i} = \dfrac{\partial \theta_i}{\partial \mu_i} \dfrac{\partial \mu_i}{\partial \eta_i}$, $b'(\theta_i) = \mu_i$ und wegen

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i) \tag{3.10}$$

vereinfacht sich die Scorefunktion (3.9) zu

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i$$

$$= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(y_i)} \frac{\partial \mu_i}{\partial g(\mu_i)} \mathbf{x}_i$$
$$= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(y_i)} \frac{\mathbf{x}_i}{g'(\mu_i)}. \tag{3.11}$$

- Eigenschaften des MLE: Der MLE ist asymptotisch erwartungstreu, d.h.

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\beta}.$$

Weiters haben Fahrmeir und Kaufmann (1985) gezeigt, dass der Schätzer sogar asymptotisch normalverteilt ist:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \quad \xrightarrow{d} \quad N_p\left(\mathbf{0}, n\mathbf{I}(\boldsymbol{\beta})^{-1}\right), \tag{3.12}$$

wobei $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$, mit

$$\mathbf{W} = \mathrm{diag}\left(\frac{1}{a_i(\phi)V(\mu_i)(g'(\mu_i))^2}\right), \tag{3.13}$$

die sogenannte *Fisher Informationsmatrix*, und dem Erwartungswert der negativen Hessematrix ausgewertet im wahren Parameter $\boldsymbol{\beta}$ entspricht. Bei vorliegender Poissonverteilung der Responses und Verwendung des log-Links $g(\boldsymbol{\mu}) = \log \boldsymbol{\mu}$ vereinfacht sich die Matrix $\mathbf{W}$ beispielsweise zu

$$\mathbf{W} = \mathrm{diag}(\mu_1, \ldots, \mu_n). \tag{3.14}$$

3.1.2 Einige Mitglieder der Exponentialfamilie im Überblick

	Normal $N(\mu, \sigma^2)$	Poisson $P(\mu)$	Gamma $G(\mu, \nu)$
Range von y	$(-\infty, \infty)$	$0, 1, \ldots, \infty$	$(0, \infty)$
ϕ	σ^2	1	ν^{-1}
$a(\phi)$	ϕ	ϕ	ϕ
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$-\log(-\theta)$
$c(y, \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y!$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$
$\mu(\theta) = b'(\theta)$	θ	$\exp(\theta)$	$-1/\theta$
Kanonischer Link	identischer Link $g(\mu) = \mu$	log-Link $g(\mu) = \log \mu$	reziproker Link $g(\mu) = \mu^{-1}$
$V(\mu)$	1	μ	μ^2

3.2 Iteratively (re)weighted least squares

Aufgrund der Nichtlinearität der Scoregleichungen kann der MLE für $\boldsymbol{\beta}$ nur iterativ berechnet werden, eine Methode ist die *iteratively (re)weighted least squares Methode (IRLS)*. Sie ist eine Variation der Newton-Raphson Methode zur Findung von Nullstellen und wurde erstmals erwähnt von Fisher (1935). Die folgenden Details zur Iterationsvorschrift basieren auf Green (1984).

Nach einer Reparametrisierung der Scoregleichungen

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{u} = \mathbf{0}, \tag{3.15}$$

wobei $\mathbf{u}$ den $n \times 1$ Vektor $\dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta}} = \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu})$ (wobei im Falle eines kanonischen Links $\boldsymbol{\Delta} = \mathbb{I}_n$ gilt) und $\mathbf{X}$ die $n \times p$ Matrix $\dfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}$ darstellt, liefert die Newton-Raphson Methode folgende Iterationsvorschrift zur Lösung des Gleichungssystems

$$\left(-\frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) = \mathbf{X}^T\mathbf{u} \tag{3.16}$$

bzw.

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \mathbf{X}^T \mathbf{u}. \tag{3.17}$$

Die negative Hessematrix der log-Likelihood Funktion kann nun geschrieben werden als

$$\left(-\frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) = -\sum_{i=1}^{n} \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta}_i} \frac{\partial^2 \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right)^T \frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right). \tag{3.18}$$

Die Terme auf der rechten Seite von (3.18) werden nun durch ihre Erwartungen ersetzt (an der Stelle des aktuellen Wertes von $\boldsymbol{\eta}$), nämlich

$$\mathbb{E}\left(\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \eta_i} \right) = 0$$

$$\mathbb{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) = \mathbb{E}\left(\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta}} \left(\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\eta}} \right)^T \right) =: \mathbf{A}.$$

Somit resultiert als Iterationsvorschrift

$$(\mathbf{X}^T \mathbf{A} \mathbf{X})(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) = \mathbf{X}^T \mathbf{u}, \tag{3.19}$$

wobei t die aktuelle Iterationszahl beschreibt. Diese Iterationstechnik wird auch als *Fisher Scoring Technik* bezeichnet.

Da die Designmatrix $\mathbf{X}$ grundsätzlich als Matrix mit vollem Rang p angenommen wird und $\mathbf{A}$ eine positiv definite Diagonalmatrix ist, beschreibt obige Vorschrift somit ein nicht-singuläres $p \times p$ Gleichungssystem für $\boldsymbol{\beta}^{(t+1)}$. Bei Konvergenz des Algorithmus resultiert als Grenzwert der Schätzer $\hat{\boldsymbol{\beta}}$.

3.3 Güte der Modellanpassung

Hat man ein Modell gefunden, stellt sich die Frage nach der Güte desselben. Grundsätzlich gibt es zwei Extreme. Seien n Beobachtungen gegeben, können Modelle mit bis zu n Parametern gefittet werden. Das maximale Modell, bei dem jedes μ_i frei wählbar ist und somit der MLE $\hat{\mu}_i$ gleich der Beobachtung y_i ist, wird als *saturiertes* oder *volles Modell* bezeichnet. Es hat eine perfekte Datenanpassung, lässt aber keinerlei Rückschlüsse auf aussagekräftige Abhängigkeiten der Response von den Prädiktoren zu. Im Gegenteil dazu stellt das minimalste Modell das *Null Modell* dar, welches jedem y_i den *overall mean* μ als Erwartungswert zuordnet und somit die Responses $y_1, \ldots, y_n$ als Zufallsstichprobe modelliert.
Beide Varianten treffen im Allgemeinen nicht die Erwartungen an ein optimales GLM, jedoch dient das saturierte Modell als Basis für die Bewertung von p-parametrigen Modellen.

3.3.1 Deviance

Die Deviance ist ein Maß, das auf dem log-Likelihood Quotiententest basiert (vgl. McCullagh und Nelder, 1989). Betrachtet man die log-Likelihood Funktion in Termen von $\boldsymbol{\mu}$ und nicht von $\boldsymbol{\theta}$, so bezeichnet $l(\hat{\boldsymbol{\mu}}, \phi|\mathbf{y})$ die log-Likelihood Funktion maximiert nach $\boldsymbol{\beta}$ unter dem p-Parameter-Modell. Sei weiters $l(\mathbf{y}, \phi|\mathbf{y})$ die maximale log-Likelihood Funktion des saturierten Modells und somit ein fester Wert gegeben $\mathbf{y}$, so kann die Güte der Modellanpassung als proportional zur doppelten Differenz der log-Likelihood Funktionen angesehen werden.

Definition 3.2. *Sei $a_i(\phi) = \phi \cdot a_i$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$ und $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$ dann wird*

$$\begin{aligned}\frac{1}{\phi} D(\hat{\boldsymbol{\mu}}|\mathbf{y}) &= -2\left(\log L(\hat{\boldsymbol{\mu}}, \phi|\mathbf{y}) - \log L(\mathbf{y}, \phi|\mathbf{y})\right) \\ &= -2\sum_{i=1}^{n} \frac{1}{a_i}\left(y_i(\hat{\theta}_i - \tilde{\theta}_i) - \left(b(\hat{\theta}_i) - b(\tilde{\theta}_i)\right)\right)/\phi \qquad (3.20)\end{aligned}$$

*als **skalierte Deviance** des aktuellen p-parametrigen Modells bezeichnet.*

Grundsätzlich gilt also, dass einfachere Modelle (d.h. Modelle mit weniger Parameter) eine größere Deviance haben. Im Falle der Normalverteilung entspricht die Deviance genau der minimalen Fehlerquadratsumme und ist eine χ^2_{n-p}-verteilte Größe mit Erwartungswert $n - p$.
In Spezialfällen kann die Deviance weiter vereinfacht werden, was folgender Satz beschreibt.

Satz 3.1 (Nelder und Wedderburn, 1972). *Ist die Linkfunktion des generalisierten linearen Modells (a) $\boldsymbol{\eta} = \boldsymbol{\mu}^\alpha$ oder (b) $\boldsymbol{\eta} = \log \boldsymbol{\mu}$ und beinhaltet das Modell einen Intercept β_0, so gilt*

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)\hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)} = 0, \qquad (3.21)$$

wobei $\hat{\boldsymbol{\mu}}$ den MLE von $\boldsymbol{\mu}$ bezeichnet und $\alpha \in \mathbb{R}$.

Beweis. O.b.d.A. sei $a_i(\phi)$ fest und unabhängig von μ_i. Bezeichne $\hat{\mu}_i$ den MLE von μ_i unter dem betrachteten Modell.

(b): Sei $\eta_i = g(\mu_i) = \log(\mu_i)$, dann gilt für die Scorefunktion

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_0} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \frac{x_{i0}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \frac{1}{1/\mu_i} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mu_i}{a_i(\phi)V(\mu_i)} = 0 \quad \text{für } \mu_i = \hat{\mu}_i.\end{aligned}$$

(a): Sei nun $\eta_i = g(\mu_i) = \mu_i^\alpha$, dann gilt wiederum für die Scorefunktion

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}, \quad j = 0, \ldots, p-1 \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \frac{x_{ij}}{\alpha \mu_i^{\alpha-1}} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mu_i}{a_i(\phi)V(\mu_i)} \frac{x_{ij}}{\alpha \mu_i^{\alpha}} = 0 \quad \text{für } \mu_i = \hat{\mu}_i.\end{aligned}$$

Mittels Umformulierung des nachstehenden Terms folgt die Behauptung, dass

$$\begin{aligned}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)\hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)} &= \sum_{i=1}^{n} \frac{\alpha \hat{\mu}_i^\alpha (y_i - \hat{\mu}_i)\hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)\alpha\hat{\mu}_i^\alpha} \\ [\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}] &= \sum_{i=1}^{n}\sum_{j=0}^{p-1} \alpha x_{ij}\hat{\beta}_j \frac{(y_i - \hat{\mu}_i)\hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)\alpha\hat{\mu}_i^\alpha} \\ &= \sum_{j=0}^{p-1} \alpha\hat{\beta}_j \underbrace{\left\{ \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)\hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)} \frac{x_{ij}}{\alpha\hat{\mu}_i^\alpha} \right\}}_{=0} = 0.\end{aligned}$$

□

Mit diesem Wissen kann die Deviance von den log-linearen Modellen der späteren Kapitel vereinfacht werden. Insbesondere ergibt sich bei Einhalten der im Satz genannten Voraussetzungen für die Poisson-Verteilung

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0 \tag{3.22}$$

und für die Gamma-Verteilung

$$\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} = 0. \tag{3.23}$$

3.3.2 Pearson-Statistik

Ein weiteres Gütemaß ist die sogenannte *Pearson-Statistik*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}, \tag{3.24}$$

wobei $a_i(\phi) = \phi \cdot a_i$ gilt.
Auch hier ist im Normalverteilungsfall X^2 gleich der minimierten Fehlerquadratsumme. Mit Hilfe der Pearson-Statistik kann auch das Problem der Schätzung des möglich unbekannten Dispersionsparameters ϕ gelöst werden:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)} = \frac{1}{n-p} X^2 \tag{3.25}$$

3.3.3 Parametertest

Betrachtet man zwei ineinandergeschachtelte Modelle (*nested models*), so kann man die überschüssigen Parameter mit Hilfe der Deviance-Differenz auf deren Signifikanz testen. Es werden folgende Hypothesen getestet, wobei $q < p$:

$$\begin{aligned} H_0 &: \boldsymbol{\eta} = \beta_{q+1}\mathbf{x}_{q+1} + \cdots + \beta_p \mathbf{x}_p \\ H_1 &: \boldsymbol{\eta} = \beta_1 \mathbf{x}_1 + \cdots + \beta_q \mathbf{x}_q + \beta_{q+1}\mathbf{x}_{q+1} + \cdots + \beta_p \mathbf{x}_p \end{aligned}$$

bzw.

$$\begin{aligned} H_0 &: \beta_1 = \cdots = \beta_q = 0 \\ H_1 &: \beta_1, \cdots, \beta_q \quad \text{beliebig.} \end{aligned}$$

Seien $\hat{\boldsymbol{\beta}}_0$ und $\hat{\boldsymbol{\beta}}_1$ bzw. $\hat{\boldsymbol{\mu}}_0$ und $\hat{\boldsymbol{\mu}}_1$ die Parameterschätzungen der Modelle unter H_0 bzw. H_1, so beschreibt die Likelihood-Quotienten Teststatistik

$$-2\left(\log L(\hat{\boldsymbol{\beta}}_0|\mathbf{y}) - \log L(\hat{\boldsymbol{\beta}}_1|\mathbf{y})\right) = \frac{1}{\phi}\left(D(\hat{\boldsymbol{\mu}}_0|\mathbf{y}) - D(\hat{\boldsymbol{\mu}}_1|\mathbf{y})\right)$$

eine Deviance-Reduktion. Diese Teststatistik ist unter H_0 und weiteren Regularitätsbedingungen asymptotisch χ^2_q verteilt, sofern der Dispersionsparameter bekannt ist.

Trifft dies nicht zu, also ist ϕ unbekannt, so gilt approximativ

$$\frac{\left(D(\hat{\boldsymbol{\mu}}_0|\mathbf{y}) - D(\hat{\boldsymbol{\mu}}_1|\mathbf{y})\right)/q}{\hat{\sigma}^2} \sim F_{q,n-p},$$

mit $\hat{\sigma}^2 = D(\hat{\boldsymbol{\mu}}_1|\mathbf{y})/(n-p)$.

3.3.4 AIC

Eine weitere Möglichkeit der Modellbewertung ist das AIC (Akaike's Information Criterion), erstmals vorgestellt von Akaike (1973), welches auf dem Konzept des Informationsverlustes basiert. Für eine Übersicht der gängigen Versionen des AIC wird auf Millar (2011) verwiesen.
Grundsätzlich wird die Minderung der Vorhersagekraft des gefitteten Modells gegenüber dem wahren Modell bewertet, wodurch ein minimaler AIC zum präferierten Modell führt. Definiert ist diese Größe als

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p, \tag{3.26}$$

wobei p die Gesamtanzahl aller Parameter im betrachteten Modell beschreibt. Modellselektion mittels AIC hat den Vorteil, dass die betrachteten Modelle nicht ineinander geschachtelt sein müssen, auch verschiedene Linkfunktionen oder Fehlerstrukturen sind erlaubt. Es muss aber darauf geachtet werden, dass die Berechnung der maximalen log-Likelihood Funktion immer auch alle konstanten Terme enthält.

Eine bias-korrigierte Version des AIC ist das sogenannte korrigierte AIC (**AICc**), beschrieben in Hurvich und Tsai (1989). Das ursprüngliche AIC kann vorallem im Fall von kleinen Stichproben zu overfitting führen, was durch Hinzufügen eines sogenannten *Strafterms* asymptotisch verbessert werden kann. Dies führt zu

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(p+1)(p+2)}{n-p-2}. \tag{3.27}$$

Diese Korrektur hat keinen Einfluss auf die Varianz des Schätzers, da der Strafterm von nicht-stochastischer Natur ist. Liegt ein lineares Modell vor, so ist das AICc asymptotisch effizient.

3.4 Residuen

Auch Residuen zählen zu den häufig verwendeten Mitteln, um die Güte eines Modells zu beurteilen. Aus ihnen können unter anderem Rückschlüsse auf die durch das vorliegende Modell erklärte Varianz der betrachteten Daten gezogen werden.

Im Falle normalverteilter Responses können diese dargestellt werden als *Datum = fitted value + Residuum*

$$\mathbf{y} = \hat{\boldsymbol{\mu}} + (\mathbf{y} - \hat{\boldsymbol{\mu}}).$$

Beim GLM entsprechen die **Pearson Residuen** den oben genannten gewöhnlichen Residuen:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i V(\hat{\mu}_i)}}. \tag{3.28}$$

Es gilt, $\sum_{i=1}^{n}(r_i^P)^2 = X^2$.

Eine weitere Variante sind die **Deviance Residuen**, welche den Beitrag der Beobachtung zur Deviance miteinbeziehen, d.h.

$$r_i^D = sign(y_i - \hat{\mu}_i)\sqrt{2\left(l(y_i|y_i) - l(\hat{\mu}_i|y_i)\right)}. \tag{3.29}$$

Es gilt, $\sum_{i=1}^{n}(r_i^D)^2 = D(\hat{\boldsymbol{\mu}}|\mathbf{y})$.

4 Generalisierte lineare Modelle für Anzahlen

Den Hauptteil dieser Arbeit stellt die Spezifikation und Schätzung von Regressionsmodellen für Zählvariablen, sogenannte *Counts* oder zu Deutsch *Anzahlen*, dar. Als Counts werden Variablen bezeichnet, deren Wertebereich aus den nicht negativen ganzen Zahlen besteht, d.h. es werden Modelle mit nach oben unbeschränkten Responses der Form $y = 0, 1, 2, \ldots$ betrachtet. Ein breites Spektrum an Methoden zur Modellierung von Counts wird unter anderem von Winkelmann (2008) und Myers et al. (2010) vorgestellt.

Beispiele für solche Zählvariablen sind die Anzahl von eingehenden Telefonanrufen in einem fixen Zeitintervall oder die Anzahl von Keimen zu einer bestimmten Zeit bzw. an einem bestimmten Ort. Zweiteres Szenario wird ausführlich im Kapitel 5 behandelt.

4.1 Poisson-Modell

Das Poisson-Modell gilt als Standardmodell für Zähl-Daten (ähnlich dem Normalverteilungsmodell für reellwertige stetige Daten), da es den Vorteil mit sich bringt, den nicht negativen Integercharakter der Responses explizit abzubilden.

Zum besseren Verständnis folgt ein kurzer Überblick der wesentlichen Eigenschaften der Poisson-Verteilung.

Definition 4.1. *Sei Y eine Zufallsvariable mit diskreter Verteilung über $\mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$, so folgt Y genau dann einer Poisson-Verteilung mit Parameter μ, $Y \sim Poisson(\mu)$, wenn sich die zugehörige Wahrscheinlichkeitsfunktion darstellen lässt als*

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \qquad \mu \in \mathbb{R}^+,\ y = 0, 1, 2, \ldots.$$

Charakteristisch für die Poisson-Verteilung ist die Übereinstimmung von Erwartungswert und Varianz (und auch aller höheren Kumulanten), nämlich

$$\mathbb{E}(Y) = \mathrm{Var}(Y) = \mu.$$

Diese Eigenschaft wird auch als *Equidispersion* bezeichnet.

4.1.1 Modelleigenschaften

Wie schon im vorigen Kapitel erwähnt, ist der kanonische Link der Poisson-Verteilung der log-Link, also

$$\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \quad \text{für } i = 1, \ldots, n.$$

Die Likelihood-Funktion ergibt sich als

$$L(\boldsymbol{\mu}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp\left(-\sum_{i=1}^{n} \mu_i\right)}{\prod_{i=1}^{n} y_i!} \tag{4.1}$$

$$\log L(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log(y_i!). \tag{4.2}$$

Ignoriert man den für die Maximierung nach $\boldsymbol{\beta}$ konstanten Term, so erhält man

$$\log L(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^{n} \left(y_i \log(\mu_i) - \mu_i\right), \tag{4.3}$$

die sogenannte *Profile log-Likelihood-Funktion.*
Für den resultierenden MLE $\hat{\boldsymbol{\beta}}$ gilt natürlich

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} N_p\left(\boldsymbol{\beta}, \left(\mathbf{X}^T \operatorname{diag}(\mu_1, \ldots, \mu_n)\mathbf{X}\right)^{-1}\right).$$

Per Definition der Poisson-Verteilung ist der Dispersionsparameter $\phi = 1$, somit folgt für die (skalierte) Deviance

$$D(\hat{\boldsymbol{\mu}}|\mathbf{y}) = 2\sum_{i=1}^{n} \left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right). \tag{4.4}$$

Ist im Modell eine Konstante (Intercept) enthalten, hält laut Satz 3.1 die Eigenschaft $\sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0$ und damit ist

$$D(\hat{\boldsymbol{\mu}}|\mathbf{y}) = 2\sum_{i=1}^{n} y_i \log \frac{y_i}{\hat{\mu}_i}. \tag{4.5}$$

Für große $\boldsymbol{\mu}$ kann die Deviance auch durch die Pearson-Statistik approximiert werden:

$$D(\hat{\boldsymbol{\mu}}|\mathbf{y}) \approx \sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right)^2 / \hat{\mu}_i = X^2. \tag{4.6}$$

Unterscheiden sich Deviance und Pearson-Statistik wesentlich, sollte daher das zugrunde liegende Modell hinterfragt werden.

4.1.2 Missspezifikation

In realen Datensätzen wird oftmals die zentrale Annahme des Poisson-Modells verletzt, nämlich die Äquivalenz von Poisson-Erwartung und Varianz. In diesen Fällen spricht man einerseits von *Überdispersion*, wenn die Variabilität der Daten den Erwartungswert übersteigt, und andererseits von *Unterdispersion*, sollte die Datenvarianz proportional geringer ausfallen als die Erwartung. Grundsätzlich kann man formulieren, dass Über- bzw. Unterdispersion dann auftritt, wenn die Abbildung von Erwartung auf Varianz nicht der identischen Abbildung entspricht:

$$\mathrm{Var}(y|\mathbf{x}) = f\left[\mathbb{E}(y|\mathbf{x})\right] = f\left[\exp(\mathbf{x}^T\boldsymbol{\beta})\right], \quad f \text{ beliebig.}$$

Bei Zähl-Daten kann man nach Winkelmann und Zimmermann (1991) obige Gleichung für Überdispersion folgendermaßen anschreiben,

$$\mathrm{Var}(y|\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) + \sigma^2\left[\mathbb{E}(y|\mathbf{x})\right]^{k+1}, \quad \sigma^2 \in \mathbb{R}^+,\ k \in \mathbb{R}.$$

Für $k = 1$ resultiert hier beispielsweise die Varianzfunktion des Negativ-Binomial-Modells, welches in Kapitel 4.3 näher behandelt wird.

Unterdispersion tritt in der Praxis eher selten auf und wird im Rahmen dieser Arbeit nicht näher betrachtet.

4.2 Überdispersion

Nicht berücksichtigte Überdispersion kann zu unterschätzten Standardfehlern und daher zu falschen Inferenzaussagen bzgl. der Signifikanz der Regressionsparameter führen. Eine Reihe von Tests auf Überdispersion wird ausführlich in Winkelmann (2008) und Cameron und Trivedi (1990) beschrieben. Die einfachsten Indikatoren für eine vorliegende Überdispersion werden nachfolgend erwähnt.

Graphischer Test:
Eine „straight forward"-Methode zur Erkennung von Überdispersion ist die Gegenüberstellung von geschätztem Erwartungswert $\hat{\mathbb{E}}(y_i) = \hat{\mu}_i$ und geschätzter Varianz $\widehat{\mathrm{Var}}(y_i) = (y_i - \hat{\mu}_i)^2$ in einem Scatterplot, ähnlich zur Residuenanalyse beim klassischen linearen Modell. Bei vorliegender Equidispersion sollten diese Punkte um die 45° Linie streuen.

Per Regression:
Eine Alternative basierend auf der Theorie der Regressionsanalyse liegt in der Schätzung der folgenden Regressionsgleichung

$$\widehat{\mathrm{Var}}(y_i) = \hat{\mu}_i + \gamma \cdot h(\hat{\mu}_i),$$

wobei $h()$ eine vorab spezifizierte Funktion darstellt. Hierbei sollte die Hypothese $H_0 : \ \gamma = 0$ nicht verworfen werden können, wenn Equidispersion gegeben ist und somit das Poisson-Modell korrekt gewählt wurde.

4.3 Negativ-Binomial-Modell

Die Negativ-Binomial-Verteilung ist eine oft verwendete Alternative zum Poisson-Modell, vorallem bei vorliegender Überdispersion, da ihre Varianzstruktur extra-Poisson Variabilität widerspiegelt. Ausführliche Studien dieser Modellklasse finden sich unter anderem in Lawless (1987), Ismail und Jemain (2007) und Hilbe (2011).

4.3.1 Die Negativ-Binomial-Verteilung

Die Negativ-Binomial-Verteilung kann, wie nachfolgend beschrieben, als Mischung von Poisson- und Gamma-Verteilungen hergeleitet werden. Salopp gesagt betrachtet man eine Poissonverteilung, dessen Erwartungswert ebenfalls eine Zufallsvariable ist und einer Gammaverteilung folgt. Je nach Parametrisierung resultieren verschiedene Varianzstrukturen.
Sei dazu θ Gamma-verteilt mit Dichtefunktion

$$\begin{aligned} f(\theta|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\theta\beta}, \quad \alpha>0,\beta>0,\theta\in\mathbb{R}^+ \\ \mathbb{E}(\theta) &= \alpha/\beta, \\ \mathrm{Var}(\theta) &= \alpha/\beta^2. \end{aligned} \tag{4.7}$$

Durch eine Reparametrisierung von (4.7) mit $\beta = \alpha/\mu$ ergibt sich $\mathbb{E}(\theta) = \mu$ und $\mathrm{Var}(\theta) = \mu^2/\alpha$. Sei weiters $Y|\theta$ Poisson-verteilt mit bedingter Erwartung $\mathbb{E}(Y|\theta) = \theta$. Betrachtet man nun diese gemischte Poisson-Verteilung und integriert über θ, so ergibt sich als marginale Verteilung von Y eine Negativ-Binomial-Verteilung:

$$\begin{aligned} P(Y=y|\alpha,\mu) &= \int_0^\infty P(Y=y|\theta)f(\theta)\ d\theta \\ &= \int_0^\infty \frac{\mathrm{e}^{-\theta}\theta^y}{y!}\frac{(\alpha/\mu)^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\frac{\theta\alpha}{\mu}}\ d\theta \\ &= \frac{\alpha^\alpha}{\mu^\alpha y!\Gamma(\alpha)}\int_0^\infty \mathrm{e}^{-\theta\left(\frac{\mu+\alpha}{\mu}\right)}\theta^{y+\alpha-1}\ d\theta \end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha^\alpha \Gamma(y+\alpha)}{\mu^\alpha y! \Gamma(\alpha)} \left(\frac{\mu}{\mu+\alpha}\right)^{y+\alpha} \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha} \left(\frac{\mu}{\mu+\alpha}\right)^{y}, \qquad (4.8)
\end{aligned}
$$

d.h. $Y \sim NegBin(\alpha, \mu)$.
Unter Verwendung des Satzes der iterierten Erwartung $\mathbb{E}(Y) = \mathbb{E}_\theta\left[\mathbb{E}(Y|\theta)\right]$ und der Varianzzerlegung $\mathrm{Var}(Y) = \mathbb{E}_\theta\left[\mathrm{Var}(Y|\theta)\right] + \mathrm{Var}_\theta\left[\mathbb{E}(Y|\theta)\right]$ ergeben sich die folgenden Momente

$$
\begin{aligned}
\mathbb{E}(Y) &= \mu \\
\mathrm{Var}(Y) &= \mu + \mu^2/\alpha.
\end{aligned} \qquad (4.9)
$$

Der Parameter α ist also maßgeblich für die Ausprägung der Überdispersion verantwortlich. Für $\alpha \to \infty$ resultiert als Grenzsituation die Poissonverteilung.
Setzt man α als bekannt voraus, kann die Verteilungsfunktion wie folgt umformuliert werden.

$$
\begin{aligned}
P(Y = y|\alpha, \mu) &= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha} \left(\frac{\mu}{\mu+\alpha}\right)^{y} \\
&= \exp\left\{\log\left(\frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\right) + \alpha \log\left(\frac{\alpha}{\mu+\alpha}\right) + y \log\left(\frac{\mu}{\mu+\alpha}\right)\right\} \\
&= \exp\Bigg\{\underbrace{y \cdot \log\left(\frac{\mu}{\mu+\alpha}\right)}_{=:\ \theta} - \underbrace{(-\alpha) \cdot \log\left(\frac{\alpha}{\mu+\alpha}\right)}_{=:\ b(\theta)} \\
&\qquad + \underbrace{\log\left(\frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\right)}_{=:\ c(y,\phi)}\Bigg\}.
\end{aligned}
$$

Mit $\phi = 1,\ a(\phi) = \phi$ und $b(\theta) = -\alpha \cdot \log(1 - \exp(\theta))$ gehört die $\mathrm{NegBin}(\alpha, \mu)$-Verteilung also zur einparametrigen linearen Exponentialfamilie. Somit kann für festes α die Theorie der generalisierten linearen Modelle Anwendung finden.

4.3.2 Modelleigenschaften

Die Wahl der Linkfunktion als log-Link ist im Fall von Zählvariablen vorteilhaft, also $\mathbb{E}(y_i|\mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, da damit auch die Vergleichbarkeit mit dem Poisson-Modell gegeben ist. Sie entspricht jedoch nicht dem für die

Negativ-Binomial-Verteilung kanonischen Link. Die log-Likelihood-Funktion ergibt sich nun als

$$\begin{aligned} l(\boldsymbol{\mu}, \alpha|\mathbf{y}) &= \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \cdot \log\left(\frac{\mu_i}{\mu_i + \alpha}\right) + \alpha \cdot \log\left(\frac{\alpha}{\mu_i + \alpha}\right) + \log\left(\frac{\Gamma(\alpha + y_i)}{y_i!\Gamma(\alpha)}\right) \right\}. \end{aligned} \tag{4.10}$$

Mit $\Gamma(y+\alpha)/\Gamma(\alpha) = \alpha(\alpha+1)\cdots(\alpha+y-1)$ für jedes $\alpha > 0$ und $y \in \mathbb{N}$ kann (4.10) vereinfacht werden zu

$$\begin{aligned} l(\boldsymbol{\mu}, \alpha|\mathbf{y}) = \sum_{i=1}^{n} \Bigg\{ & y_i \cdot \log\left(\frac{\mu_i}{\mu_i + \alpha}\right) + \alpha \cdot \log\left(\frac{\alpha}{\mu_i + \alpha}\right) + \sum_{r=0}^{y_i - 1} \log(\alpha + r) \\ & - \log(y_i!) \Bigg\}, \end{aligned} \tag{4.11}$$

wobei für $y = 0$ gilt, dass die innere Summe ebenfalls gleich Null ist.
Somit erhält man den MLE für $\hat{\beta}_j$, indem die folgenden Scoregleichungen gelöst werden. Für α fix, kann dafür der IRLS-Algorithmus angewendet werden. Als Scorefunktion resultiert

$$\frac{\partial l(\boldsymbol{\mu}|\mathbf{y}, \alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{1 + \mu_i/\alpha}, \quad j = 0, 1, \ldots, p-1 \tag{4.12}$$

und weiters

$$\begin{aligned} \frac{-\partial^2 l(\boldsymbol{\mu}|\mathbf{y}, \alpha)}{\partial \beta_j \partial \beta_k} &= \frac{\partial}{\partial \beta_k} \left(\sum_{i=1}^{n} \frac{(\mu_i - y_i)x_{ij}}{1 + \mu_i/\alpha} \right) \\ &= \sum_{i=1}^{n} \frac{\mu_i x_{ik} x_{ij}(1 + \mu_i/\alpha) - (\mu_i - y_i)x_{ij}x_{ik}\frac{\mu_i}{\alpha}}{(1 + \mu_i/\alpha)^2} \\ &= \sum_{i=1}^{n} \frac{\mu_i(1 + y_i/\alpha)x_{ij}x_{ik}}{(1 + \mu_i/\alpha)^2}, \quad j,\ k = 0, 1, \ldots, p-1. \end{aligned} \tag{4.13}$$

Soll auch α geschätzt werden, werden weitere Scoregleichung benötigt.

$$\begin{aligned} \frac{\partial l(\boldsymbol{\mu}, \alpha|\mathbf{y})}{\partial \alpha} = \sum_{i=1}^{n} \Bigg\{ & y_i \frac{\alpha + \mu_i}{\mu_i} \frac{\mu_i}{(\alpha + \mu_i)^2} + \log\left(\frac{\alpha}{\alpha + \mu_i}\right) \\ & + \alpha \frac{\alpha + \mu_i}{\alpha} \frac{\alpha + \mu_i - \alpha}{(\alpha + \mu_i)^2} + \sum_{r=0}^{y_i - 1} \frac{1}{\alpha + r} \Bigg\} \end{aligned}$$

$$= \sum_{i=1}^{n} \left\{ \log\left(\frac{\alpha}{\alpha+\mu_i}\right) + \frac{\mu_i + y_i}{\alpha+\mu_i} + \sum_{r=0}^{y_i-1} \frac{1}{\alpha+r} \right\}, \tag{4.14}$$

$$\begin{aligned}
\frac{-\partial^2 l(\boldsymbol{\mu}, \alpha|\mathbf{y})}{\partial \beta_j \partial \alpha} &= \frac{\partial}{\partial \alpha} \left(\sum_{i=1}^{n} \frac{(\mu_i - y_i) x_{ij}}{1 + \mu_i/\alpha} \right) \\
&= \sum_{i=1}^{n} \frac{-(\mu_i - y_i) x_{ij} \mu_i (-1) \alpha^{-2}}{(1+\mu_i/\alpha)^2} \\
&= \sum_{i=1}^{n} \frac{\mu_i(\mu_i - y_i) x_{ij}}{\alpha^2 (1+\mu_i/\alpha)^2}, \quad j = 0, 1, \ldots, p-1, \qquad (4.15)
\end{aligned}$$

$$\begin{aligned}
\frac{-\partial^2 l(\boldsymbol{\mu}, \alpha|\mathbf{y})}{\partial \alpha^2} &= \frac{\partial}{\partial \alpha} \left(\sum_{i=1}^{n} \left\{ -\log\left(\frac{\alpha}{\alpha+\mu_i}\right) - \frac{\mu_i + y_i}{\alpha+\mu_i} + \sum_{r=0}^{y_i-1} \frac{1}{-(\alpha+r)} \right\} \right) \\
&= \sum_{i=1}^{n} \left\{ -\frac{\alpha+\mu_i}{\alpha} \frac{\alpha+\mu_i-\alpha}{(\alpha+\mu_i)^2} - \frac{-\mu_i - y_i}{(\alpha+\mu_i)^2} + \sum_{r=0}^{y_i-1} \frac{1}{(\alpha+r)^2} \right\} \\
&= \sum_{i=1}^{n} \left\{ -\frac{\alpha\mu_i + \mu_i^2 - \alpha\mu_i - \alpha y_i}{\alpha(\alpha+\mu_i)^2} + \sum_{r=0}^{y_i-1} \frac{1}{(\alpha+r)^2} \right\} \\
&= \sum_{i=1}^{n} \left\{ \sum_{r=0}^{y_i-1} \frac{1}{(\alpha+r)^2} - \frac{\mu_i^2 - \alpha y_i}{\alpha(\alpha+\mu_i)^2} \right\}. \qquad (4.16)
\end{aligned}$$

Die Berechnung der MLEs $(\hat{\boldsymbol{\beta}}, \hat{\alpha})$ erfolgt nun simultan. Die Prozedur startet mit einem initialen Wert von α, $\alpha^{(0)}$, mit dem dann $\log L(\boldsymbol{\mu}|\mathbf{y}, \alpha^{(0)})$ nach $\boldsymbol{\beta}$ maximiert wird und zum Schätzer $\boldsymbol{\beta}^{(1)}$ führt. Im zweiten Schritt fixiert man $\boldsymbol{\beta}$ mit $\boldsymbol{\beta}^{(1)}$, maximiert $\log L(\alpha|\mathbf{y}, \boldsymbol{\mu}(\boldsymbol{\beta}^{(1)}))$ nach α und erhält den Schätzer $\alpha^{(1)}$. Für die Maximierung kann man zum Beispiel eine Newton-Raphson Iteration verwenden,

$$\alpha^{(t+1)} = \alpha^{(t)} - \frac{\partial l(\boldsymbol{\mu}, \alpha|\mathbf{y})}{\partial \alpha} \left(\frac{\partial^2 l(\boldsymbol{\mu}, \alpha|\mathbf{y})}{\partial \alpha^2} \right)^{-1}. \tag{4.17}$$

Durch Iteration dieser abwechselnden Fixierung von α und $\boldsymbol{\beta}$ resultieren dann bei Konvergenz die MLEs $(\hat{\boldsymbol{\beta}}, \hat{\alpha})$.

Unter $\alpha > 0$ und gewissen Regularitätsbedingungen, sodass für $n \to \infty$ $n^{-1}\mathbf{I}(\boldsymbol{\beta}, \alpha)$ gegen eine positiv definite Matrix strebt, ist auch $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{\alpha} - \alpha)$ asymptotisch normalverteilt mit Erwartungswertvektor $\mathbf{0}$ und Kovarianzmatrix $n\mathbf{I}(\boldsymbol{\beta}, \alpha)^{-1}$. Die Kovarianzmatrix setzt sich hierbei zusammen als

$$n\mathbf{I}(\boldsymbol{\beta}, \alpha)^{-1} = n \begin{bmatrix} \mathbf{I}(\boldsymbol{\beta}|\alpha)^{-1} & \mathbf{0}_p \\ \mathbf{0}_p^T & \mathbf{I}(\alpha|\boldsymbol{\beta})^{-1} \end{bmatrix}, \tag{4.18}$$

wobei

$$\mathbf{I}_{jk}(\boldsymbol{\beta}|\alpha) = \mathbb{E}\left(\frac{-\partial^2 l(\boldsymbol{\mu}|\mathbf{y},\alpha)}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{\mu_i x_{ij} x_{ik}}{1+\mu_i/\alpha}, \quad j,\ k = 0,1,\ldots,p-1 \tag{4.19}$$

und

$$\mathbf{I}(\alpha|\boldsymbol{\beta}) = \mathbb{E}\left(\frac{-\partial^2 l(\boldsymbol{\mu},\alpha|\mathbf{y})}{\partial\alpha^2}\right) = \sum_{i=1}^{n}\left\{\sum_{r=0}^{\infty}\frac{1}{(\alpha+r)^2}\mathbb{P}(y_i \geq r) - \frac{\mu_i^2 - \alpha\mu_i}{\alpha(\alpha+\mu_i)^2}\right\}. \tag{4.20}$$

Ein weiterer Ansatz, vorgeschlagen von Breslow (1984), ist die Schätzung von α durch die Momentenmethode. Hierbei wird die Pearson-Statistik mit ihren Freiheitsgraden in Relation gesetzt,

$$\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i(1+\hat{\mu}_i/\alpha)} = n-p. \tag{4.21}$$

Um die Prozedur zu starten sollte $\alpha^{(0)} = \infty$ gewählt und $\boldsymbol{\beta}^{(0)}$ als MLE eines Poisson-Modells berechnet werden. Die resultierenden Werte werden nun in (4.21) eingesetzt und $\alpha^{(1)}$ bestimmt und damit wieder ein Negativ-Binomial-Modell gefittet, um $\boldsymbol{\beta}^{(1)}$ zu erhalten. Durch Iteration ergeben sich die Schätzer $(\tilde{\boldsymbol{\beta}}, \tilde{\alpha})$ (Konvergenz vorausgesetzt), welche ebenfalls asymptotisch normalverteilt sind. Weiters zeigte Lawless (1987), dass die Momentenmethode robuster ist als die oben beschriebene adaptierte IRLS-Prozedur, jedoch ist sie weniger effizient, falls das Negativ-Binomial-Modell korrekt ist.

Die Deviance des Negativ-Binomial-Modells gestaltet sich etwas komplexer als für Standardvertreter der Exponentialfamilie, da sich für das saturierte und das p-Parameter-Modell im Allgemeinen zwei verschiedene Schätzer für den Parameter α ergeben. Im Folgenden bezeichne $\hat{\alpha}$ den MLE von α unter $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ und α^* den MLE von α unter $\hat{\boldsymbol{\mu}} = \mathbf{y}$.

$$\begin{aligned}
D(\hat{\boldsymbol{\mu}},\hat{\alpha}|\mathbf{y}) &= -2\left(l(\hat{\boldsymbol{\mu}},\hat{\alpha}|\mathbf{y}) - l(\mathbf{y},\alpha^*|\mathbf{y})\right) \\
&= -2\sum_{i=1}^{n}\left[y_i\log\left(\frac{\hat{\mu}_i}{\hat{\mu}_i+\hat{\alpha}}\right) + \hat{\alpha}\log\left(\frac{\hat{\alpha}}{\hat{\mu}_i+\hat{\alpha}}\right) + \log\left(\frac{\Gamma(\hat{\alpha}+y_i)}{y_i!\Gamma(\hat{\alpha})}\right)\right. \\
&\quad \left. - y_i\log\left(\frac{y_i}{y_i+\alpha^*}\right) - \alpha^*\log\left(\frac{\alpha^*}{y_i+\alpha^*}\right) - \log\left(\frac{\Gamma(\alpha^*+y_i)}{y_i!\Gamma(\alpha^*)}\right)\right] \\
&= -2\sum_{i=1}^{n}\left[y_i\log\left(\frac{\hat{\mu}_i(y_i+\alpha^*)}{y_i(\hat{\mu}_i+\hat{\alpha})}\right) + \hat{\alpha}\log\left(\frac{\hat{\alpha}}{\hat{\mu}_i+\hat{\alpha}}\right)\right. \\
&\quad \left. - \alpha^*\log\left(\frac{\alpha^*}{y_i+\alpha^*}\right) + \log\left(\frac{\Gamma(\hat{\alpha}+y_i)\Gamma(\alpha^*)}{\Gamma(\alpha^*+y_i)\Gamma(\hat{\alpha})}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{1+\alpha^*/y_i}{1+\hat{\alpha}/\hat{\mu}}\right) + \hat{\alpha}\log\left(\frac{\hat{\alpha}/\hat{\mu}_i}{1+\hat{\alpha}/\hat{\mu}_i}\right)\right. \\
&\quad \left. - \alpha^*\log\left(\frac{\alpha^*/y_i}{1+\alpha^*/y_i}\right) + \log\left(\frac{\Gamma(\hat{\alpha}+y_i)\Gamma(\alpha^*)}{\Gamma(\alpha^*+y_i)\Gamma(\hat{\alpha})}\right)\right] \\
&= -2\sum_{i=1}^{n}\left[y_i\log\left(1+\alpha^*/y_i\right) - y_i\log\left(1+\hat{\alpha}/\hat{\mu}\right) + \hat{\alpha}\log\left(\hat{\alpha}/\hat{\mu}_i\right)\right. \\
&\quad - \hat{\alpha}\log\left(1+\hat{\alpha}/\hat{\mu}_i\right) - \alpha^*\log\left(\alpha^*/y_i\right) + \alpha^*\log\left(1+\alpha^*/y_i\right) \\
&\quad \left. + \log\left(\frac{\Gamma(\hat{\alpha}+y_i)\Gamma(\alpha^*)}{\Gamma(\alpha^*+y_i)\Gamma(\hat{\alpha})}\right)\right] \\
&= -2\sum_{i=1}^{n}\left[(y_i-\alpha^*)\log\left(1+\frac{\alpha^*}{y_i}\right) - (y_i-\hat{\alpha})\log\left(1+\frac{\hat{\alpha}}{\hat{\mu}}\right)\right. \\
&\quad \left. + \hat{\alpha}\log\left(\frac{\hat{\alpha}}{\hat{\mu}_i}\right) - \alpha^*\log\left(\frac{\alpha^*}{y_i}\right) + \log\left(\frac{\Gamma(\hat{\alpha}+y_i)\Gamma(\alpha^*)}{\Gamma(\alpha^*+y_i)\Gamma(\hat{\alpha})}\right)\right].
\end{aligned}
\tag{4.22}
$$

4.3.3 Test Poisson- versus Negativ-Binomial-Modell

Zur Überprüfung der Adäquatheit eines Negativ-Binomial-Modells kann bei vorliegender Überdispersion ein Likelihood-Quotiententest angewendet werden.

Sei dazu $\hat{l}_r$ der maximale Wert der log-Likelihood-Funktion unter dem restriktiveren Poisson-Modell und $\hat{l}_u$ das Maximum der log-Likelihood-Funktion unter dem allgemeineren Negativ-Binomial-Modell. Weiters sei k die Anzahl an Restriktionen (hier $k = 1$, da das Poisson-Modell mit $\alpha \to \infty$ in der Familie der Negativ-Binomial-Modelle enthalten ist).

Getestet wird die Hypothese $H_0\colon \alpha = \infty$, also das Vorliegen einer Equidispersion, gegen $H_1\colon \alpha$ *beliebig* mittels folgender Teststatistik:

$$LQT = -2(\hat{l}_r - \hat{l}_u) \sim \chi^2_k$$

bzw. in diesem speziellen Fall

$$LQT = -2(\hat{l}_r - \hat{l}_u) \sim \chi^2_1.$$

Die Nullhypothese wird verworfen, wenn $LQT > \chi^2_{1-\alpha,1}$. Diese Teststatistik entspricht offensichtlich, wie in Abschnitt 3.3.3 Parametertest beschrieben, einer Deviance-Reduktion unter Annahme einer Negativ-Binomial-Verteilung.

4.4 Quasi-Likelihood Funktion

Eine weitere Möglichkeit mit Überdispersion bei Poisson-verteilten Daten umzugehen, ist die Anwendung eines Quasi-Likelihood-Ansatzes. Diese Methode wurde erstmals von WEDDERBURN (1974) vorgestellt und verlangt keine vollständige Verteilungsannahme, die Spezifikation der Erwartungswert-Varianz-Relation ist ausreichend. Schätzer für die zugehörigen Modellparameter erhält man dann durch Maximierung der Quasi-Likelihood-Funktion mittels einem modifizierten IRLS-Algorithmus.

4.4.1 Definition

Definition 4.2 (WEDDERBURN, 1974 erweitert). *Sei y eine Zufallsvariable mit Erwartung μ und Varianz $\phi V(\mu)$, wobei $a(\phi) = \phi$ und $V()$ eine bekannte Funktion darstellt, dann ist die Quasi-Likelihood Funktion (QL-Funktion), korrekterweise log-Quasi-Likelihood Funktion, $q(\mu|y)$ definiert über die Beziehung*

$$\frac{\partial q(\mu|y)}{\partial \mu} = \frac{y-\mu}{\phi V(\mu)}, \tag{4.23}$$

oder äquivalent dazu in Integraldarstellung

$$q(\mu|y) = \int^{\mu} \frac{y-t}{\phi V(t)}\, dt + Funktion\ in\ y\ (und\ \phi). \tag{4.24}$$

MCCULLAGH UND NELDER (1989) formulierten die QL-Funktion leicht abweichend von (4.24) als

$$q(\mu|y) = \int_{y}^{\mu} \frac{y-t}{\phi V(t)}\, dt. \tag{4.25}$$

Jedoch führt die Angabe einer fixen unteren Integrationsgrenze bei Integration nach t zu einem zusätzlichen Term in y und somit zur Äquivalenz von (4.24) und (4.25).

Die QL-Funktion der gesamten Stichprobe $\mathbf{y}$ ergibt sich bei unabhängigen y_i $(i = 1, \ldots, n)$, definitionsbedingt Voraussetzung bei GLM, durch Summation der einzelnen QL-Funktionen

$$q(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^{n} q(\mu_i|y_i). \tag{4.26}$$

4.4.2 Eigenschaften

Die QL-Funktion und im Speziellen die *Quasi-Scorefunktion* $\partial q/\partial \mu$ haben einige wesentliche Eigenschaften mit der log-Likelihood Funktion gemein, was die Handhabung von q erheblich erleichtert.

Satz 4.1 (Wedderburn, 1974). *Seien y und $q(\mu|y)$ gegeben laut Definition 4.2, so hat $q(\mu|y)$ die folgenden Eigenschaften:*

$$\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = 0, \tag{4.27}$$

$$\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \beta_j}\right) = 0, \qquad j = 0, \ldots, p-1 \tag{4.28}$$

$$\mathbb{E}\left(\left(\frac{\partial q(\mu|y)}{\partial \mu}\right)^2\right) = -\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right) = \frac{1}{\phi V(\mu)}, \tag{4.29}$$

$$\begin{aligned}\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \beta_j}\frac{\partial q(\mu|y)}{\partial \beta_k}\right) &= -\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \beta_j \partial \beta_k}\right) \\ &= \frac{1}{\phi V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}, \qquad j,k = 0, \ldots, p-1.\end{aligned} \tag{4.30}$$

Beweis. (4.27) folgt direkt aus der Definition von $q(\mu|y)$.

Für (4.28) wird der linke Term umformuliert und führt zur benötigten Gleichung,

$$\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \beta_j}\right) = \mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \mu}\frac{\partial \mu}{\partial \beta_j}\right) = \underbrace{\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \mu}\right)}_{\overset{(4.27)}{=}0}\frac{\partial \mu}{\partial \beta_j} = 0.$$

Um (4.29) zu zeigen, ist die Formel $\mathbb{E}(y-\mu)^2 = \mathrm{Var}(y) = \phi V(\mu)$ hilfreich. Hiermit folgt zum Einen

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

und durch Umformulierung ergibt sich der andere Teil der Gleichung,

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

$$
\begin{aligned}
&= -\mathbb{E}\left(\frac{-\phi V(\mu) - (y-\mu)\phi V'(\mu)}{\phi^2 V(\mu)^2}\right) \\
&= \frac{1}{\phi V(\mu)} + \frac{\phi V'(\mu)}{\phi^2 V(\mu)^2}\underbrace{\mathbb{E}(y-\mu)}_{=0} \\
&= \frac{1}{\phi V(\mu)}.
\end{aligned}
$$

Identität (4.30) wird auch wieder getrennt betrachtet, einerseits gilt

$$
\mathbb{E}\left(\frac{\partial q(\mu|y)}{\partial \beta_j}\frac{\partial q(\mu|y)}{\partial \beta_k}\right) = \mathbb{E}\left(\left(\frac{\partial q(\mu|y)}{\partial \mu}\right)^2\right)\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k} \overset{(4.29)}{=} \frac{1}{\phi V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}
$$

und andererseits

$$
\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \beta_j \partial \beta_k}\right) &= -\mathbb{E}\left(\frac{\partial}{\partial \beta_k}\left(\frac{\partial q(\mu|y)}{\partial \mu}\frac{\partial \mu}{\partial \beta_j}\right)\right) \\
&= -\mathbb{E}\left(\frac{\partial}{\partial \mu}\left(\frac{y-\mu}{\phi V(\mu)}\right)\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}\right) \overset{(4.29)}{=} \frac{1}{\phi V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}.
\end{aligned}
$$

□

Weiters ergibt sich für Verteilungen von y aus der Exponentialfamilie, dass in diesem Fall die log-Likelihood Funktion äquivalent zur QL-Funktion ist, was in folgendem Satz beschrieben wird.

Satz 4.2 (Wedderburn, 1974). *Für eine Beobachtung y mit $\mathbb{E}(y) = \mu$ und $Var(y) = \phi V(\mu)$ hat die log-Likelihood Funktion $l(\mu|y)$ genau dann die Eigenschaft*

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

wenn die Dichte oder 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. [2] $\Rightarrow$) Sei $\dfrac{\partial l(\mu|y)}{\partial \mu} = \dfrac{y-\mu}{\phi V(\mu)}$, dann liefert die Integration nach μ

$$
l(\mu|y) = \int^{\mu} \frac{\partial l(\mu|y)}{\partial \mu}\, d\mu = \int^{\mu} \frac{y-\mu}{\phi V(\mu)}\, d\mu
$$

[2] Beweis aus dem Vorlesungsskript „Generalisierte Lineare Modelle“ von Friedl H., 2000

$$= \frac{y}{\phi} \underbrace{\int^{\mu} \frac{1}{V(\mu)}\, d\mu}_{\theta} - \frac{1}{\phi} \underbrace{\int^{\mu} \frac{\mu}{V(\mu)}\, d\mu}_{b(\theta)}$$
$$= \frac{y\theta - b(\theta)}{\phi} + c(y, \phi).$$

Daraus folgt die Behauptung $f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$.

$\Leftarrow$) Sei nun die Dichte bzw. Wahrscheinlichkeitsfunktion von y Teil der einparametrigen Exponentialfamilie mit $a(\phi) = \phi$, daher gilt $\mathbb{E}(y) = \mu = b'(\theta)$ und $\text{Var}(y) = \phi V(\mu) = \phi b''(\theta)$ und weiters

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

Nun ist $l(\mu|y) = \frac{y\theta - b(\theta)}{\phi} + c(y, \phi)$ und θ eine Funktion von μ, somit folgt

$$\frac{\partial l(\mu|y)}{\partial \mu} = \frac{y}{\phi}\frac{d\theta}{d\mu} - \frac{b'(\theta)}{\phi}\frac{d\theta}{d\mu}$$
$$= \frac{y - \mu}{\phi V(\mu)}.$$

□

Im Folgenden werden die Erkenntnisse aus Satz 4.2 verwendet, um für vorliegende Varianzfunktionen die QL-Funktionen zu bestimmen (vgl. auch McCullagh und Nelder, 1989).

Konstante Varianz:
$V(\mu) = 1,\ \phi = \sigma^2,\ \mu \in \mathbb{R},\ y \in \mathbb{R}$ (vergleichbar mit $y \sim N(\mu, \sigma^2)$):

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

Varianz proportional zu Erwartung:
$V(\mu) = \mu,\ \phi = 1,\ 0 < \mu,\ 0 \leq y$ (vergleichbar mit $y \sim Poisson(\mu)$):

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

Varianz proportional zu quadrierter Erwartung:
$V(\mu) = \mu^2,\ \phi = 1,\ 0 < \mu,\ 0 \leq y$ (vergleichbar mit $y \sim Gamma(\mu, 1)$):

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

Varianz proportional zu Erwartung^p:
$V(\mu) = \mu^p,\ \phi = 1,\ 0 < \mu,\ 0 \leq y,\ p \geq 3$:

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

Varianz wie bei Negativ-Binomial-Verteilung mit festem α:
$V(\mu) = \mu + \mu^2/\alpha,\ \phi = 1,\ 0 < \mu,\ 0 \leq y,\ 0 < \alpha$ (vergleichbar mit $y \sim NegBin(\alpha, \mu)$):

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

Weiters formulierte Wedderburn (1974) folgendes

Korollar 4.1. *Ist die Verteilung von y in Termen von μ spezifiziert, sodass die log-Likelihood Funktion $l(\mu|y)$ definiert werden kann, und ist $Var(y) = \phi V(\mu)$, dann gilt*

$$Var\left(\frac{\partial q(\mu|y)}{\partial \mu}\right) = -\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right) \leq -\mathbb{E}\left(\frac{\partial^2 l(\mu|y)}{\partial \mu^2}\right) = Var\left(\frac{\partial l(\mu|y)}{\partial \mu}\right).$$

Beweis. Aus Satz 4.1 ist bekannt, dass

$$-\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right) = \frac{1}{\phi V(\mu)} = \frac{1}{\mathrm{Var}(y)} \tag{4.31}$$

entspricht. Nun kann (4.31) umformuliert werden zu

$$\mathrm{Var}(y) \overset{?}{\geq} -\mathbb{E}\left(\frac{\partial^2 l(\mu|y)}{\partial \mu^2}\right)^{-1}.$$

Obige Ungleichung kann unmittelbar durch die Cramér-Rao Ungleichung verifiziert werden, die besagt, dass unter gewissen Regularitätsbedingungen für eine beliebige Statistik $T(y)$

$$\mathrm{Var}(T(y)) \geq \frac{\left(\frac{\partial}{\partial \mu}\mathbb{E}(T(y))\right)^2}{\mathbb{E}\left(\frac{\partial}{\partial \mu} l(\mu|y)\right)^2}$$

gilt. Setzt man nun $T(y) = y$, so resultiert

$$\mathrm{Var}(y) \geq \frac{\left(\frac{\partial}{\partial \mu}\mu\right)^2}{\mathbb{E}\left(\frac{\partial}{\partial \mu} l(\mu|y)\right)^2} = \frac{1}{-\mathbb{E}\left(\frac{\partial^2}{\partial \mu^2} l(\mu|y)\right)}.$$

□

Gleichheit herrscht, wenn dem angenommenen QL-Modell eine einparametrige Exponentialfamilie zugrunde liegt. Folglich wird in diesem Fall die Fisher-Information $-\mathbb{E}\left(\frac{\partial^2 l(\mu|y)}{\partial \mu^2}\right)$ minimiert. Somit ist es naheliegend $-\mathbb{E}\left(\frac{\partial^2 q(\mu|y)}{\partial \mu^2}\right) = 1/\mathrm{Var}(y)$ als Maß der Information zu wählen, wenn nur die Erwartungswert-Varianz-Beziehung bekannt ist.

4.4.3 Quasi-Dichte

Liegt einer Erwartungswert-Varianz-Beziehung keine Exponentialfamilie zugrunde, so kann man ad hoc auch keine zugehörige Dichtefunktion angeben. Nelder und Lee (1992) sind diesem Problem habhaft geworden und haben die sogenannte *Quasi-Dichte* eingeführt.

Definition 4.3. *Sei $q(\mu|y)$ eine log-Quasi-Likelihood Funktion, so wird mit der Normalisierungsfunktion*

$$w(\mu) = \int_{\mathbb{R}} \exp(q(\mu|y))dy \tag{4.32}$$

die Quasi-Dichte definiert als

$$f_q(y, \mu) = \frac{\exp(q(\mu|y))}{w(\mu)}. \tag{4.33}$$

Hierbei gilt $w(\mu) = 1$, $\forall\mu$, wenn zur Varianzannahme auch ein Vertreter der Exponentialfamilie existiert.
Die zugehörige log-Likelihood Funktion ergibt sich dann als

$$l_q(\mu|y) = \log(f_q(y, \mu)) = q(\mu|y) - \log(w(\mu)) \tag{4.34}$$

und weiters

$$\frac{\partial l_q(\mu|y)}{\partial\mu} = \frac{\partial q(\mu|y)}{\partial\mu} - \frac{\partial \log(w(\mu))}{\partial\mu} \tag{4.35}$$

als Scorefunktion.
Der Score der Quasi-Dichte unterscheidet sich also vom Quasi-Score durch den letzten Term von (4.35), der sich wie folgt vereinfachen lässt (unter der Annahme, dass die Vertauschbarkeit von Integral und Differential gegeben ist),

$$\begin{aligned}
\frac{\partial \log(w(\mu))}{\partial\mu} &= \frac{1}{w(\mu)}\frac{\partial w(\mu)}{\partial\mu} \\
&= \frac{1}{w(\mu)}\int_{\mathbb{R}} \frac{\partial \exp(q(\mu|y))}{\partial\mu}dy \\
&= \frac{1}{w(\mu)}\int_{\mathbb{R}} \exp(q(\mu|y))\frac{\partial q(\mu|y)}{\partial\mu}dy \\
&= \int_{\mathbb{R}} \frac{y-\mu}{\phi V(\mu)}\frac{\exp(q(\mu|y))}{w(\mu)}dy \\
&= \int_{\mathbb{R}} \frac{y-\mu}{\phi V(\mu)} f_q(y, \mu)dy \\
&= \mathbb{E}_q\left(\frac{y-\mu}{\phi V(\mu)}\right) \\
&= \frac{\mu_q - \mu}{\phi V(\mu)}.
\end{aligned} \tag{4.36}$$

Der Parameter μ_q wird als *Quasi-Mean* bezeichnet. Ist die Differenz $\mu_q - \mu$ verglichen mit $y - \mu$ klein, so kann angenommen werden, dass sich auch der Maximum-Quasi-Likelihood-Schätzer nicht sehr vom Maximum-Likelihood-Schätzer bezüglich der Quasi-Verteilung unterscheidet.

4.4.4 Parameterschätzung

Da ϕ als unabhängig von $\boldsymbol{\mu}$ vorausgesetzt wird, erfolgt die Schätzung des Parametervektors $\boldsymbol{\beta}$ wie beim normalen GLM, mit dem Unterschied, dass die erwartete Hessematrix der Quasi-Scorefunktion verwendet wird. Der Dispersionsparameter ϕ wird mittels der Momentenmethode geschätzt, d.h.

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{1}{n-p} X^2, \tag{4.37}$$

wobei definitionsgemäß $a(\phi) = \phi$ angenommen wird.

McCullagh (1983) befasste sich auch ausführlich mit den (asymptotischen) Eigenschaften des Parameterschätzers $\hat{\boldsymbol{\beta}}$ und kam unter gewissen Regularitätsbedingungen zu den folgenden Resultaten.

- $\frac{1}{\sqrt{n}} \frac{\partial q(\boldsymbol{\mu}|\mathbf{y})}{\partial \boldsymbol{\beta}} \xrightarrow{d} N_p \left(\mathbf{0}, -\frac{1}{n} \mathbb{E} \left(\frac{\partial^2 q(\boldsymbol{\mu}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta}^2} \right) \right)$
- $\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \mathbf{0}$
- $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p \left(\mathbf{0}, -n\mathbb{E} \left(\frac{\partial^2 q(\boldsymbol{\mu}(\boldsymbol{\beta})|\mathbf{y})}{\partial \boldsymbol{\beta}^2} \right)^{-1} \right)$.

Quasi-Deviance:
Zur Bewertung der Anpassungsgüte von ineinander geschachtelten Modellen kann man, wie im Fall der Exponentialfamilie, die sogenannte *Quasi-Deviance* verwenden (vgl. Nelder und Pregibon, 1987)

$$D_q(\boldsymbol{\mu}|\mathbf{y}) = -2\phi \left(q(\boldsymbol{\mu}|\mathbf{y}) - q(\mathbf{y}|\mathbf{y}) \right) = -2 \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} \, dt. \tag{4.38}$$

Eine weitere Schreibweise ergibt sich unter Verwendung von (4.25) als

$$D_q(\boldsymbol{\mu}|\mathbf{y}) = -2\phi q(\boldsymbol{\mu}|\mathbf{y}) = -2 \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} \, dt. \tag{4.39}$$

In beiden Fällen wird die minimale Quasi-Deviance vom Maximum-Quasi-Likelihood-Schätzer erzeugt.

Hypothesentest:
Es kann auch, wie im Standard-GLM-Fall, ein Hypothesentest für geschachtelte Modelle definiert werden. Seien dazu H_A und H_B zwei geschachtelte Hypothesen mit Dimensionen $A < B$, dann gilt unter H_A,

$$D_q(\hat{\boldsymbol{\mu}}_B, \hat{\boldsymbol{\mu}}_A) = D_q(\hat{\boldsymbol{\mu}}_A|\mathbf{y}) - D_q(\hat{\boldsymbol{\mu}}_B|\mathbf{y}) \sim \chi^2_{B-A}.$$

4.4.5 Gamma-Modell

Aus praktischen Gründen kann auch ein log-lineares Quasi-Gamma-Modell zur Modellierung von Zähl-Daten mit vorliegender Überdispersion verwendet werden, da, wie in Abschnitt 4.4.2 gezeigt, ein QL-Ansatz für eine Verteilung mit Varianz proportional zur quadrierter Erwartung vergleichbar mit einer Gamma-Verteilung ist. Deshalb wird auch diese Modell-Familie im Folgenden näher betrachtet (vgl. McCullagh und Nelder, 1989).

Modelleigenschaften
Die Gamma-Verteilung gehört, wie schon im Abschnitt 3.1.2 erwähnt, zur Exponentialfamilie. Die Dichte einer Gamma-verteilten Zufallsvariable Y lässt sich schreiben als

$$f(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \mathrm{e}^{-y\beta}, \qquad \alpha,\ \beta,\ y > 0.$$

Die Reparametrisierung $\mu = \nu/\beta$ mit $\nu = \alpha$ ergibt

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

mit $\mu,\ \nu,\ y > 0$ und weiters $\mathbb{E}(y) = \mu$, $\mathrm{Var}(y) = \mu^2/\nu$ und $\phi = 1/\nu$.

Der log-Link entspricht hier nicht dem kanonischen Link der Gammaverteilung, das wäre der Reziproke. Dies birgt jedoch den Vorteil, dass der log-Link keine negativen bzw. praktisch auch keine Nullwerte für den Erwartungswert zulässt und somit zu mehr Stabilität bei den Kalkulationen führt. Die log-Likelihood-Funktion lautet

$$l(\boldsymbol{\mu}, \nu|\mathbf{y}) = -\nu \sum_{i=1}^{n} \left(\frac{y_i}{\mu_i} - \log \mu_i\right) + \nu \log \nu^n + (\nu - 1) \sum_{i=1}^{n} \log y_i - n \log \Gamma(\nu). \tag{4.40}$$

Da ein Standard-GLM vorliegt, kann auch hier der MLE für $\boldsymbol{\beta}$ durch den IRLS-Algorithmus berechnet werden. Dadurch wird implizit vorausgesetzt, dass $\nu = \phi^{-1}$ ein konstanter, vom Mean unabhängiger Parameter ist.
Die skalierte Deviance lässt sich also wie folgt darstellen

$$\frac{1}{\phi} D(\hat{\boldsymbol{\mu}}|\mathbf{y}) = -2 \sum_{i=1}^{n} \left[\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i \right]. \tag{4.41}$$

Beinhaltet das vorliegende Modell einen Intercepttterm, so summiert sich laut Satz 3.1 der zweite Teil der Summe zu Null und die Deviance vereinfacht sich unter der log-Link-Annahme zu

$$\frac{1}{\phi} D(\hat{\boldsymbol{\mu}}|\mathbf{y}) = -2 \sum_{i=1}^{n} \left[\log(y_i/\hat{\mu}_i) \right]. \tag{4.42}$$

Beide Varianten sind aber offensichtlich nur sinnvoll, wenn alle Beobachtungen echt größer Null sind. Bei Vorliegen auch nur einer Null-Beobachtung ist $\hat{\nu} = 0$ und es resultiert $\hat{\phi} = \infty$. Daher sollte in diesem Fall der Momentenschätzer als konsistenter Schätzer für ϕ bevorzugt werden,

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]^2 = \frac{X^2}{n-p}. \tag{4.43}$$

Die geschätzte Varianz-Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ ergibt sich dann aus (3.13) als

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \operatorname{diag}(1/\hat{\phi}, \ldots, 1/\hat{\phi})\mathbf{X})^{-1} = \hat{\phi}(\mathbf{X}^T\mathbf{X})^{-1}. \tag{4.44}$$

4.5 Extended Quasi-Likelihood Funktion

Die bisher vorgestellten Methoden haben ausschließlich Regressionsmodelle behandelt, bei denen die Kenntnis der Varianzfunktion Voraussetzung war. Auch der Quasi-Likelihood-Ansatz erlaubt lediglich die freie Wahl bzw. Schätzung eines konstanten Dispersionsparameters. Durch die Einführung einer *extended Quasi-Likelihood Funktion (kurz: EQL-Funktion)*, vorgestellt von Nelder und Pregibon (1987), wurde es möglich Modelle zu fitten, die auch in der Varianzfunktion $V(\boldsymbol{\mu})$ einen unbekannten (zu schätzenden) Parameter θ erlauben oder den Dispersionsparameter ϕ als Funktion von Kovariablen beinhalten.

4.5.1 Definition

Definition 4.4 (Nelder und Pregibon, 1987). *Sei y eine Beobachtung mit Erwartung μ und Varianz $\phi V(\mu)$, dann ist die extended Quasi-Likelihood Funktion $q^+(\mu, \phi|y)$ definiert als*

$$q^+(\mu, \phi|y) = -\frac{1}{2}\log\left[2\pi\phi V(y)\right] - \frac{1}{2}D_q(\mu|y)/\phi, \tag{4.45}$$

wobei D_q die Quasi-Deviance (4.38) und ϕ den Dispersionsparameter darstellt.

Wie schon die QL-Funktion, setzt auch die EQL-Funktion nicht die volle Verteilungsannahme voraus, sondern nur die Kenntnis der Gestalt der ersten beiden Momente. Liegt eine unabhängige Stichprobe $\mathbf{y} = (y_1, \ldots, y_n)^T$ vor, so ergibt sich die zugehörige EQL-Funktion als Summe der Einzelnen

$$q^+(\boldsymbol{\mu}, \phi|\mathbf{y}) = \sum_{i=1}^{n} q^+(\mu_i, \phi|y_i). \tag{4.46}$$

McCullagh und Nelder (1989) beschreiben die Konstruktion der EQL-Funktion näher. Für eine Beobachtung y soll $q^+(\mu, \phi|y)$ so konstruiert sein, dass es bei bekanntem ϕ mit der QL-Funktion $q(\mu|y)$ übereinstimmt und für unbekanntes ϕ die log-Likelihood Eigenschaften bzgl. der Ableitung nach dem Dispersionsparameter erfüllt. Diese Überlegungen führen zu folgender Darstellung

$$\begin{aligned} q^+(\mu, \phi|y) &= q(\mu|y) + h(\phi|y) \\ &\overset{(4.39)}{=} -\frac{D_q(\mu|y)}{2\phi} + h(\phi|y) \end{aligned} \tag{4.47}$$

mit $h(\phi|y) = -\frac{1}{2}h_1(\phi) - h_2(y)$.
Um nun der log-Likelihood Eigenschaft bzgl. ϕ zu genügen, muss für den Quasi-Score $\mathbb{E}(\partial q^+/\partial\phi) = 0$ gelten, also

$$\begin{aligned} \mathbb{E}\left(\frac{\partial q^+(\mu, \phi|y)}{\partial\phi}\right) &= \mathbb{E}\left(\frac{1}{2}\frac{D_q(\mu|y)}{\phi^2} - \frac{1}{2}h_1'(\phi)\right) \\ &= \frac{1}{2\phi^2}\mathbb{E}(D_q(\mu|y)) - \frac{1}{2}h_1'(\phi) \overset{!}{=} 0 \end{aligned}$$

und in weiterer Folge

$$\mathbb{E}(D_q(\mu|y)) = \phi^2 h_1'(\phi). \tag{4.48}$$

Eine Taylorentwicklung zweiten Grades für $D_q(\mu|y)$ führt zur Approximation

$$\mathbb{E}(D_q(\mu|y)) \approx \phi$$

(vgl. THALER, 2009). Aus (4.48) ergibt sich damit $h_1(\phi) = \log(\phi) + const$ und die EQL-Funktion ist somit gegeben als

$$q^+(\mu, \phi|y) = -\frac{1}{2}\frac{D_q(\mu|y)}{\phi} - \frac{1}{2}\log(\phi). \tag{4.49}$$

Diese Version unterscheidet sich von (4.45) nur durch die additive Konstante $-\frac{1}{2}\log(2\pi V(y))$.

Bezieht man auch höhere Momente in die Berechnung der Taylorentwicklung mit ein (vgl. MCCULLAGH UND NELDER, 1989 bzw. THALER, 2009), so erhält man durch eine Sattelpunkt-Approximation die in der Definition beschriebene Darstellung (4.45).

NELDER UND LEE (1992) definierten auch eine extended Quasi-Deviance D^+ als

$$D^+(\boldsymbol{\mu}|\mathbf{y}) = -2q^+(\boldsymbol{\mu}, \phi|\mathbf{y}). \tag{4.50}$$

4.5.2 Parameterschätzung

Da die EQL-Funktion $q^+(\boldsymbol{\mu}, \phi|\mathbf{y})$ eine lineare Funktion der QL-Funktion $q(\boldsymbol{\mu}|\mathbf{y})$ bzw. deren Deviance und der erste Term von $\boldsymbol{\mu}$ bzw. $\boldsymbol{\beta}$ unabhängig ist, sind die Schätzer des Parametervektors $\boldsymbol{\beta}$, die man durch Maximierung von $q^+(\boldsymbol{\mu}, \phi|\mathbf{y})$ nach $\boldsymbol{\beta}$ erhält, ident mit den Maximum-Quasi-Likelihood Schätzern von $q(\boldsymbol{\mu}|\mathbf{y})$.

Für die Spezialfälle der Korrespondenz von $q^+(\boldsymbol{\mu}, \phi|\mathbf{y})$ zu einer Normal- oder Invers-Gauß-Verteilung ist, wie auch im Quasi-Likelihood-Fall für alle Vertreter der einparametrigen Exponentialfamilie, eine Äquivalenz von EQL- und log-Likelihood Funktion gegeben, somit stimmt auch $\hat{\phi}$ mit dem MLE von ϕ überein (vgl. NELDER UND PREGIBON, 1987). Bei vorliegender Gammaverteilung unterscheidet sich die EQL-Funktion von der log-Likelihood-Funktion durch einen Faktor, der nur von ϕ abhängt. Wobei im Falle der Poisson-, Binomial- und Negativ-Binomial-Verteilungen die korrespondierende EQL-Funktion durch Ersetzen aller k-Faktoriellen mit deren Stirling-Approximation

$$k! \approx \sqrt{2\pi k}\; k^k e^{-k}$$

generiert werden kann.

Durch Maximierung der EQL-Funktion nach ϕ erhält man auch einen Schätzer für den Dispersionsparameter

$$\hat{\phi} = \frac{1}{n} D_q(\hat{\boldsymbol{\mu}}|\mathbf{y}), \tag{4.51}$$

die mittlere Quasi-Deviance.

Im Folgenden wird nun die Berechnung der Schätzer von ϕ und dem möglicherweise vorliegenden unbekannten Varianzparameter θ näher betrachtet.

Modelle mit unbekannten Parametern in der Varianzfunktion
Mittels der EQL-Funktion können, wie schon erwähnt, Modelle mit unbekanntem Varianzparameter θ gefittet werden. Dazu wird angenommen, dass die vorliegende Varianzfunktion aus einer parametrisierten Familie $\mathcal{F}_\theta$, zum Beispiel der Exponentialfamilie, stammt mit $\text{Var}(y) = \phi V_\theta(\mu) = V(\mu, \theta)$. Die zugehörige EQL-Funktion einer einzelnen Beobachtung lautet somit

$$q_\theta^+(\mu, \phi|y) = -\frac{1}{2} \log\left[2\pi\phi V_\theta(y)\right] - \frac{1}{2\phi} D_\theta(y|\mu), \tag{4.52}$$

wobei D_θ gegeben ist durch

$$D_\theta(y|\mu) = -2 \int_y^\mu \frac{y - \mu}{V_\theta(u)}\, du. \tag{4.53}$$

Um Inferenzaussagen über θ treffen zu können, wird von Nelder und Pregibon (1987) die Verwendung von *Profile-Quasi-Likelihood-Intervallen* vorgeschlagen. Diese Intervalle erhält man durch Maximierung von q_θ^+ mit festgehaltenem θ, was zu den Schätzern $\hat{\boldsymbol{\beta}}(\theta)$ und $\hat{\phi}(\theta)$ führt. Dann variiert man θ über einem zuvor festgelegten Intervall Θ. Hierbei sei q_{max}^+ das Maximum aller zuvor berechneten Werte, also

$$q_{max}^+ = \max_{\theta \in \Theta} q_\theta^+. \tag{4.54}$$

Das Profile-Quasi-Likelihood-Intervall (θ_L, θ_R) entspricht dann jenem Intervall, welches garantiert, dass

$$q_{PL}(\theta) := q_{max}^+ - q_\theta^+ > c, \tag{4.55}$$

wobei c beispielsweise einem $\frac{1}{2}\chi_1^2$-Quantil entspricht. Sei θ^* jener Wert, der q_{max}^+ maximiert, so beschreiben $\hat{\boldsymbol{\beta}}(\theta^*)$ und $\hat{\phi}(\theta^*)$ die gewünschten endgültigen Schätzer der Modellparameter $(\boldsymbol{\beta}, \phi)$. Nelder und Pregibon (1987)

schlagen zur effizienteren Berechnung des Profile-Quasi-Likelihood-Intervalls auch einen Bootstrap-Ansatz vor.

THALER (2009) beschreibt die zur Modellfindung annehmbaren Varianzfamilien genauer. Häufig Anwendung findet der Potenzansatz $V_\theta(\mu) = \mu^\theta$.

Modelle mit nicht konstantem Dispersionsparameter

In Quasi-Likelihood-Modellen spielte die Annahme eines konstanten Dispersionsparameters ϕ eine zentrale Rolle. In manchen Anwendungen kann es aber von Vorteil sein, ϕ als Funktion von Kovariablen zu modellieren. Man betrachtet also Modelle für $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$. Hierzu wird folgendes Modell betrachtet (vgl. NELDER UND PREGIBON, 1987)

$$\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu},\ \mathrm{Var}(\mathbf{y}) = \mathrm{diag}(\phi_1, \ldots, \phi_n)V(\boldsymbol{\mu}),\ g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},\ h(\boldsymbol{\phi}) = \mathbf{Z}\boldsymbol{\beta}^*,$$

wobei $h()$ die Linkfunktion des Dispersionsparameters darstellt und $\boldsymbol{\beta}^*$ genau wie $\boldsymbol{\beta}$ unbekannt ist. Eine typische Wahl für $h()$ ist der identische oder der log-Link. Wie beim Standard-QL-Ansatz wird auch hier vorausgesetzt, dass keine funktionale Abhängigkeit zwischen $\boldsymbol{\mu}$ und $\boldsymbol{\phi}$ herrscht (die gesamte Abhängigkeit der Varianz von $\mathbf{y}$ von $\boldsymbol{\mu}$ wird durch $V(\boldsymbol{\mu})$ widergespiegelt). Wird also dieselbe Designmatrix für die Modellierung von Mean und Dispersion verwendet ($\mathbf{X} = \mathbf{Z}$), kann dies zu ungewollten Abhängigkeiten führen.
Da es sich hier offensichtlich um zwei Sub-Modelle handelt, eines für den Mean und eines für den Dispersionsparameter, wird zur Schätzung des oben genannten gemeinsamen Modells eine Sägezahn-Iteration angewandt. Hierzu wird zuerst $\boldsymbol{\phi}$ am jeweiligen Schätzer $\hat{\boldsymbol{\phi}}$ festgehalten und der Parameterschätzer $\boldsymbol{\beta}$ standardmäßig berechnet ($\hat{\boldsymbol{\phi}}$ fließt als Gewicht in die Berechnung mit ein). NELDER ET AL. (1998) nehmen als Startwert $\boldsymbol{\phi} = \mathbf{1}$ an. Im zweiten Schritt wird nun $\boldsymbol{\mu}$ am aktuellen Schätzer $\hat{\boldsymbol{\mu}}$ fixiert, um das QL-Modell des Dispersionsparameters zu fitten. Zum besseren Verständnis wird dieses Modell genauer formuliert (vgl. NELDER UND PREGIBON, 1987). Sei hierzu $\mathbf{d} = (d_1, \ldots, d_n)^T$ mit $d_i = D(\hat{\mu}_i | y_i)$, dann ergibt sich folgende Formulierung,

$$\mathbb{E}(\mathbf{d}) = \boldsymbol{\phi},\ \mathrm{Var}(\mathbf{d}) = \psi V_\phi(\boldsymbol{\phi}),\ h(\boldsymbol{\phi}) = \mathbf{Z}\boldsymbol{\beta}^*. \tag{4.56}$$

Im Dispersionsmodell werden also die Komponenten der Deviance (erhalten aus dem vorhergehenden Schritt) als Response und $V_\phi(\phi_i) = \phi_i^2$ angenommen. Die abwechselnde Iteration dieser zwei Schritte führt schlussendlich zu Schätzern $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^*)$, welche die QL-Funktion $q^+(\boldsymbol{\mu}, \boldsymbol{\phi}|\mathbf{y})$ maximieren. Laut NELDER ET AL. (1998) führen in den meisten Fällen schon vier bis fünf Iterationen zur gewünschten Konvergenz.

EQL-AIC

Um nun verschiedene Modelle vergleichen zu können, wurde von HURVICH UND TSAI (1995) auch für die Familie der EQL-Modelle ein AIC entwickelt:

$$AIC_{EQL} = n\left[\log(\hat{\phi}) + 1\right] + \frac{2n(p+1)}{n-p-2}, \tag{4.57}$$

wobei $\hat{\phi} = \frac{1}{n} D_q(\hat{\boldsymbol{\mu}}|\mathbf{y})$.

5 Praktische Anwendung für Verteilung von Sporenkonzentrationen

Dieses Kapitel beschäftigt sich nun eingehend mit der Umsetzung der vorab vorgestellten theoretischen Grundlagen. Hierzu werden Daten aus dem Bereich der Mikrobiologie herangezogen, die vom Institut für Hygiene, Mikrobiologie und Umweltmedizin der Universität Graz bereitgestellt wurden[3]. Beide Beispiele behandeln die Abhängigkeit von Sporenkonzentrationen in Innenräumen bzw. Weinkellern von weiteren Messparametern. Die Bestimmung der Sporenkonzentration, gemessen in Anzahl koloniebildender Einheiten pro Kubikmeter Luft (KBE/m^3), wird dabei mit zwei verschiedenen Messgeräten vorgenommen, nämlich einerseits mit dem Andersen® ACFM Kaskaden-Impaktor und andererseits dem MAS100® Merck Air Sampler.

ACFM Kaskaden-Impaktor	**MAS100 Air Sampler**
Ein sechsstufiger Multi-Düsen-Impaktor (Staubmessgerät), mit dem Partikel in der Luft (Bakterien und Pilze) gemessen werden können. Der Durchmesser der Düsenöffnungen nimmt von Stufe 1 bis 6 (von oben nach unten) kontinuierlich ab, wobei jede Stufe aus einer Siebplatte bestückt mit einem Nährboden besteht. Die 6 Stufen des ACFM erlauben die Messungen von unterschiedlichsten Partikelgrößen.	Ein einstufiger Impaktor bestückt mit einer Siebplatte und darunter liegendem Nährboden. Der Luftdurchsatz bei diesem Gerät ist weitaus höher als beim ACFM, jedoch ist die messbare Partikelgröße eingeschränkter.
	

Quelle: http://www.hygiene-graz.at/

[3]http://www.hygiene-graz.at/, Projektleitung: Mag. Dr. Doris Haas

Zudem wurden unabhängig voneinander zwei verschiedene Nährmedien pro Messgerät verwendet:

- Dichloran Glycerol Agar (DG18)
- Malzextrakt Agar (MEA).

Der Nährboden DG18 dient vorrangig der Feststellung von trockenheitliebenden (xerophilen) Schimmelpilzen und MEA begünstigt die Anzucht von mesophilen Schimmelpilzarten, d.h. jene, die in einem mittleren Temperaturbereich bevorzugt anwachsen. Die Proben werden bei $25°C$ 7-10 Tage lang gelagert und danach die Sporenkonzentration gruppiert nach den vorherrschenden Pilzgattungen bestimmt.

In den folgenden Anwendungsbeispielen wird die Summe aller Keime zur Modellierung herangezogen und nicht nach einzelnen Pilzgattungen klassifiziert. Die statistische Auswertung der Daten erfolgte mit der freien Statistiksoftware *R Version 3.0.1 (2013-05-16)*[4]. Eine gute Einführung in das Arbeiten mit R bietet Crawley (2007), eine Vertiefung in den Bereich der statistischen Modellierung von Daten ist bei Aitkin, Francis und Hinde (2009) zu finden. Die in diesem Kapitel folgenden Plots wurden größtenteils mit Hilfe des Grafikpakets *ggplot2* erstellt, welches in Wickham (2009) ausführlich beschrieben wird.

5.1 Weinkellerstudie

Im Rahmen dieser Studie wurden Luftkeimmessungen in Weinkellern von Februar bis September bei 20 österreichischen Weinbauern durchgeführt. Die Keller unterscheiden sich grob im Baustoff (Ziegel- oder Betonkeller), im Verwendungstyp (Produktions- oder Lagerkeller) und in der Klimatisierung (Klimaanlage: ja oder nein). Weiters wurden Temperatur und Luftfeuchtigkeit im Keller und außerhalb bzw. eine etwaige Taupunktunterschreitung gemessen, sowie das Baujahr (klassifiziert in 3 Kategorien: vor 1950, 1950-1979 und ab 1980) und der Schimmelpilzbefall an Wänden, Decke und Boden (leicht: $< 20\%$, mittel: $20 - 80\%$ bzw. schwer: $> 80\%$) festgehalten. Zu jeder Innen-Keimmessung wurde auch eine Vergleichsmessung in der Außenluft durchgeführt, um die Sporenbelastung und deren mögliche Auswirkungen auf Mensch und Wein im Kellerinneren abschätzen zu können. Mikrobiologische Details dieser Studie sind in Galler (2011) nachzulesen.

[4] http://www.r-project.org/

5.1.1 Explorative Datenanalyse

Um schon vorab ein Gefühl für die richtige Verteilungsannahme zu bekommen, werden zunächst die Varianz-Erwartungswert-Verhältnisse der jeweiligen Innenmessungen betrachtet.

	$Var/Mean$	$Var/Mean^2$	$Var/Mean^3$
$ACFM\ DG$	4273.289	1.631069	0.000622
$ACFM\ MEA$	3953.549	1.591069	0.000640
$MAS\ DG$	2226.812	1.122913	0.000566
$MAS\ MEA$	1923.439	1.208089	0.000758

Die Annahme einer Poissonvarianz (1. Spalte) scheint bei allen vier Messgruppen wenig sinnvoll und auch eine quadratische Varianzannahme (2.Spalte) ist zumindest beim ACFM wahrscheinlich nicht haltbar. Der weit unter 1 liegende Quotient von Varianz zu kubischem Mittel lässt wiederum bei Annahme von $\mathrm{Var}(\mu) = \mu^3$ auf Unterdispersion schließen.
Die nachfolgenden Plots zeigen einerseits die Verteilungen der Innen- und Außenmessungen der Sporenkonzentrationen (Abb. 5.1 und 5.2) und andererseits die Abhängigkeit der vier Innenmessungen von den einzelnen Haupteffekten (Abb. 5.3 bis 5.14). Im Weiteren konzentriert sich diese Arbeit nur auf die Modellierung der Sporenkonzentration in der Innenluft der verschiedenen Weinkeller ohne Miteinbeziehung der Sporenkonzentrationen der Außenluft.

ACFM:

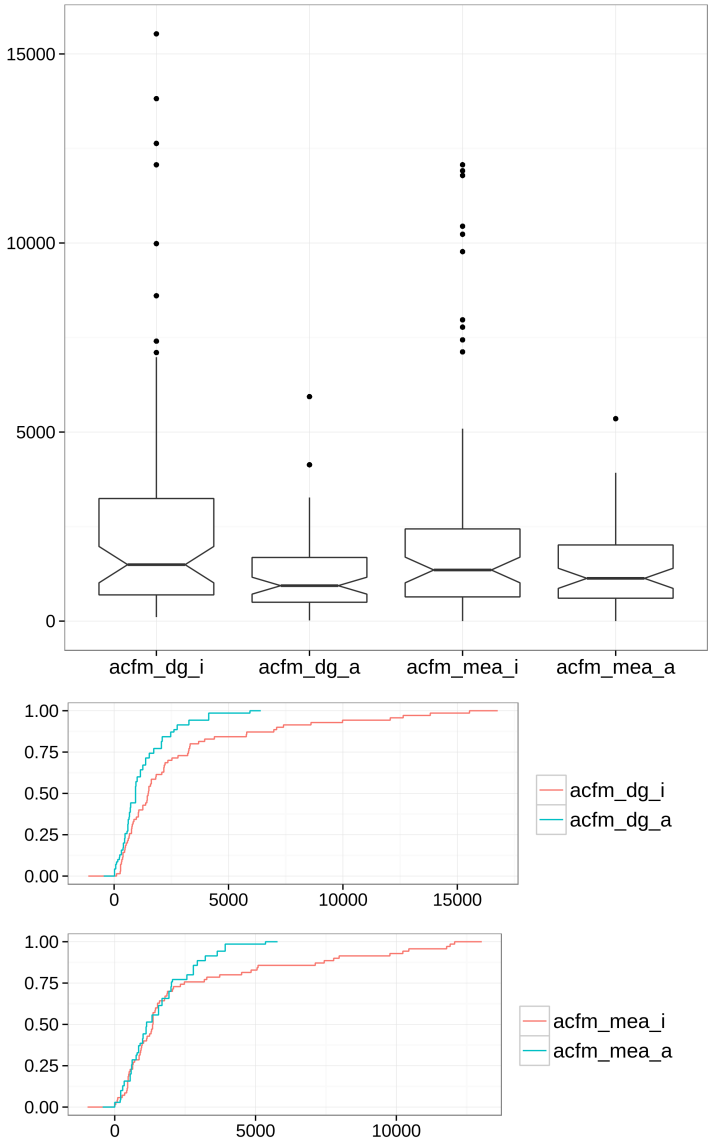

Abb. 5.1: *oben*: Boxplots der Innen- und Außenmessungen auf den zwei Nährböden. Grundsätzlich ist zu erkennen, dass sich die Mediane der Sporenkonzentration innen und außen mit MEA nicht signifikant ($\alpha = 0.05$) unterscheiden. Xerophile Pilze scheinen im Median im Kellerinneren häufiger vorzukommen als an der Außenluft.
unten: empirische Verteilungsfunktionen der DG18- und MEA-Messungen gruppiert nach Innen- und Außenmessungen. Die Verteilung der Außenmessungen ist im Allgemeinen stochastisch kleiner als jene der Innenmessungen, unabhängig vom Nährboden.

MAS100:

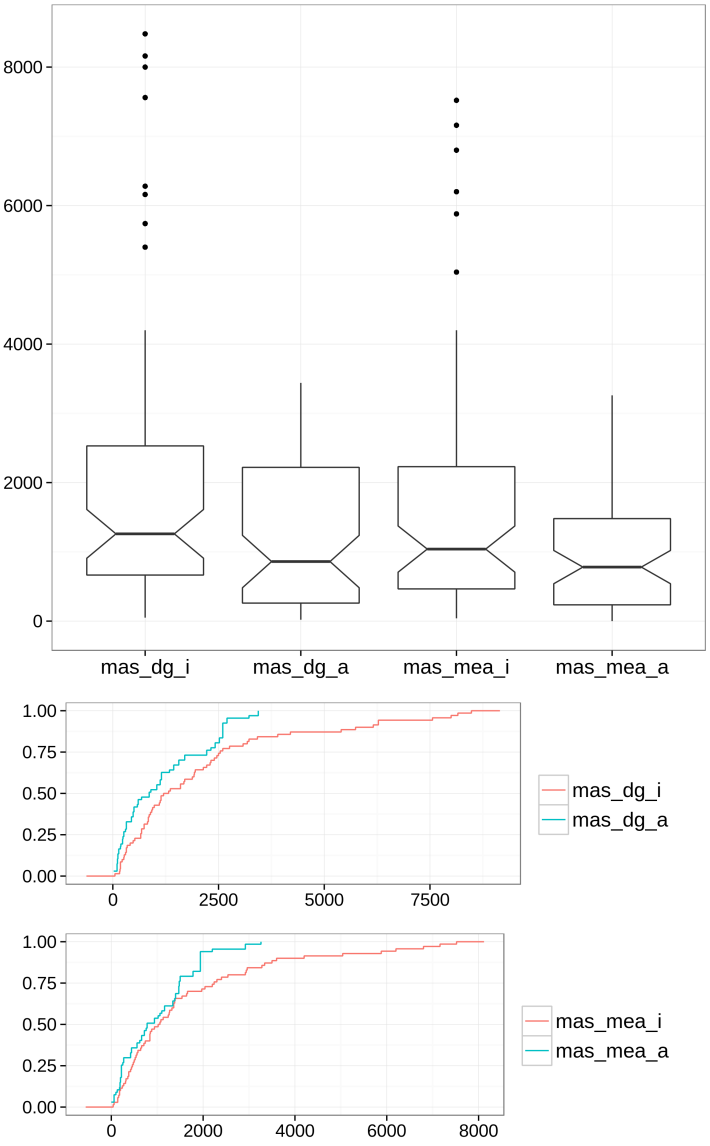

Abb. 5.2: *oben*: Boxplots der Innen- und Außenmessungen auf den zwei Nährböden. Auch bei diesem Messgerät unterscheiden sich die Mediane der Sporenkonzentration innen und außen bei MEA nicht signifikant ($\alpha = 0.05$).
unten: empirische Verteilungsfunktionen der DG18- und MEA-Messungen gruppiert nach Innen- und Außenmessungen. Die Verteilung der Außenmessungen ist im Allgemeinen stochastisch kleiner als jene der Innenmessungen, unabhängig vom Nährboden.

ACFM mit DG18: Faktorielle Messgrößen

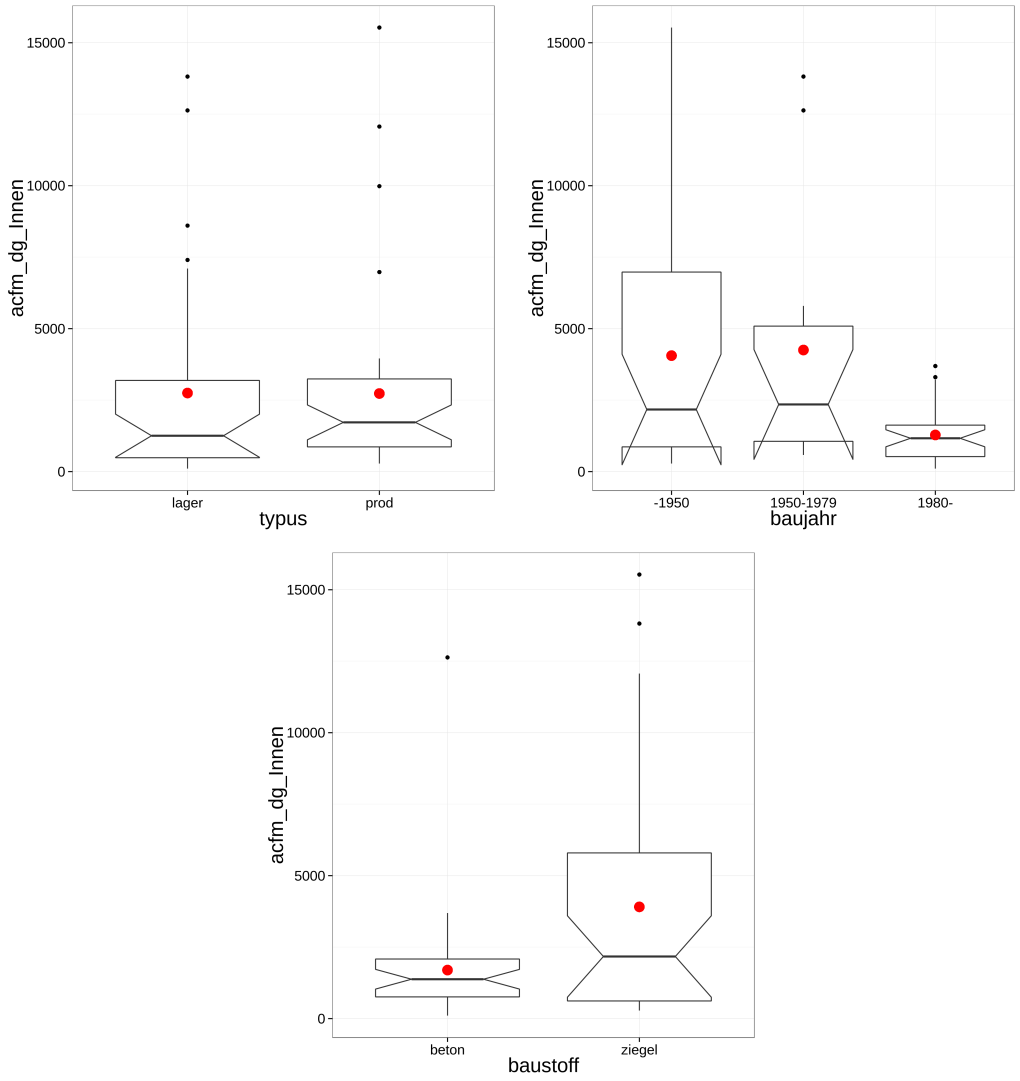

Abb. 5.3: Abhängigkeit der Sporenkonzentration von jeweils Kellertypus, Baujahr und Baustoff dargestellt durch Boxplots, der rote Punkt kennzeichnet den Mittelwert. Im Median scheint ein Ziegelkeller und auch vor 1980 erbaute Keller höhere Sporenkonzentrationen zu begünstigen.

ACFM mit DG18: Faktorielle Messgrößen

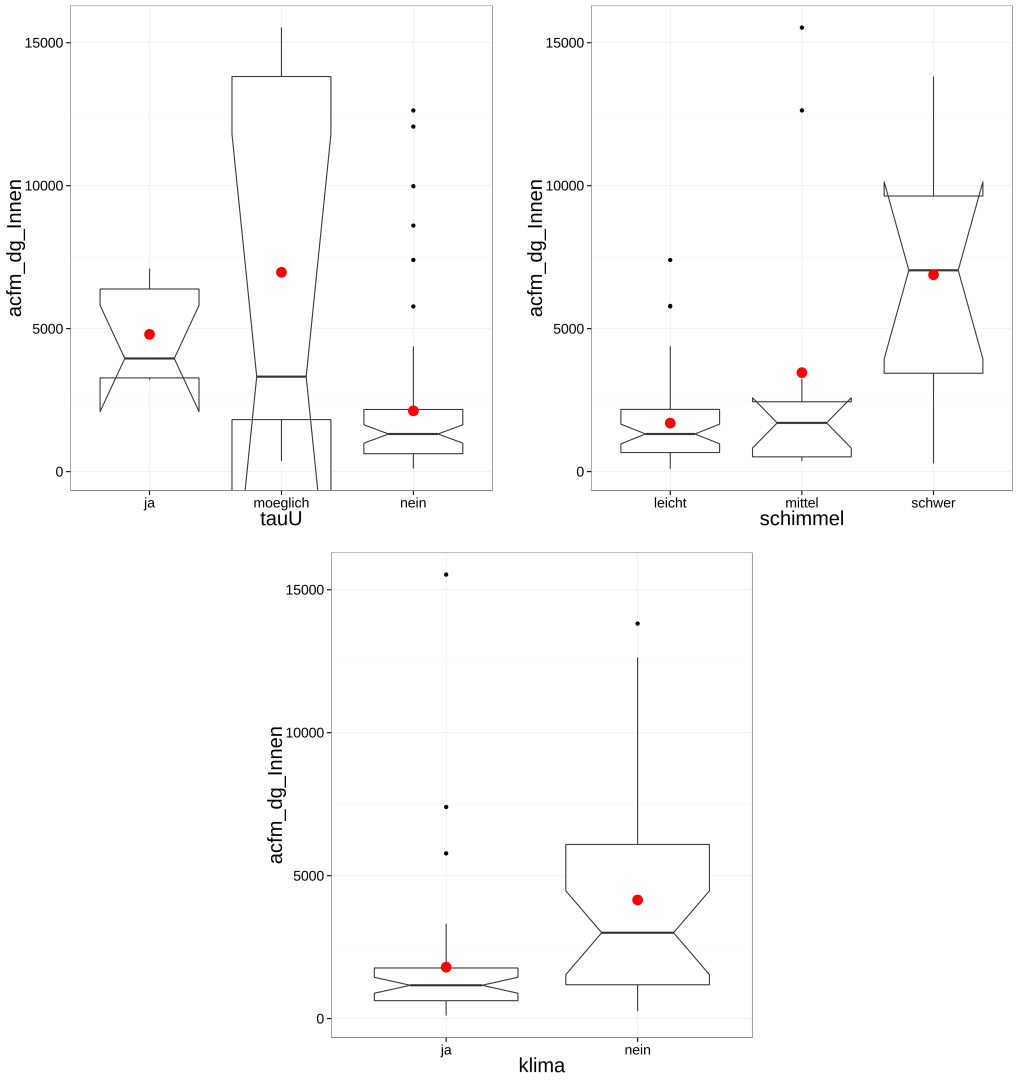

Abb. 5.4: Abhängigkeit der Sporenkonzentration von jeweils Kellertypus, Baujahr, Baustoff, Taupunktunterschreitung, Schimmelbefall und Klimatisierung dargestellt durch Boxplots. Großen Einfluss scheint vorallem der vorliegende Schimmelbefall sowie die Klimatisierung zu haben.

AFCM mit DG18: stetige Messgrößen

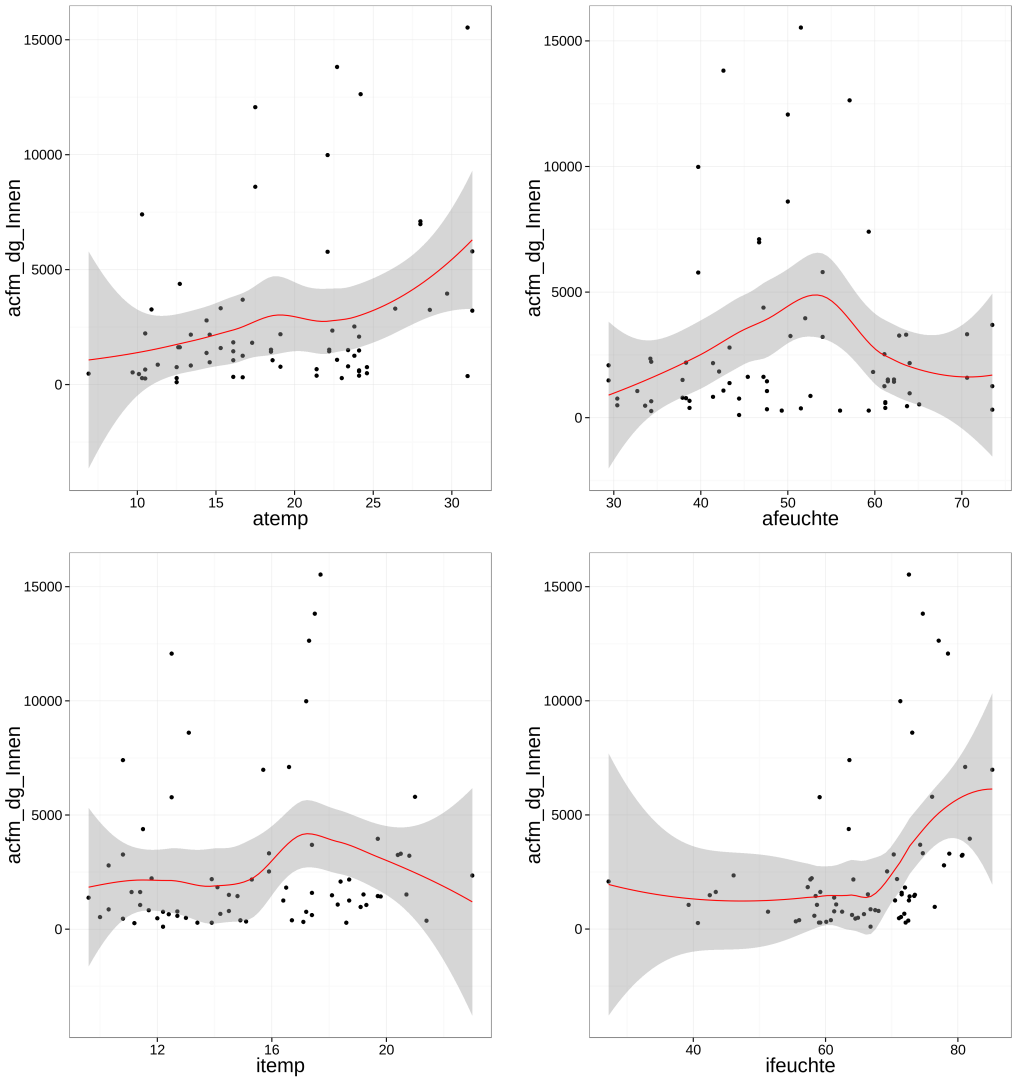

Abb. 5.5: Scatterplot der DG18-Innenmessung mit Außentemperatur, Außenfeuchtigkeit, Innentemperatur und Innenfeuchtigkeit. Die rote Linie beschreibt eine loess-Glättungsfunktion (local polynomial regression) mit quadratischen Polynomen. Außentemperatur zeigt hierbei einen linearen Trend, wohingegen die Abhängigkeit zur Außenfeuchte bzw. Innentemperatur eher quadratischen Charakter hat.

ACFM mit MEA: Faktorielle Messgrößen

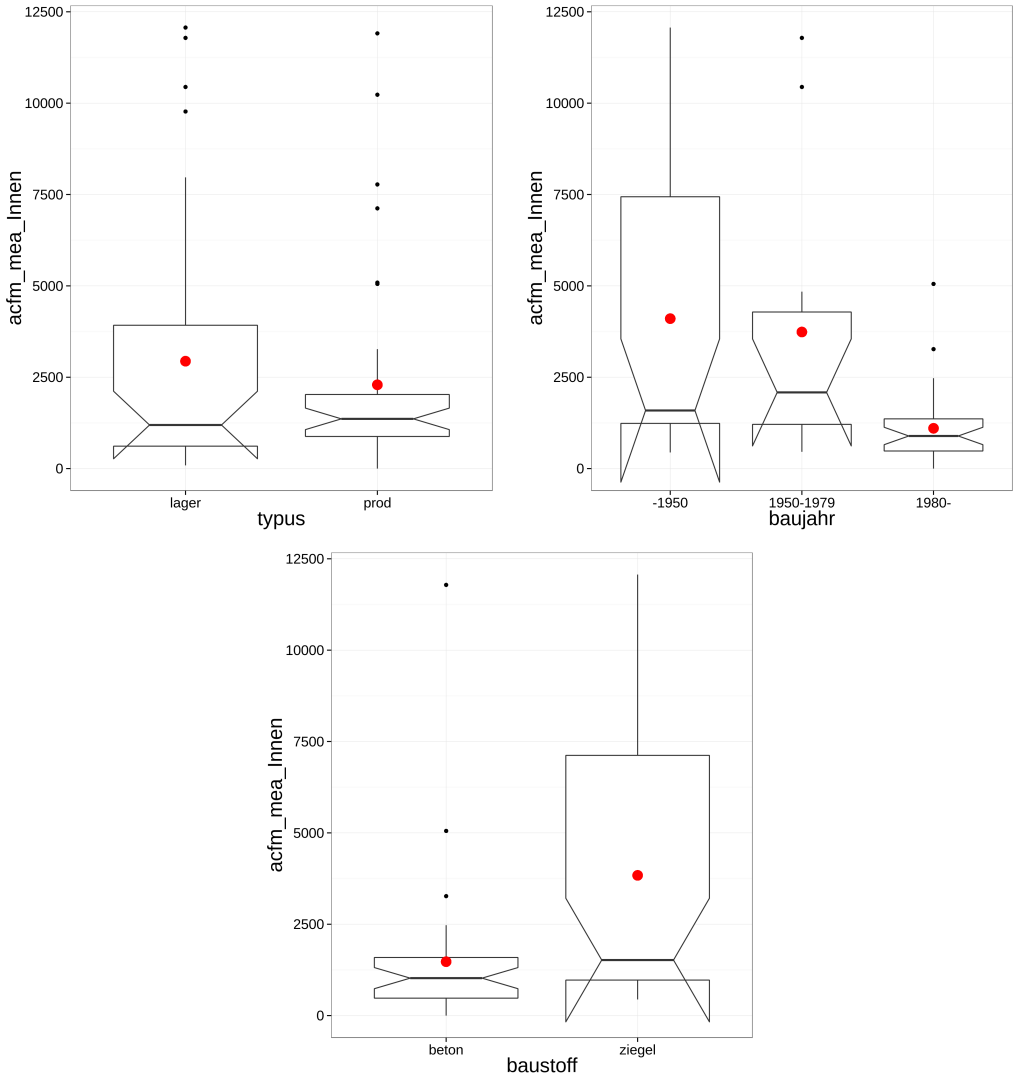

Abb. 5.6: Auch bei mesophilen Pilze zeigt der Baustoff und das Alter des Kellers einen signifikanten Einfluss auf zumindest den Median der Sporenkonzentration. Weiters ist auch eine stärkere Streuung bei Lagerkellern zu erkennen.

ACFM mit MEA: Faktorielle Messgrößen

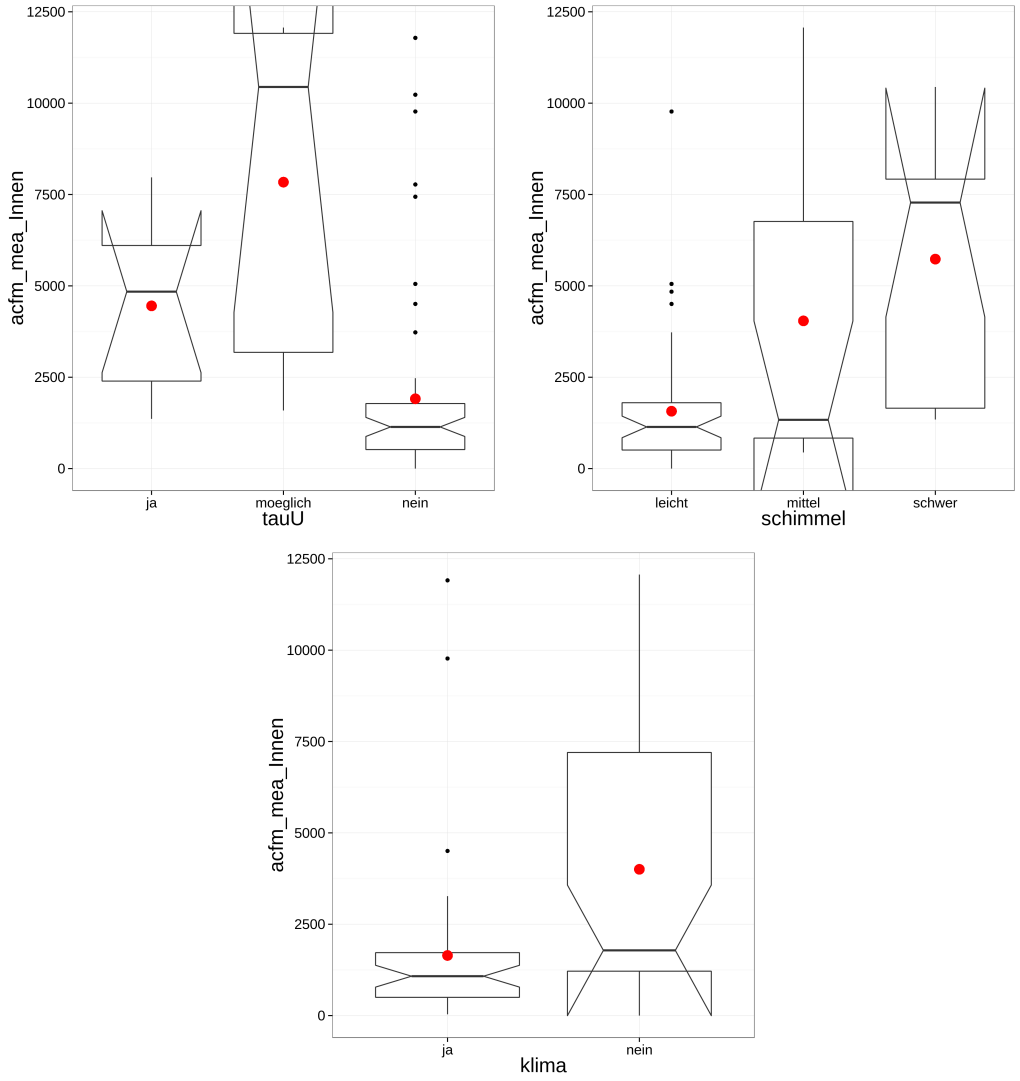

Abb. 5.7: Großen Einfluss scheint auch bei mesophilen Pilzen ein vorliegender Schimmelbefall, eine Taupunktunterschreitung sowie die Klimatisierung zu haben.

ACFM mit MEA: stetige Messgrößen

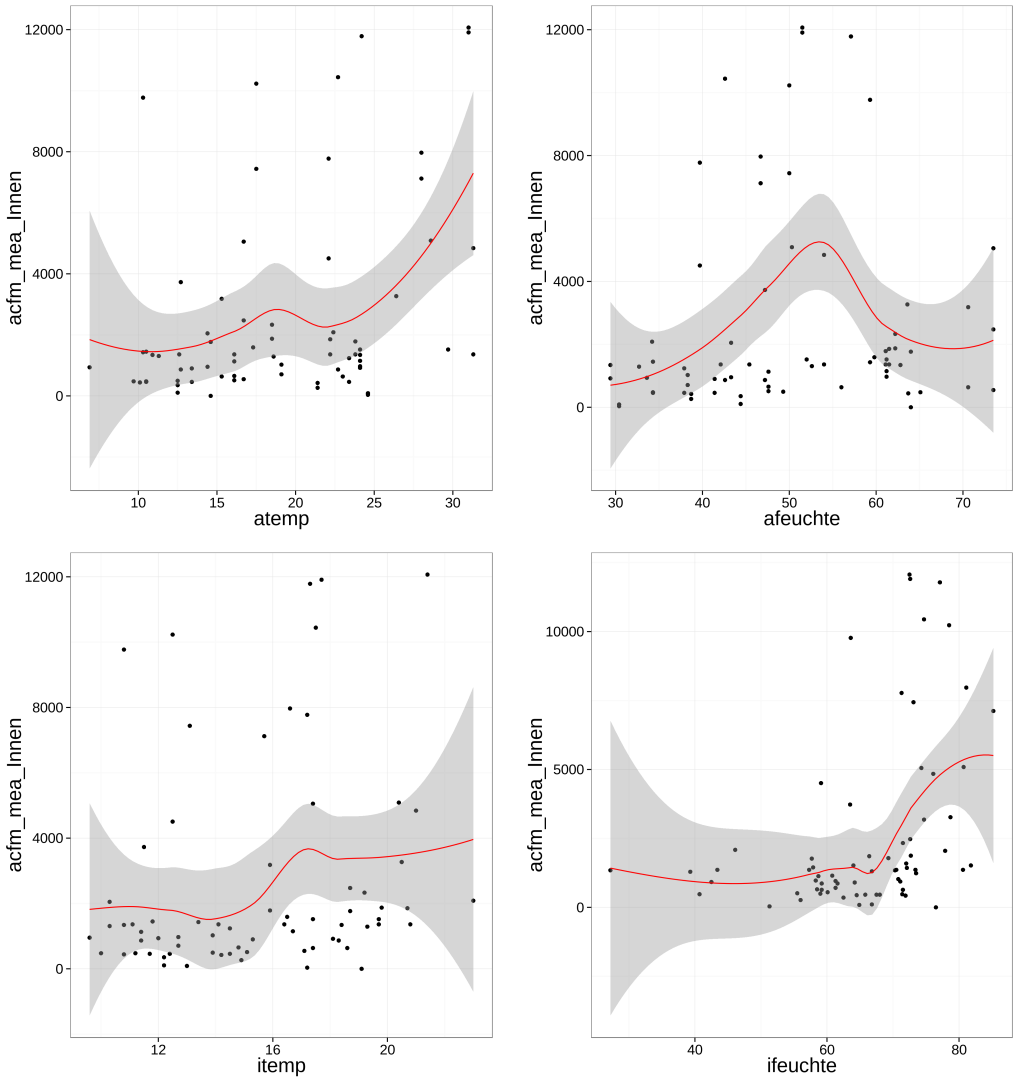

Abb. 5.8: Scatterplot inkl. loess-Glättungsfunktion. Hier zeigen Außen- und Innentemperatur einen eher linearen Trend, die Außenfeuchte hat wieder quadratischen Charakter.

MAS100 mit DG18: Faktorielle Messgrößen

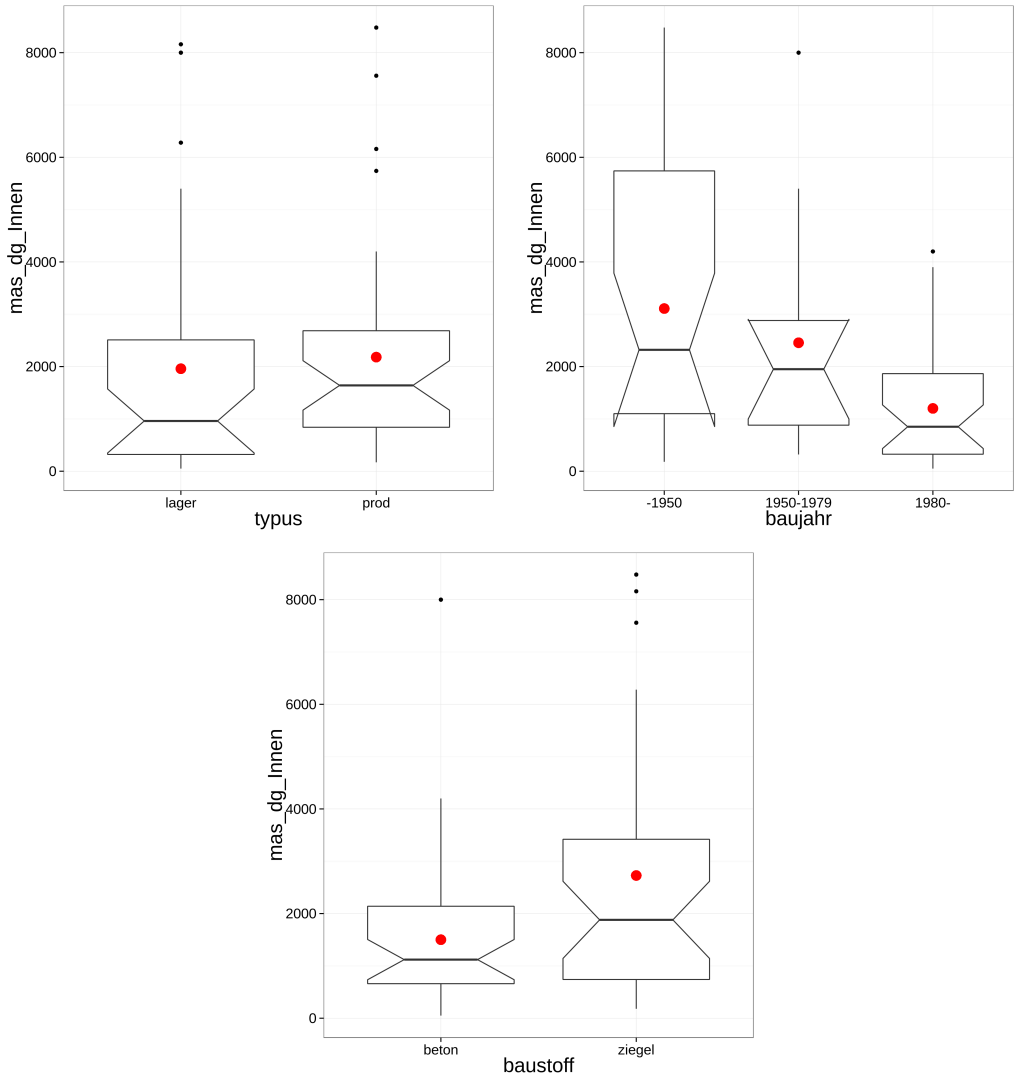

Abb. 5.9: Das Baujahr weist, wie auch schon beim ACFM, einen linearen Trend auf bzw. bei Ziegelkellern eine im Median höhere Sporenkonzentration als bei Betonbauweise.

MAS100 mit DG18: Faktorielle Messgrößen

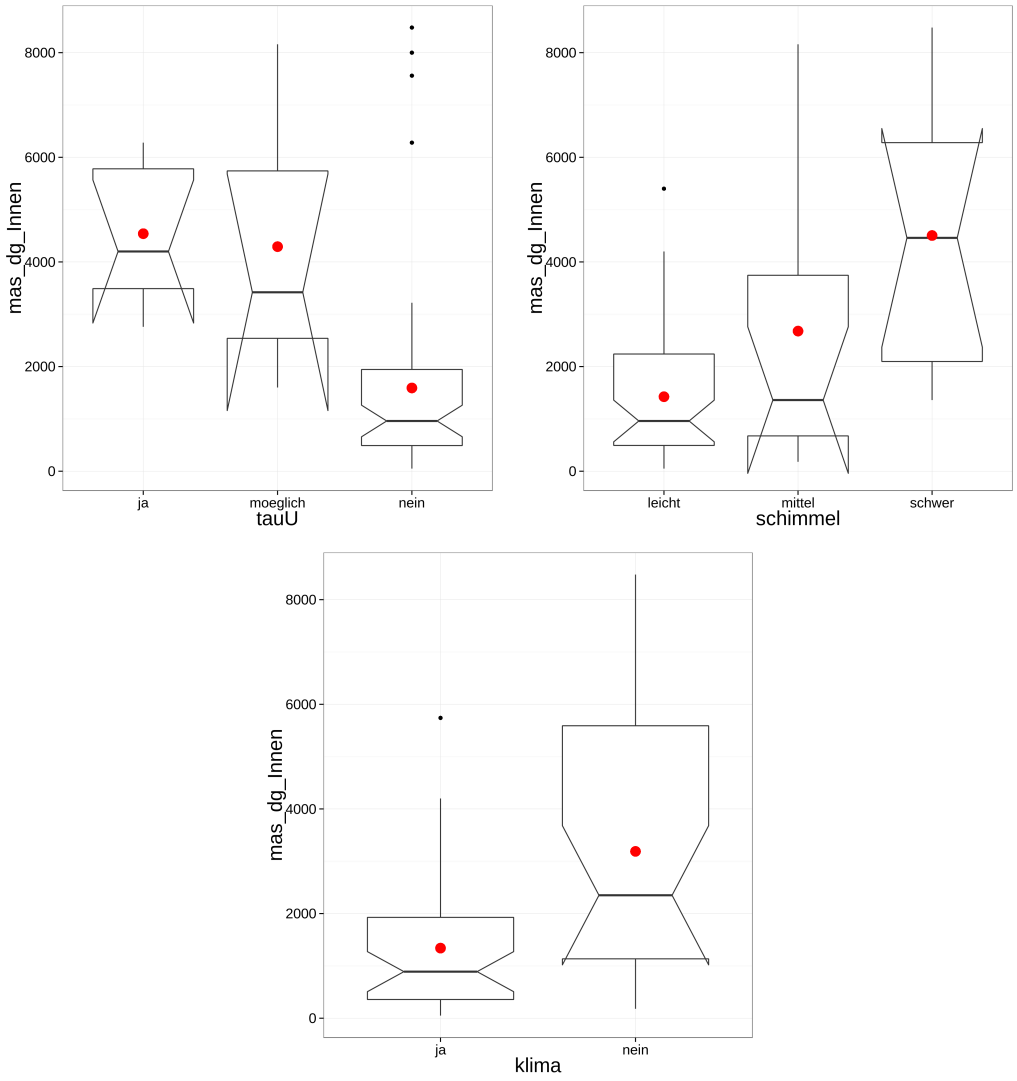

Abb. 5.10: Großen Einfluss hat auch bei diesem Messgerät der vorliegende Schimmelbefall, die Klimatisierung, sowie die Taupunktunterschreitung.

MAS100 mit DG18: stetige Messgrößen

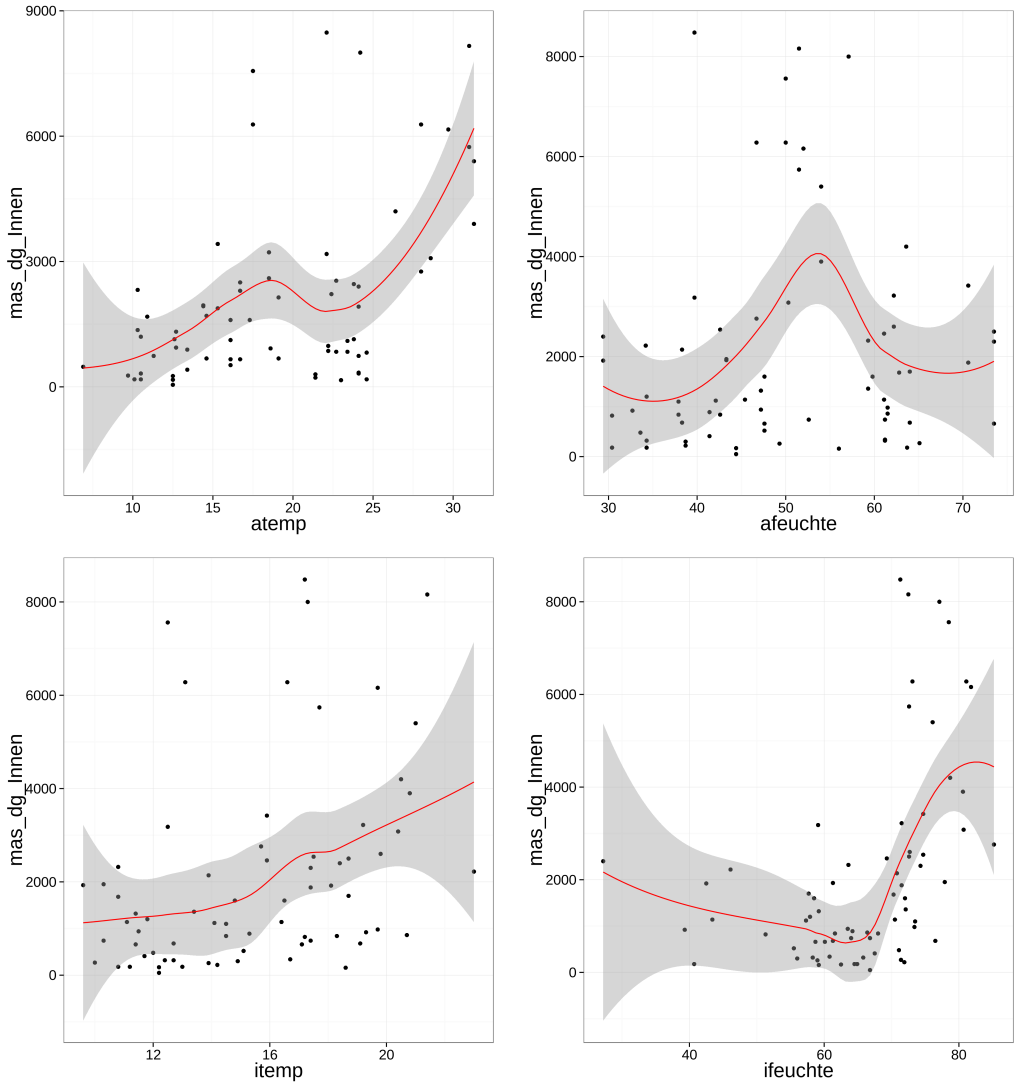

Abb. 5.11: Außen- und Innentemperatur zeigen einen eher linearen Trend, wohingegen die Feuchtemessungen auch hier quadratischen Charakter hat.

MAS100 mit MEA: Faktorielle Messgrößen

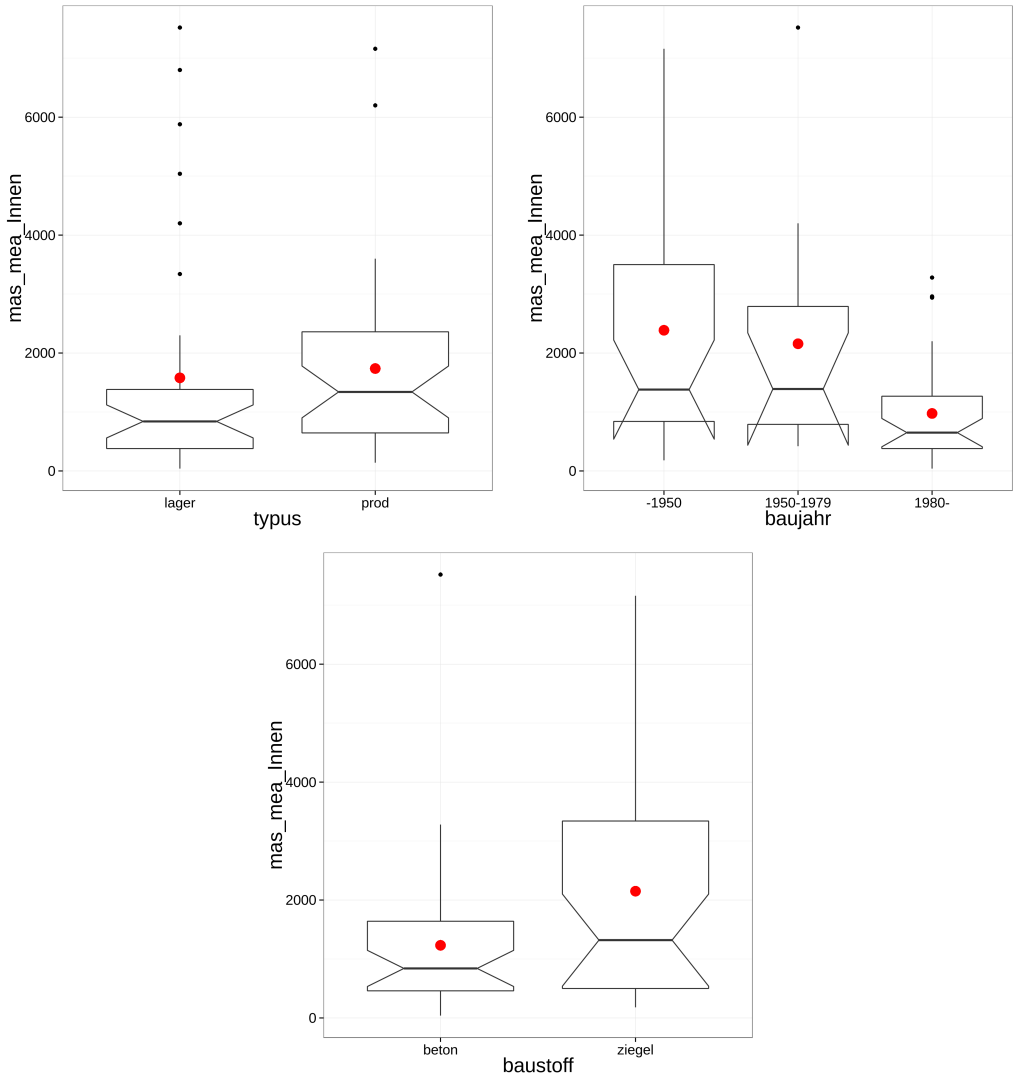

Abb. 5.12: Bei mesophilen Pilzgattungen scheint auch bei diesem Messgerät der vorliegende Typ, das Alter und der Baustoff Einfluss auf die Sporenbelastung zu haben.

MAS100 mit MEA: Faktorielle Messgrößen

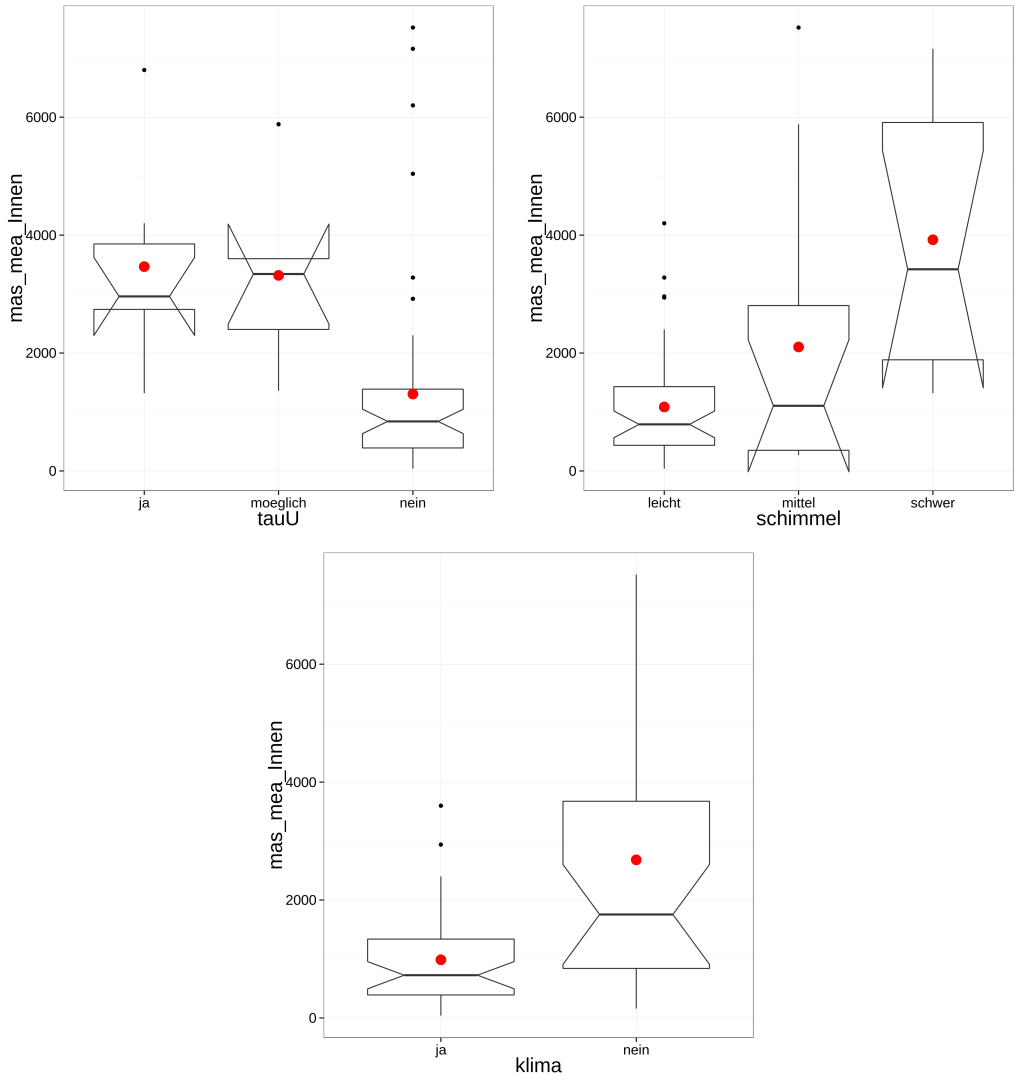

Abb. 5.13: Signifikant auf den Median wirkt sich auch der vorliegende Schimmelbefall, die Klimatisierung, sowie die Taupunktunterschreitung aus.

MAS100 mit MEA: stetige Messgrößen

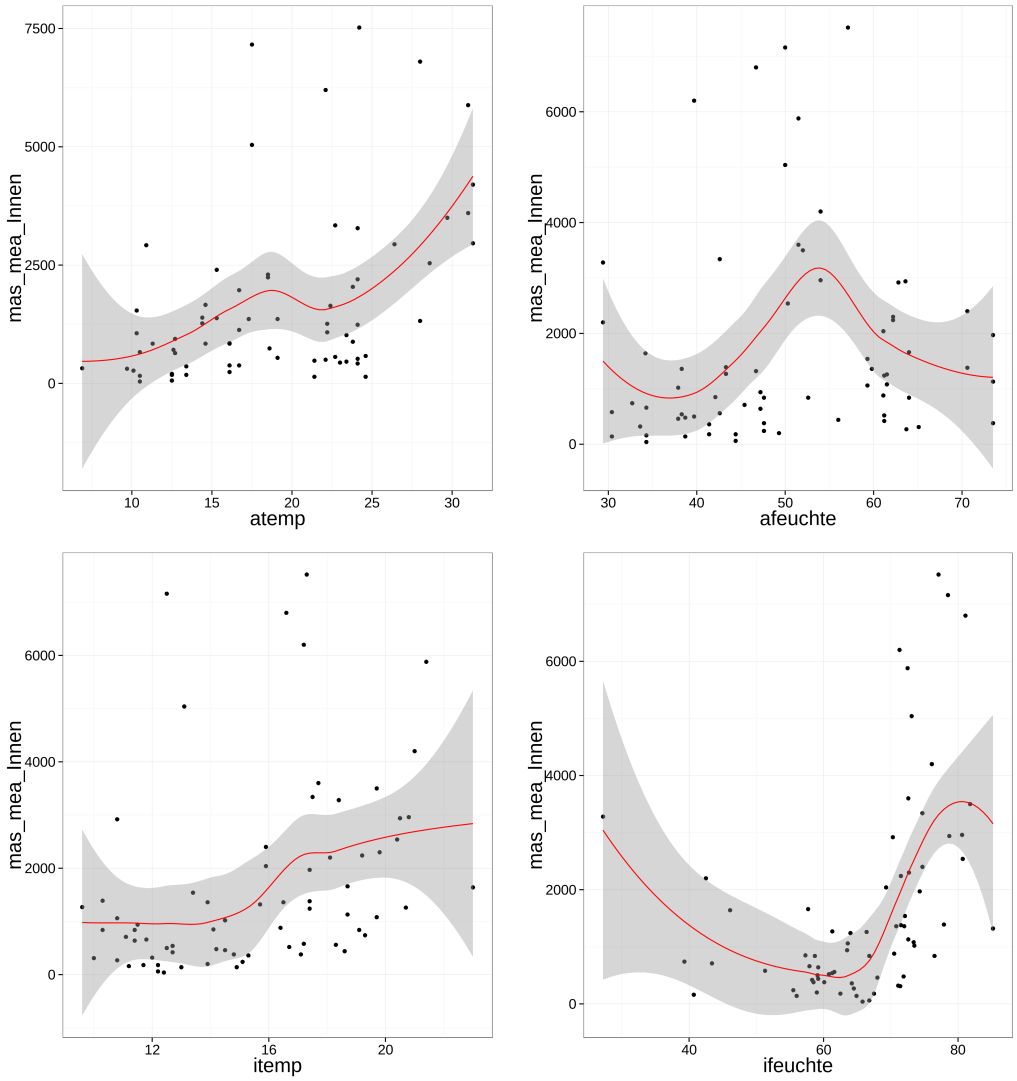

Abb. 5.14: Außen- und Innentemperatur sowie Taupunkt zeigen einen eher linearen Trend, wohingegen die Außenfeuchte wieder quadratischen Charakter hat. Innenfeuchte scheint zumindest quadratisch einzufließen.

5.1.2 Regressionsmodelle für ACFM mit DG18

Da es sich bei den gegebenen Responses (Sporenkonzentrationen in KBE/m^3 Luft) um Zählvariablen handelt, werden die Untersuchungen mittels einem Poisson-Modell mit log-Link gestartet (auch wenn die vorab berechnete em-

pirische Varianz der Daten beträchtlich größer ist als ihr Mittelwert, dient dieser Schritt dem besseren Verständnis).

```
> y<-acfm_dg_i
> family<-poisson(link="log")
```

Da Pearson X^2-Statistik standardmäßig nicht in R implementiert ist, wird diese Funktion selbst definiert als

```
X2<-function(model){
  sum(residuals(model,type="pearson")^2)}
```

und in weiterer Folge auch zur Bewertung der Modelle verwendet.

Wegen der geringen Datenmenge von 70 Datenpunkten bei 14 Haupteffekten ist ein volles Modell (mit Interaktionen bis zum Grad 14) nicht modellierbar und auch Interaktionen von mehr als zwei Prädiktoren sind in der Praxis schwer zu erklären, daher beschränken sich die betrachteten GLMs auf Modelle mit maximal zweifachen Interaktionen. Im Sinne einer niedrigen Modellkomplexität wird auch auf kubische oder Terme höherer Ordnung verzichtet, was auch durch die Plots in Abschnitt 5.1.1 als logisch bestätigt wird. Die erste Modellfindung erleichtert die Funktion *step*(), welche die Variablenauswahl in GLMs durch Vergleich der AIC-Werte automatisiert (durch Setzen der Option *direction* = *„both“* wird mittels Vorwärts- und Rückwärtsselektion das optimale AIC bestimmt). Es resultiert nach einer weiteren manuellen Varianzanalyse das folgende Poisson-Modell.

```
> class(main.eff< y~typus+baujahr+baustoff+atemp+I(atemp^2)
  +afeuchte+I(afeuchte^2)+itemp+I(itemp^2)+ifeuchte+I(ifeuchte^2)
  +tauU+schimmel+klima)

> mod_pois_me<-glm(main.eff,family=poisson())
> mod_pois<-step(mod_pois_me,scope=.~.^2,direction="both",trace=T)
> summary(mod_pois.1)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.735e+02  8.929e+00  19.433  < 2e-16
typusprod                    1.459e+01  1.461e+00   9.991  < 2e-16
baujahr1950-1979            -1.522e+02  6.297e+00 -24.167  < 2e-16
baujahr>1980                -5.685e+01  5.683e+00 -10.004  < 2e-16
baustoffziegel              -4.950e+00  6.398e-01  -7.737 1.02e-14
atemp                        2.863e+00  2.984e-01   9.594  < 2e-16
I(atemp^2)                   2.468e-02  1.827e-02   1.351 0.176775
... ... ...
afeuchte:I(afeuchte^2)      -1.425e-04  1.194e-05 -11.931  < 2e-16
```

```
atemp:itemp                  -3.914e-01  1.660e-02 -23.577  < 2e-16
atemp:klimanein               3.129e-01  2.584e-02  12.111  < 2e-16
atemp:I(atemp^2)             -3.158e-03  3.396e-04  -9.298  < 2e-16
ifeuchte:I(ifeuchte^2)        4.739e-05  1.255e-05   3.777 0.000158
typusprod:afeuchte            5.167e-03  2.451e-03   2.108 0.035011
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.1231e+05  on 69  degrees of freedom
Residual deviance: 3.1326e+00  on  4  degrees of freedom
AIC: 776.16

> X2(mod_pois.1) #Pearson's X^2
[1] 3.117827
```

Das optimale Poisson-Modell hat zwar eine Deviance, die in der Größenordnung ähnlich den Freiheitsgraden ist und auch mit der Pearson-Statistik gut übereinstimmt, jedoch ist es mit 66 Prädiktoren heillos überbestimmt. Auch die folgenden Reisduenplots (Abb. 5.15) lassen am Poissonansatz zweifeln. Der sogenannte Scale-Location-Plot stellt die fitted values den standardisierten (varianzstabilisierten) absoluten Residuen gegenüber und sollte, bei korrekt spezifizierter Varianz, keinen Trend (fallend oder steigend) aufweisen. Die standardisierten Pearson- oder Deviance-Residuen r_i^{*P} bzw. r_i^{*D} berechnen sich grundsätzlich als

$$r_i^* = \frac{r_i}{\sqrt{\hat{\phi}(1-h_i)}}.$$

Hierbei ist $h_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$, das i-te Diagonalelement der Hutmatrix, und wird als *leverage* bezeichnet. Im Normalverteilungsfall entspricht $\sigma^2(1-h_i)$ der exakten Varianz der rohen Residuen. Eine automatische Berechnung der standardisierten Residuen (und noch weiterer) bietet die Funktion *glm.diag*() aus dem Paket *boot*.

Als nächstes wird versucht ein Quasi-Poisson-Modell zu fitten, d.h. es wird eine Varianzannahme $\text{Var}(y) = \phi\mu$ mit $\phi > 0$ getroffen. Da schon im Poisson-Modell erkannt wurde, dass für ein vollständiges quadratisches Modell zu wenig Messwerte vorhanden sind, starten wir die Analysen mit der Betrachtung der Einzelabhängigkeiten aller Haupteffekte. Der Einfachheit halber bedienen wir uns dafür folgender Funktion.

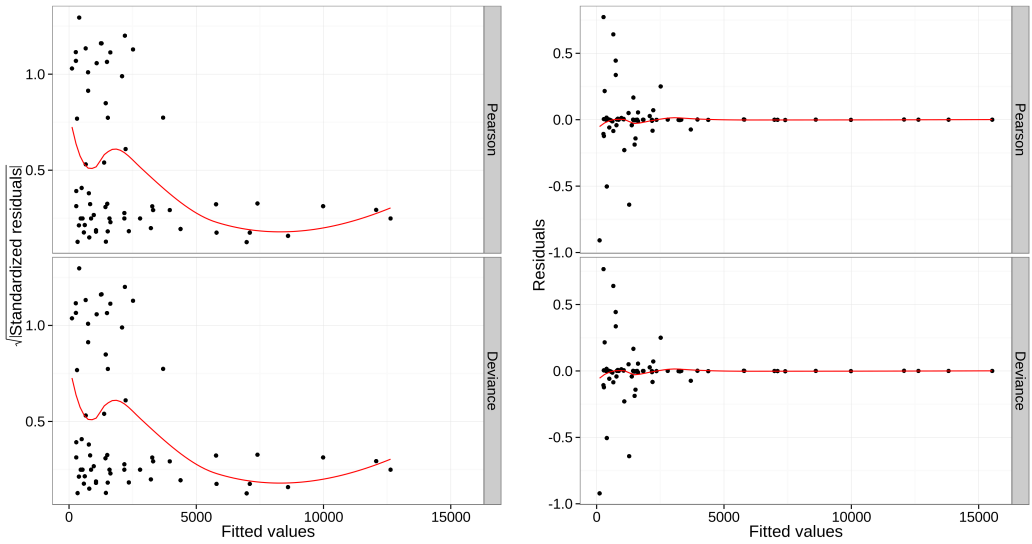

Abb. 5.15: Links: Der Scale-Location-Plot zum optimalen Poisson-Modell zeigt einen eindeutigen Trend und lässt somit an der Varianzannahme zweifeln.
Rechts: Die unskalierten Residuen sind zwar größtenteils nahe Null, jedoch erkennt man im Bereich der geringeren Sporenkonzentration eine beträchtliche Streuung.

```
select_main.eff<-function(y,family,test){
  list(anova(glm(y~typus,family=family),test=test),
       anova(glm(y~baujahr,family=family),test=test),
       anova(glm(y~baustoff,family=family),test=test),
       anova(glm(y~atemp+I(atemp^2),family=family),test=test),
       anova(glm(y~afeuchte+I(afeuchte^2),family=family),test=test),
       anova(glm(y~itemp+I(itemp^2),family=family),test=test),
       anova(glm(y~ifeuchte+I(ifeuchte^2),family=family),test=test),
       anova(glm(y~tauU,family=family),test=test),
       anova(glm(y~schimmel,family=family),test=test),
       anova(glm(y~klima,family=family),test=test),
}
```

Der Aufruf dieser Funktion mit $y = acfm\_dg\_i$, $family = quasipoisson()$ und $test =$ „F“ liefert die in folgender Tabelle dargestellten Ergebnisse. Die Bewertung mit + und - stellt den Einfluss des Haupteffekts auf die Sporenkonzentration dar. Hierbei ist anzumerken, dass der log-Link den Defaultwert dieser Familie darstellt und somit im Funktionsaufruf nicht explizit genannt werden muss.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp^2**	**Afeuchte**
	$+(1950-79), -(>1980)$	$+(Ziegel)$	$+$		$+$

Afeuchte^2	**Itemp**	**Itemp^2**	**Ifeuchte**	**Ifeuchte^2**	**TaupunktU**
$-$			$+$		$+(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Die Ergebnisse der Einzelanalysen decken sich größtenteils mit den Ergebnissen der explorativen Datenanalyse. Die signifikanten Zweifachinteraktionen werden durch ein analoges Schema ermittelt und es resultiert nach einigen weiteren Varianzanalysen das folgende Modell.

```
glm(formula = y ~ baujahr + baustoff + atemp + ifeuchte + schimmel +
    klima + baustoff:schimmel + baujahr:klima + baujahr:schimmel,
    family = quasipoisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-78.388  -17.443   -3.913   18.069   78.782

Coefficients: (3 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.245906   0.898493   9.177 9.33e-13
baujahr1950-1979               -1.598526   0.649434  -2.461 0.016946
baujahr>1980                   -1.735624   0.594361  -2.920 0.005033
baustoffziegel                 -0.539595   0.552934  -0.976 0.333323
atemp                           0.043569   0.012905   3.376 0.001342
ifeuchte                       -0.005104   0.010160  -0.502 0.617381
schimmelmittel                 -1.132084   0.825880  -1.371 0.175919
schimmelschwer                  2.161091   0.543555   3.976 0.000203
klimanein                      -1.626332   0.449822  -3.616 0.000643
baustoffziegel:schimmelmittel   1.795013   0.841156   2.134 0.037237
baustoffziegel:schimmelschwer         NA         NA      NA       NA
baujahr1950-1979:klimanein      3.383094   0.828887   4.081 0.000143
baujahr>1980:klimanein          2.193161   0.566948   3.868 0.000288
baujahr1950-1979:schimmelmittel 1.511227   0.767141   1.970 0.053797
baujahr>1980:schimmelmittel           NA         NA      NA       NA
baujahr1950-1979:schimmelschwer-1.437557   0.705128  -2.039 0.046210
baujahr>1980:schimmelschwer           NA         NA      NA       NA
---
```

```
(Dispersion parameter for quasipoisson family taken to be 1160.757)

    Null deviance: 212309  on 69  degrees of freedom
Residual deviance:  66725  on 56  degrees of freedom
```

Dieses Modell spiegelt bei Betrachtung der Deviance die Variabilität der Daten recht gut wieder (die Residual Deviance entspricht in *R* der unskalierten Deviance). Fasst man jedoch den Dispersionsparameter ins Auge, so erscheint das Modell doch noch verbesserungswürdig. Es ist anzunehmen, dass eine etwas spezifischere Varianzannahme die Daten besser beschreibt. Auch die zugehörigen Residuenplots (Abb. 5.16) zeigen eine große Streuung und steigenden Trend im Scatter-Location-Plot, was auf Überdispersion hinweist. Hier ist anzumerken, dass *R* keinen AIC ausgibt, da die Likelihood-Funktion bei *glm*() mit Quasi-Familien nicht explizit berechnet wird.

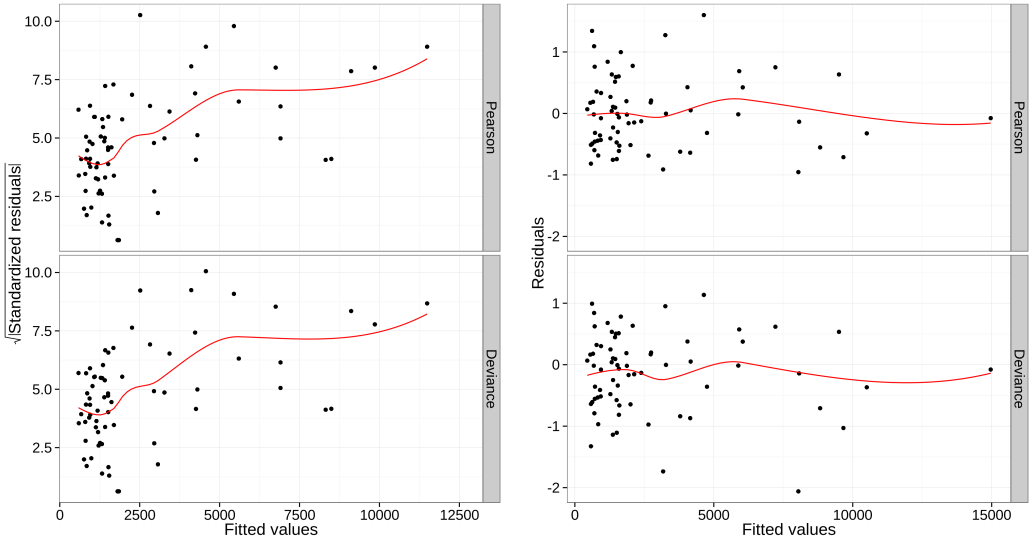

Abb. 5.16: Pearson- und Deviance-Residuen (standardisiert und unskaliert) des finalen Quasi-Poisson-Modells.

Um die vorliegenden Daten besser zu modellieren, wird nun die Möglichkeit einer Negativ-Binomial-Verteilung der Responses untersucht, also eine Datenvarianz von $V(\mu) = \mu + \mu^2/\alpha$. Hierzu wird die Funktion *glm.nb()* aus dem Package MASS verwendet (vgl. auch Venables und Ripley, 2002), welche iterativ die Funktion *glm*() mit fixiertem α (in *R theta* genannt) aufruft und dann wiederum α mittels Maximum-Likelihood und fixiertem $\boldsymbol{\beta}$ schätzt. Der Dispersionsparameter wird, wie beim Poisson-Modell, als 1 angenommen. Für Objekte der Klasse *negbin* (erzeugt von *glm.nb*()) sind alle

grundlegenden Funktionen, wie *anova*(), *summary*(), *update*() und *step*(), implementiert.
Um wieder vorab das Modell zu verkleinern, werden die Haupteffekte getrennt betrachtet.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp²**	**Afeuchte**
	$+(1950-79), -(>1980)$	$+(Ziegel)$	$+$		$+$

Afeuchte²	**Itemp**	**Itemp²**	**Ifeuchte**	**Ifeuchte²**	**TaupunktU**
$-$			$-$	$+$	$+(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Das endgültige Modell ergibt sich als das Folgende. Der Startwert für α (Argument *init.theta* bei *glm.nb*()) wird bei Funktionsaufruf automatisch berechnet, falls er nicht per Hand gesetzt ist, und entspricht dem Momentenschätzer bei Verwendung eines Poisson-Modells.

```
glm.nb(formula = y ~ typus + baujahr + atemp + schimmel + klima +
    atemp:schimmel + baujahr:klima + typus:baujahr, trace = 10,
    maxit = 200, init.theta = 2.483613003, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2443  -0.8981  -0.0664   0.4980   1.7888

                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 7.89783    0.41773  18.907  < 2e-16 ***
typusprod                  -0.26356    0.28443  -0.927 0.354113
baujahr1950-1979           -1.53478    0.52779  -2.908 0.003638 **
baujahr>1980               -1.76865    0.32495  -5.443 5.25e-08 ***
atemp                       0.01851    0.01754   1.055 0.291298
schimmelmittel             -1.85363    0.63920  -2.900 0.003732 **
schimmelschwer              0.31805    0.82167   0.387 0.698698
klimanein                  -0.43102    0.37209  -1.158 0.246709
atemp:schimmelmittel        0.09049    0.02844   3.182 0.001464 **
atemp:schimmelschwer        0.03416    0.03475   0.983 0.325669
baujahr1950-1979:klimanein  2.16801    0.59587   3.638 0.000274 ***
baujahr>1980:klimanein      0.84055    0.45869   1.832 0.066880 .
typusprod:baujahr1950-1979  0.78264    0.61963   1.263 0.206565
typusprod:baujahr>1980      1.05696    0.36469   2.898 0.003753 **
---
```

```
(Dispersion parameter for NegBin(2.4836) family taken to be 1)

    Null deviance: 207.047  on 69  degrees of freedom
Residual deviance:  74.607  on 56  degrees of freedom
AIC: 1199.2

              Theta:  2.484
          Std. Err.:  0.396

 2 x log-likelihood:  -1169.169

> X2(mod_negbin.3) #Pearson's X^2
[1] 57.66221
```

Der hohe Wert der Residual Deviance und die Diskrepanz zu Pearson's X^2-Statistik lassen am Negativ-Binomial-Ansatz zweifeln. Die Residuen (Abb. 5.17) verhalten sich schon wesentlich besser als bei den vohergehenden Modellen. Der Abwärtstrend bei sehr großen Sporenkonzentrationen könnte auch ausreißerbedingt sein.

Zur Absicherung wird noch ein Modell gerechnet, diesmal aber mit fixiertem α und zu schätzendem Dispersionsparameter ϕ. Das gelingt durch Aufruf der normalen $glm()$-Funktion mit $family = negative.binomial(theta = theta, link = \text{„}log\text{“})$ und Übergabe des zuvor berechneten Wertes für α als $theta$.

```
> alpha<-mod_negbin.3$theta
> glm(mod_negbin.3$call[2],family=negative.binomial(theta=alpha))
```

Dieser Aufruf führt zum selben Modell wie mit $glm.nb()$, die Standardfehler, Deviance (= 74.6) und AIC (= 1197.2) unterscheiden sich nur marginal. Der Dispersionsparameter resultiert als $\hat{\phi} = 1.03$ und verändert somit die ursprüngliche Varianzannahme nicht maßgeblich. Obwohl das AIC des überbestimmten Poisson-Modells kleiner ausfällt als das des Negativ-Binomial-Modells, bestätigt auch der folgende Likelihood-Quotiententest die vorherrschende Überdispersion. Dieser ist im Package *pscl*[5] implementiert und vergleicht das übergebene Negativ-Binomial-Modell mit dem durch dieselbe Formel gefitteten Poisson-Modell.

```
> odTest(mod_negbin.3)
Likelihood ratio test of H0: Poisson, as restricted NB model
```

[5] Jackman S., u.a.: *pscl - Political Science Computational Laboratory, Stanford University* - Version 1.04.4

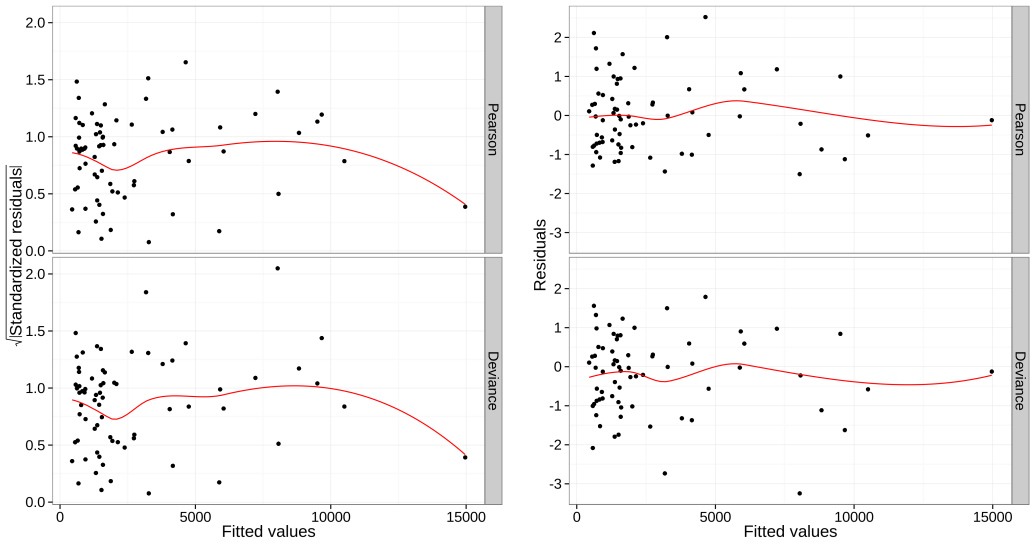

Abb. 5.17: Pearson- und Deviance-Residuen des finalen Negativ-Binomial-Modells.

```
Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  68896.2469 p-value = < 2.2e-16
```

Nun wird die Möglichkeit einer quadratischen Datenvarianz $V(\mu) = \phi\mu^2$ untersucht. Hierzu kann man entweder wieder einen Quasi-Likelihood-Ansatz mittels *family = quasi(link = „log“, variance = „mu^2“)* wählen oder der Einfachheit halber ein Gamma-Modell mit *family = Gamma(link = „log“)* verwenden. Dies hat den Vorteil, dass für nicht-Quasi-Modelle die Funktion *step()* anwendbar ist und somit die Vorabauswahl der Parameter erleichtert wird. Gestartet wird wieder mit der Auswahl der Haupteffekte, deren Einfluss in folgender Tabelle dargestellt wird.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp²**	**Afeuchte**
	$+(1950-79), -(>1980)$	$+(Ziegel)$	$+$		$+$

Afeuchte²	**Itemp**	**Itemp²**	**Ifeuchte**	**Ifeuchte²**	**TaupunktU**
$-$			$-$	$+$	$+(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Zu den Haupteffekten zeigen sich noch einige Interaktionen als signifikant und führen schlussendlich zu folgendem Modell.

```
glm(formula = y ~ typus + baujahr + atemp + schimmel + klima +
```

```
    atemp:schimmel + baujahr:klima + typus:baujahr,
    family = Gamma(link ="log"))
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.89756    0.42395  18.628  < 2e-16 ***
typusprod                  -0.26370    0.28870  -0.913 0.364951
baujahr1950-1979           -1.53475    0.53540  -2.867 0.005837 **
baujahr>1980               -1.76869    0.32976  -5.364 1.61e-06 ***
atemp                       0.01852    0.01779   1.041 0.302383
schimmelmittel             -1.85386    0.64862  -2.858 0.005973 **
schimmelschwer              0.31796    0.83415   0.381 0.704512
klimanein                  -0.43072    0.37766  -1.140 0.258939
atemp:schimmelmittel        0.09049    0.02886   3.136 0.002731 **
atemp:schimmelschwer        0.03415    0.03528   0.968 0.337221
baujahr1950-1979:klimanein  2.16771    0.60458   3.586 0.000706 ***
baujahr>1980:klimanein      0.84017    0.46552   1.805 0.076488 .
typusprod:baujahr1950-1979  0.78266    0.62860   1.245 0.218288
typusprod:baujahr>1980      1.05710    0.37006   2.857 0.005999 **
---
(Dispersion parameter for Gamma family taken to be 0.4153174)

    Null deviance: 83.502  on 69  degrees of freedom
Residual deviance: 30.113  on 56  degrees of freedom

> X2(mod_gamma.2) #Pearson's X^2
[1] 23.25778
```

Hier fällt auf, dass einerseits die skalierte Deviance mit einem Wert von 72.5 im Vergleich zu den Freiheitsgraden noch immer relativ groß ist und auch die Pearson-Statistik verhältnismäßig kleiner ausfällt als die unskalierte Deviance. Die Residuen (Abb. 5.18) verhalten sich ähnlich dem negativ-binomialen Modell, wobei sie im unskalierten Fall etwas weniger streuen.

Um die sequentielle Suche nach der optimalen Varianzfamilie zu erleichtern, wird nun ein extended Quasi-Likelihood-Ansatz angewandt. Das passende Paket dazu wird in der Arbeit von Thaler (2009) vorgestellt und nennt sich *„EQL“*. An die Funktion $eql()$ wird ein Modell als Objekt der Klasse $formula$, der zu durchsuchende Parameterraum in Form einer Liste und eine Varianzfamilie übergeben. Es können auch noch weitere Parameter gewählt werden, wie zum Beispiel die Schätzmethode von ϕ, im Fall eines eindimensionalen zu schätzenden θ die Möglichkeit einer Interpolation zwischen den berechneten EQL-Werten und allgemeine Parameter der $glm()$-Routine. Die Varianzfamilie kann entweder als $powerVarianceFamily()$, entsprechend $V(\mu|\theta) = \mu^\theta$, oder als $extBinomialVarianceFamily$ mit $V(\mu|k,l) = \mu^k(1-\mu)^l$ angenommen werden. Wird eine allgemeinere Familie benötigt, so kann

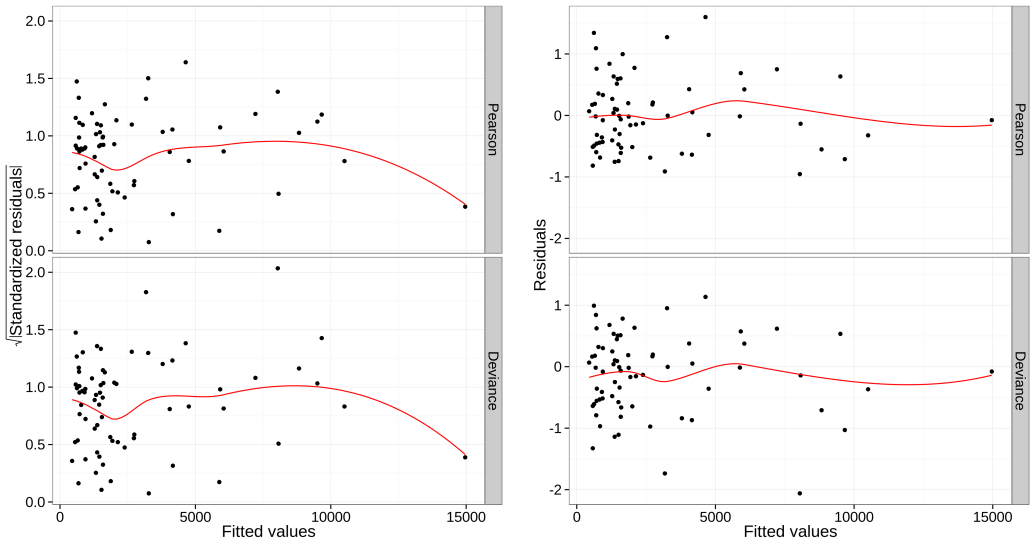

Abb. 5.18: Pearson- und Deviance-Residuen des finalen Gamma-Modells.

auch eine beliebige parametrisierte Varianzfunktion mittels $varianceFamily()$ verwendet werden. Die Plotfunktion der Klasse *eql* stellt bei eindimensionalem θ die jeweilig berechneten Werte in einem Profile-Plot dar oder bei einem mehrdimensionalen Parametervektor $\boldsymbol{\theta}$ in Form eines Kontur-Plots.

In diesem Fall wird eine Power-Varianzannahme getroffen und das finale Modell mit quadratischer Varianzfunktion (im nachfolgendem Codefragment als „min“ bezeichnet) mit einem allgemeineren Modell (mit „max“ bezeichnet) verglichen. Als Parameterraum für θ wird eine Sequenz von 50 Werten zwischen 1 und 3 angenommen. Die Funktion $eql()$ liefert die folgenden optimalen Schätzer für Annahme einer vorliegenden Power-Varianzfamilie.

```
> ps<-list(seq(1,3,length=50))

# Minimales Modell: p=14
> eql_pv_min<-eql(min,param.space=ps,family=powerVarianceFamily())
> eql_pv_min
EQL-Max: -582.1751 at: theta = 2.020
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -597.6  -591.0  -586.2  -587.4  -583.2  -582.2


# Allgemeineres Modell: p=25
> eql_pv_min<-eql(max,param.space=ps,family=powerVarianceFamily())
> eql_pv_max
```

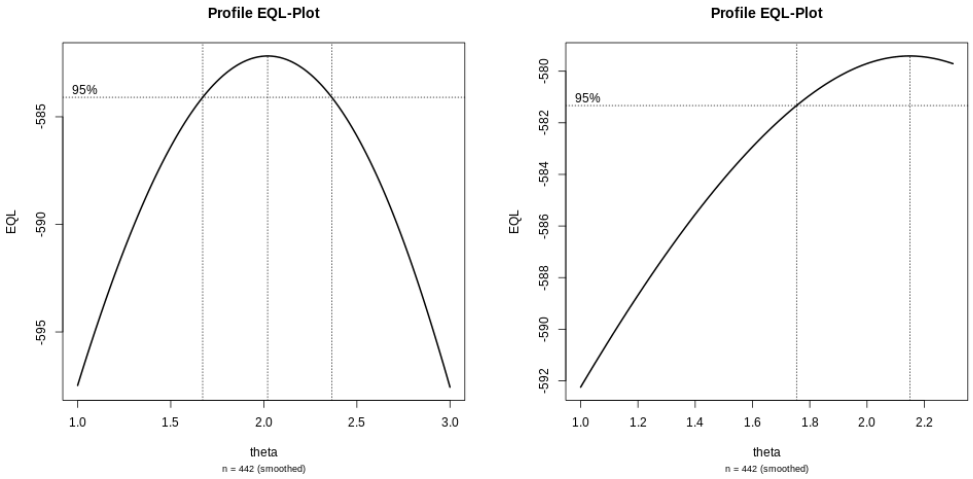

Abb. 5.19: Profilplot der EQL-Funktion in Abhängigkeit von θ inkl. 95%-Konfidenzintervall um den Maximalwert.
Links: minimales Modell. Rechts: allgemeines Modell

```
EQL-Max: -579.4109 at: theta = 2.143
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -613.3  -600.6  -596.4  -593.3  -583.3  -579.4
```

Beide EQL-Modelle weisen auf ein optimales Modell mit quadratischer Varianzfunktion hin, wobei zu erwähnen ist, dass die Konvergenz des allgemeinen Modells mit 25 Parametern nur bis zu einem θ von 2.3 gewährleistet ist. Die Profilplots (Abb. 5.19) bestätigen die Wahl des quadratischen Varianzansatzes, sie zeigen für allgemeines wie auch minimales Modell den Wert 2 im 95%-Konfidenzintervall. Die Verkleinerung des Datensatzes auf ACFM DG-Innenmesswerte < 10.000 bzw. < 7000 KBE/m^3 führt zu keiner maßgeblichen Verbesserung der Modellgüte und ist auch praktisch nicht gerechtfertigt. Aufgrund der EQL-AICs der beiden minimalen Modelle könnte man sich für jenes mit $\theta = 2.02$ entscheiden und somit einen Tweedie-Ansatz[6] wählen, der außerhalb der Exponentialfamilie liegt und deshalb auch in dieser Arbeit nicht näher behandelt wird.

	$\theta = 2.02$	$\theta = 2$
AIC_{EQL}	42.93989	53.21589

[6]Die Familie der Tweedie-Verteilungen gehört zur Klasse der *exponentiellen Dispersionsmodelle*, welche von JØRGENSEN (1987) vorgestellt wurden. Sie beinhaltet Verteilungen mit Power-Varianzfunktionen der Form $V(\mu) = \mu^k$ mit $k \in \mathbb{R}\backslash(0,1)$.

Betrachtet man zusätzlich noch die Standardfehler der bisher betrachteten Modelle (inkl. Tweedie-Modell), so erkennt man, dass kaum ein Unterschied besteht und somit auch die Wahl eines Modells mit quadratischer Varianzfunktion gerechtfertigt werden kann. Genauso gut könnte man sich auch auf das Modell mit negativ-binomialer Varianzannahme einigen, da dieselben Parameter beschrieben werden und auch die Modellgüte ähnlich passend ist. Die 3. Spalte zeigt die Standardfehler des quadratischen Modells unter Verwendung der Quasi-Familie bei $glm()$-Aufruf und soll nur nochmals veranschaulichen, dass kein Unterschied in den Ergebnissen besteht.

	Tweedie	Gamma	Quasi^2	NegBin
(Intercept)	0.42443	0.42395	0.42395	0.41773
typusprod	0.29022	0.28870	0.28870	0.28443
baujahr1950-1979	0.53286	0.53540	0.53540	0.52779
baujahr>1980	0.33013	0.32976	0.32976	0.32495
atemp	0.01777	0.01779	0.01779	0.01754
schimmelmittel	0.64865	0.64862	0.64862	0.63920
schimmelschwer	0.83979	0.83415	0.83415	0.82167
klimanein	0.37931	0.37766	0.37766	0.37209
atemp:schimmelmittel	0.02892	0.02886	0.02886	0.02844
atemp:schimmelschwer	0.03554	0.03528	0.03528	0.03475
baujahr1950-1979:klimanein	0.60398	0.60458	0.60458	0.59587
baujahr>1980:klimanein	0.46639	0.46552	0.46552	0.45869
typusprod:baujahr1950-1979	0.62583	0.62860	0.62860	0.61963
typusprod:baujahr>1980	0.37022	0.37006	0.37006	0.36469

Somit resultiert folgendes Modell für die erwartete Anzahl KBE/m^3 beim ACFM Kaskaden-Impaktor mit DG18 als Nährboden:

$$\begin{aligned}\hat{\mathbb{E}}(y) = \exp\{&7.9 - 0.3 \cdot \mathbb{I}(Typus = prod) - 1.5 \cdot \mathbb{I}(Baujahr = 1950 - 1979) \\ &- 1.8 \cdot \mathbb{I}(Baujahr = > 1980) + 0.02 \cdot aTemp - 1.9 \cdot \mathbb{I}(Schimmel = mittel) \\ &+ 0.3 \cdot \mathbb{I}(Schimmel = schwer) - 0.4 \cdot \mathbb{I}(Klima = nein) \\ &+ 0.09 \cdot aTemp \cdot \mathbb{I}(Schimmel = mittel) \\ &+ 0.03 \cdot aTemp \cdot \mathbb{I}(Schimmel = schwer) \\ &+ 2.2 \cdot \mathbb{I}(Baujahr = 1950 - 1979) \cdot \mathbb{I}(Klima = nein) \\ &+ 0.8 \cdot \mathbb{I}(Baujahr = > 1980) \cdot \mathbb{I}(Klima = nein) \\ &+ 0.8 \cdot \mathbb{I}(Typus = prod) \cdot \mathbb{I}(Baujahr = 1950 - 1979) \\ &+ 1.1 \cdot \mathbb{I}(Typus = prod) \cdot \mathbb{I}(Baujahr =< 1980)\}.\end{aligned}$$

Die Deutung der Modelle erfolgt am Einfachsten durch die Betrachtung der Plots in Abb. 5.20, die die Auswirkungen aller möglichen Indikatorkom-

binationen in Abhängigkeit von der Außentemperatur auf die Sporenkonzentration zeigen.

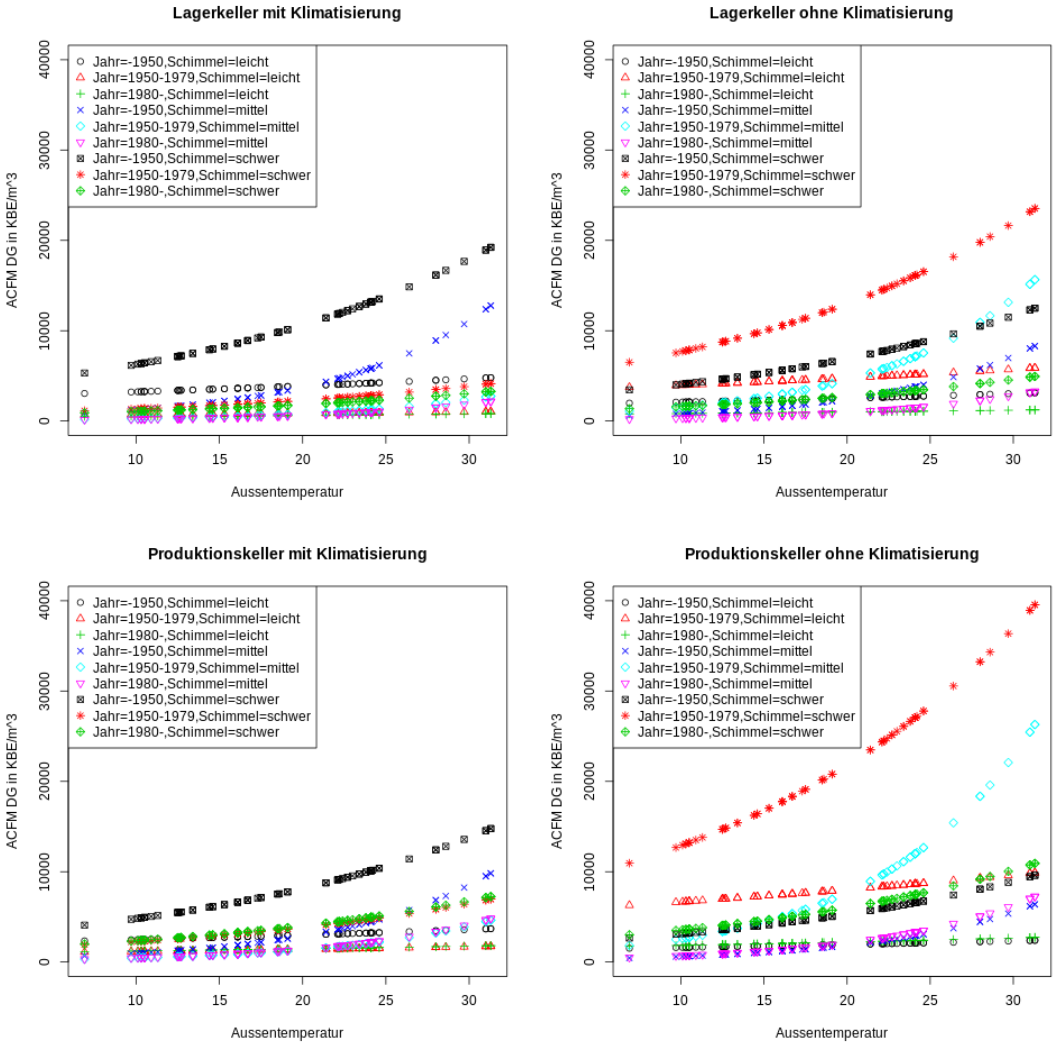

Abb. 5.20: Erwartete Anzahl KBE/m^3 bei ACFM mit DG18 als Nährboden. Grundsätzlich zeigen Keller ohne Klimatisierung, die zwischen 1950 und 1979 gebaut wurden und schweren Schimmelbefall aufweisen, mit steigender Außentemperatur die größte Sporenkonzentration in der Innenluft. Bei Kellern mit Klimatisierung zeigen vorallem Keller mit Baujahr vor 1950 die stärkste Keimbelastung.

5.1.3 Regressionsmodelle für MAS100 mit DG18

Nun werden die Innenmessungen der Sporenkonzentrationen mit dem MAS Air Sampler mit DG18 als Nährboden näher untersucht. Es liegen dieselben erklärenden Daten zugrunde wie für die Untersuchungen im vorigen Abschnitt mit ACFM als Messtool.

Aufgrund der empirischen Voruntersuchungen und um die Parameterauswahl zu erleichtern wird mit einem quadratischen Varianz-Modell gestartet. Die folgenden Haupteffekte zeigen sich als signifikant.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp^2**	**Afeuchte**
	$-(1950-79), -(>1980)$	$+(Ziegel)$	$+$		

Afeuchte^2	**Itemp**	**Itemp^2**	**Ifeuchte**	**Ifeuchte^2**	**TaupunktU**
	$+$		$-$	$+$	$-(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Das unter der quadratischen Varianzannahme optimale Modell ergibt sich als:

```
glm(formula = y ~ typus + baujahr + baustoff + atemp + ifeuchte +
    schimmel + klima + atemp:schimmel + baustoff:ifeuchte
    + typus:ifeuchte, family = Gamma(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75363  -0.54493  -0.06412   0.40008   1.44518

                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.56295    1.19605   2.143 0.036485 *
typusprod               4.37809    1.26020   3.474 0.000996 ***
baujahr1950-1979       -0.72030    0.33253  -2.166 0.034576 *
baujahr1980-           -1.20054    0.36978  -3.247 0.001975 **
baustoffziegel          3.61244    1.70724   2.116 0.038808 *
atemp                   0.03781    0.01905   1.985 0.052096 .
ifeuchte                0.06779    0.01814   3.736 0.000440 ***
schimmelmittel         -2.63432    0.64362  -4.093 0.000138 ***
schimmelschwer         -0.32967    0.80910  -0.407 0.685227
klimanein               0.65269    0.21953   2.973 0.004340 **
atemp:schimmelmittel    0.09977    0.03003   3.322 0.001579 **
atemp:schimmelschwer    0.02259    0.03822   0.591 0.556903
```

```
baustoffziegel:ifeuchte -0.05706    0.02629  -2.170 0.034259 *
typusprod:ifeuchte      -0.05842    0.01863  -3.136 0.002731 **
---
(Dispersion parameter for Gamma family taken to be 0.4496163)

    Null deviance: 75.551  on 69  degrees of freedom
Residual deviance: 29.878  on 56  degrees of freedom
```

Die Güte spiegelt sich in einer skalierten Deviance von 66.45 und einem X^2 von 25.18 als recht gut wider und auch die Residuenplots in Abb. 5.21 bekräftigen diese Aussage.

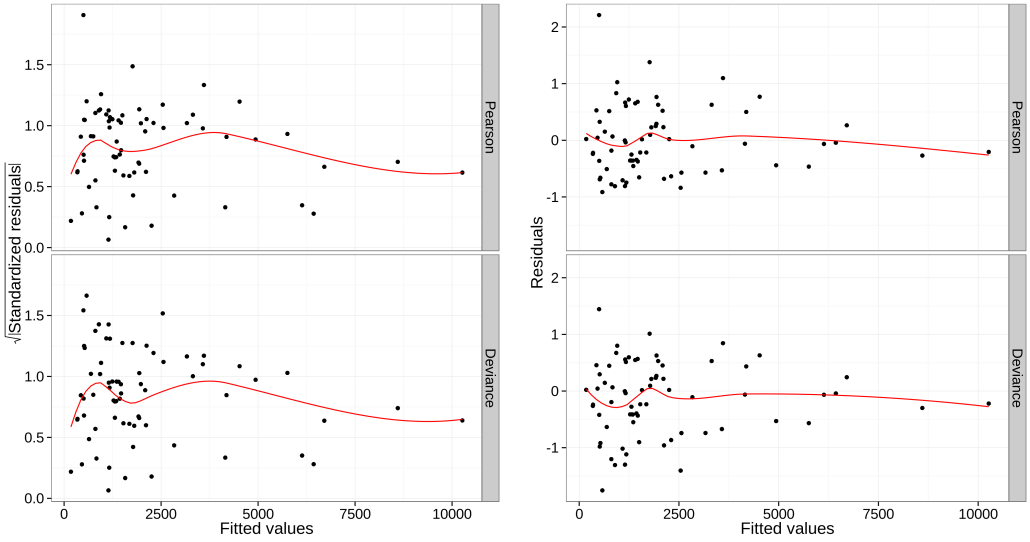

Abb. 5.21: Pearson- und Deviance-Residuen des finalen Gamma-Modells.

Das oben genannte Modell wird nun auch unter dem Aspekt des EQL-Ansatzes betrachtet und führt unter der Annahme einer vorliegenden Varianzfunktion aus der Power-Varianzfamilie zu einem optimalen θ von 1.606.

```
EQL-Max: -564.7702 at: theta = 1.606
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -589.1  -577.8  -570.4  -572.5  -566.2  -564.8
```

Auch in diesem Fall befindet sich $\theta = 2$ noch innerhalb des Konfidenzintervalls, was das zuvor berechnete Modell als passend bestätigt. Im Vergleich zu den Ergebnissen der Untersuchungen zum ACFM mit DG18 benötigt man zur Schätzung der erwarteten Sporenkonzentration mit MAS100 zusätzlich

noch Informationen über die Luftfeuchtigkeit im Kellerinneren und den Baustoff des Weinkellers. Genaueren Aufschluss über die Zusammenhänge der einzelnen Parameter mit der Sporenkonzentration geben Abb. 5.22 und Abb. 5.23. Generell höher keimbelastet zeigen sich auch hier Produktionskeller mit Baujahr vor 1950.

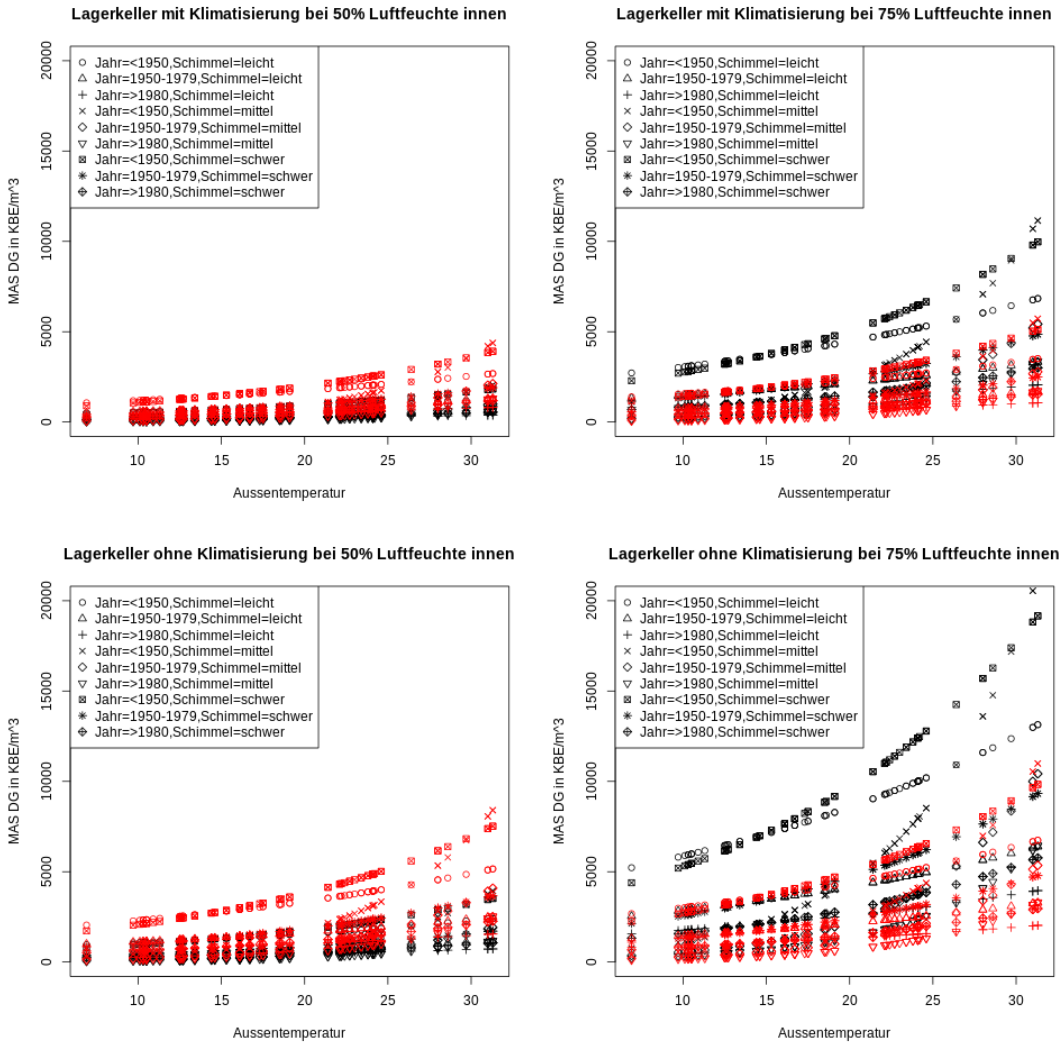

Abb. 5.22: Erwartete Anzahl KBE/m^3 bei MAS100 mit DG18 als Nährboden. Rot dargestellt sind Ziegel- und schwarz Betonkeller.
Unabhängig vom Baustoff steigt die Sporenkonzentration proportional zur Außentemperatur. Zusätzlich zeigt sich, dass mit hoher Luftfeuchte die Keimbelastung in Betonkellern im Vergleich zu Weinkellern mit Ziegelbauweise stärker zunimmt.

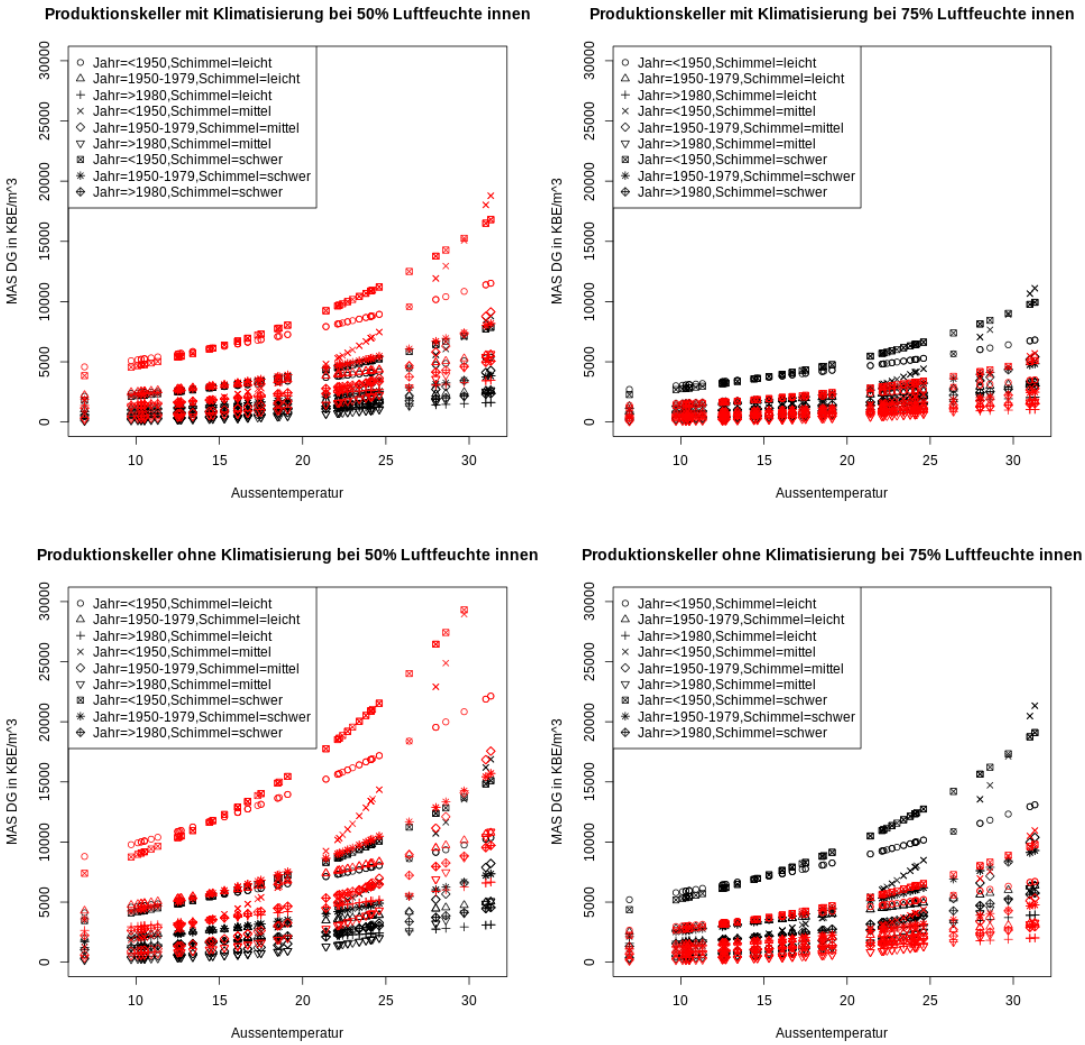

Abb. 5.23: Erwartete Anzahl KBE/m^3 bei MAS100 mit DG18 als Nährboden. Rot dargestellt sind Ziegel- und schwarz Betonkeller.
Wie bei Lagerkellern steigt auch bei Produktionskellern die Sporenkonzentration proportional zur Außentemperatur. Hohe Luftfeuchte begünstigt die Keimbelastung in Betonkellern, jedoch werden die höchsten Konzentrationen bei moderater Luftfeuchte von 50% in Ziegelkellern gefunden.

5.1.4 Regressionsmodelle für ACFM mit MEA

Bei der Untersuchung der KBE/m^3 von mesophilen Schimmelpilzarten, gemessen mittels ACFM und MEA als Nährboden, zeigen sich bei Einzelanalysen mittels quadratischen und negativ-binomialen Ansätzen die folgenden

Haupteffekte gleichermaßen als signifikant.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp²**	**Afeuchte**
	$-(1950-79), -(>1980)$	$+(Ziegel)$	$+$		$+$

Afeuchte²	**Itemp**	**Itemp²**	**Ifeuchte**	**Ifeuchte²**	**TaupunktU**
$-$			$-$	$+$	$+(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Nach zusätzlicher Analyse der einzelnen möglichen Interaktionen und Verkleinerung des Modells mittels *step*() resultiert das unter der Negativ-Binomial-Annahme optimale Modell als:

```
glm.nb(formula = y ~ typus + baujahr + baustoff + atemp + afeuchte +
    schimmel + atemp:schimmel + baustoff:afeuchte + typus:baujahr,
    maxit = 200, init.theta = 1.877248145, link = log)

                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 6.508451   0.719999   9.040  < 2e-16 ***
typusprod                  -0.291622   0.310770  -0.938 0.348046
baujahr1950-1979           -0.086167   0.378564  -0.228 0.819945
baujahr1980-               -1.661226   0.467129  -3.556 0.000376 ***
baustoffziegel              1.988598   0.936967   2.122 0.033806 *
atemp                       0.001197   0.019905   0.060 0.952064
afeuchte                    0.033615   0.009012   3.730 0.000192 ***
schimmelmittel             -2.645928   0.676109  -3.913 9.10e-05 ***
schimmelschwer             -0.428866   0.866037  -0.495 0.620455
atemp:schimmelmittel        0.127571   0.032431   3.934 8.37e-05 ***
atemp:schimmelschwer        0.051386   0.040760   1.261 0.207411
baustoffziegel:afeuchte    -0.040716   0.016763  -2.429 0.015145 *
typusprod:baujahr1950-1979 -0.224054   0.637910  -0.351 0.725415
typusprod:baujahr1980-      0.895434   0.402256   2.226 0.026012 *
---
(Dispersion parameter for NegBin(1.8772) family taken to be 1)

    Null deviance: 186.550  on 69  degrees of freedom
Residual deviance:  76.989  on 56  degrees of freedom
AIC: 1198.5
```

Die Residuen des Negativ-Binomial- und quadratischen Modells unterscheiden sich nur minimal (siehe Abb. 5.24), lediglich im Verhältnis skalierte Deviance zu Freiheitsgraden schneidet das Negativ-Binomial-Modell besser ab

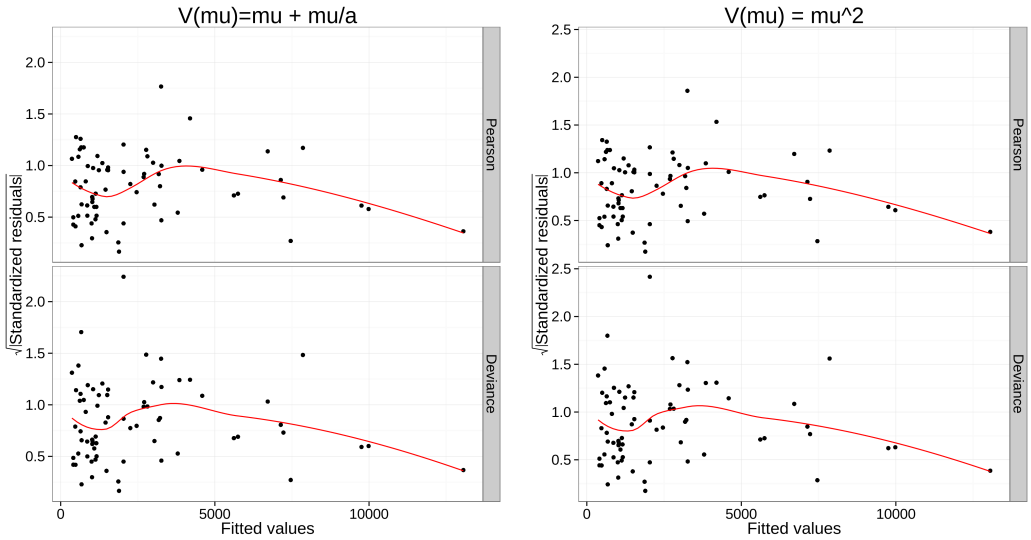

Abb. 5.24: Residuenplots des Negativ-Binomial- und quadratischen Modells

(das Modell mit quadratischer Varianzannahme schlägt sich mit einer skalierten Deviance von 97.38 bei 56 Freiheitsgraden zu Buche).

Um mehr Aufschluss über die Power-Varianzfamilie zu bekommen, die den Daten zugrunde liegt, wird wieder der EQL-Ansatz herangezogen und der Funktion *eql*() die zuvor optimierte Modellformel übergeben.

```
EQL-Max: -575.6068 at: theta = 1.619
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -618.3  -599.1  -585.4  -589.4  -578.1  -575.6
```

Bei mesophilen Schimmelpilzarten scheint die Verteilung der Sporenkonzentration nicht Teil der Exponentialfamilie zu sein, da, wie in Abb. 5.25 ersichtlich, weder $\theta = 1$ noch $\theta = 2$ im Konfidenzintervall enthalten sind. Es resultiert also, wie schon im Beispiel zuvor, ein Vertreter der Tweedie-Varianzfamilie als Optimum unter dem EQL-Ansatz. Für $\theta \in (1, 2)$ gehört die korrespondierende Varianzfunktion zu einer zusammengesetzten Poisson-Verteilung (vgl. für mehr Details JØRGENSEN, 1992). Um ein Tweedie-Modell zu fitten, benötigt man die R-Packages *tweedie* und *statmod*. Der Funktion *glm*() wird dann als Family ein Objekt der Klasse *tweedie* mit den Parametern *var.power* und *link.power* übergeben, welche den Exponenten der Power-Varianzfamilie und die zu verwendende Linkfunktion enthalten (*link.power* = 0 entspricht hierbei dem log-Link).

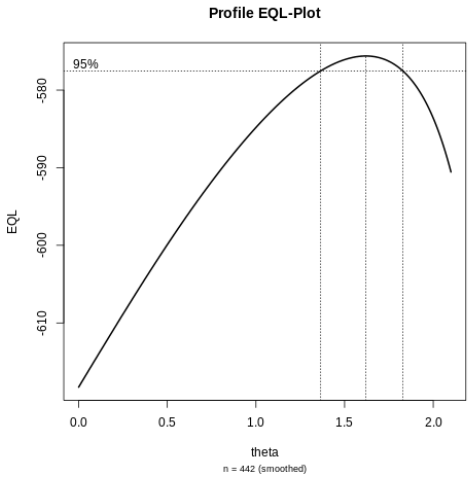

Abb. 5.25: Profilplot der EQL-Funktion in Abhängigkeit von θ

```
> ## zusammengesetzte Poisson-Verteilung
> family<-tweedie(var.power=1.6,link.power=0)
> summary(mod_tweedie.5)
Call:
glm(formula = y ~ baujahr + atemp + afeuchte + schimmel + klima +
    afeuchte:schimmel + atemp:schimmel,
    family = tweedie(var.power = 1.6, link.power = 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.9286  -2.1122  -0.4558   1.2102   7.0687

                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.217429   0.587261  10.587 3.59e-15 ***
baujahr1950-1979        -0.157706   0.269969  -0.584 0.561374
baujahr1980-            -1.081156   0.265935  -4.065 0.000146 ***
atemp                    0.005157   0.019409   0.266 0.791414
afeuchte                 0.031097   0.008086   3.846 0.000301 ***
schimmelmittel          -0.743069   1.403231  -0.530 0.598451
schimmelschwer           5.489375   2.005840   2.737 0.008224 **
klimanein                0.453859   0.215607   2.105 0.039631 *
afeuchte:schimmelmittel -0.041046   0.022943  -1.789 0.078831 .
afeuchte:schimmelschwer -0.107535   0.032813  -3.277 0.001773 **
atemp:schimmelmittel     0.131759   0.028615   4.605 2.30e-05 ***
atemp:schimmelschwer     0.003542   0.037713   0.094 0.925502
---
```

```
(Dispersion parameter for Tweedie family taken to be 9.459495)

    Null deviance: 1927.69  on 69  degrees of freedom
Residual deviance:  619.76  on 58  degrees of freedom
```

Auffallend ist, dass sich bei diesem Ansatz die signifikanten Parameter geändert haben und nun weniger erklärende Variablen für eine gute Modellanpassung von Nöten sind. Die skalierte Deviance des Tweedie-Modells beträgt 65.52 bei 58 Freiheitsgraden, wobei die Streuung der Residuen stärker ausgeprägt ist als beim Negativ-Binomial-Modell (Abb. 5.26). Ein AIC kann mittels der Funktion $AICtweedie()$ berechnet werden und lässt sich mit dem des Negativ-Binomial-Modells vergleichen. Die Größenordnung der Standardfehler ist auch unauffällig, somit wird dieses Modell dem negativ-binomialen vorgezogen.

	$NegBin$	$Tweedie$
AIC	1198.5	1186.8

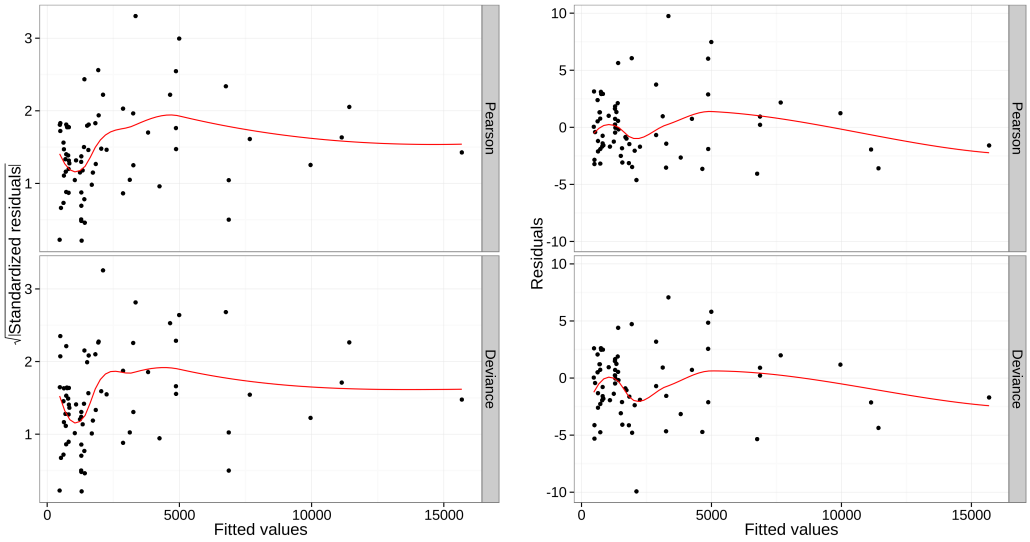

Abb. 5.26: Residuenplots des Tweedie-Modells mit $\theta = 1.6$

Abbildung 5.27 stellt die gefundenen Abhängigkeiten übersichtlich dar. Grundsätzlich zeigen Keller ohne Klimatisierung eine höhere Sporenkonzentration als klimatisierte Weinkeller und steigende Außenluftfeuchtigkeit scheint einen senkenden Einfluss auf die Schimmelbelastung zu haben. Die Außentemperatur hat auch einen wesentlichen Einfluss.

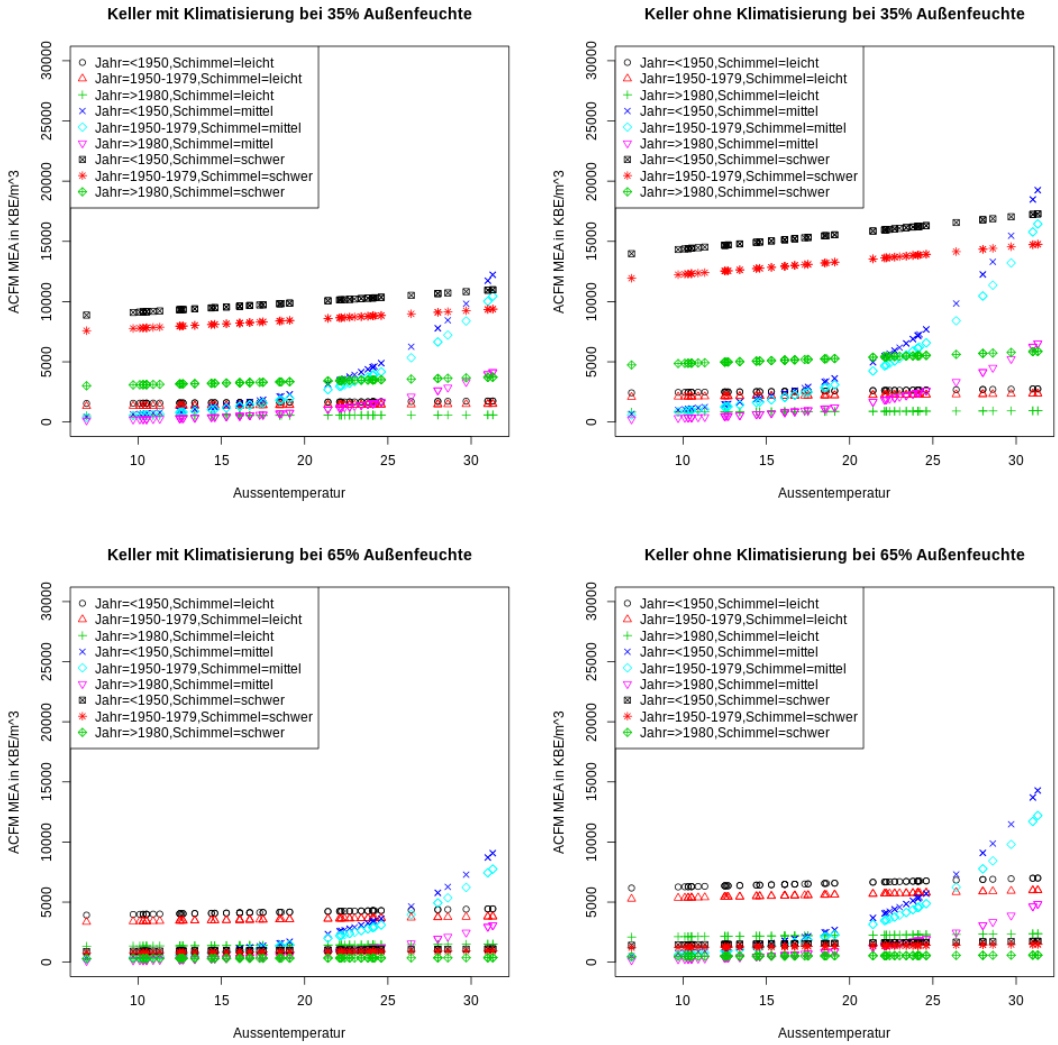

Abb. 5.27: Erwartete Anzahl KBE/m^3 bei ACFM mit MEA als Nährboden aufgeteilt nach Klimatisierung und Außenluftfeuchtigkeit.
Vorallem Keller mit starkem Schimmelbefall, die vor 1980 erbaut wurden, weisen bei geringer Außenfeuchte die höchsten Sporenkonzentrationen auf. Auffällig ist, dass bei Außentemperaturen um die 30°C die Keimbelastung in Kellern mit nur mittlerem sichtbaren Schimmelbefall markant ansteigt.

5.1.5 Regressionsmodelle für MAS100 mit MEA

Der MAS100 Air Sampler wurde auch mit einem MEA Nährboden bestückt und damit die Sporenkonzentration in der Weinkellerluft bestimmt. Als Einstieg werden wieder die Einflüsse der einzelnen Parameter auf die Anzahl

koloniebildender Einheiten betrachtet, wobei der quadratische sowie negativbinomiale Ansatz dieselben Abhängigkeiten zeigen.

Typus	**Baujahr**	**Baustoff**	**Atemp**	**Atemp²**	**Afeuchte**
	$-(1950-79), -(>1980)$	$+(Ziegel)$	$+$		

Afeuchte²	**Itemp**	**Itemp²**	**Ifeuchte**	**Ifeuchte²**	**TaupunktU**
	$+$		$-$	$+$	$-(moegl), -(nein)$

Schimmel	**Klima**
$+(mittel), +(schwer)$	$+(nein)$

Nach Hinzunahme der einzeln signifikanten Interaktionen und nochmaliger Variablenselektion mittels *anova*() ergibt sich das folgende Modell als optimal unter Annahme einer quadratischen Varianzfunktion.

```
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.9326368  1.8861162   5.266 2.13e-06 ***
typusprod                 -0.1773415  0.2479315  -0.715 0.477303
baujahr1950-1979           0.1615831  0.2949384   0.548 0.585895
baujahr1980-              -1.0120899  0.2846894  -3.555 0.000760 ***
atemp                      0.0725967  0.0125461   5.786 3.06e-07 ***
afeuchte                  -0.1021304  0.0436744  -2.338 0.022833 *
ifeuchte                  -0.0868600  0.0276930  -3.137 0.002685 **
schimmelmittel             0.2492200  0.2689474   0.927 0.357948
schimmelschwer             1.3333609  0.2929710   4.551 2.78e-05 ***
afeuchte:ifeuchte          0.0019540  0.0006444   3.032 0.003628 **
typusprod:baujahr1950-1979 -0.0411329  0.4813624  -0.085 0.932197
typusprod:baujahr1980-      1.1686370  0.3231312   3.617 0.000627 ***
---
(Dispersion parameter for Gamma family taken to be 0.3455652)

    Null deviance: 77.817  on 69  degrees of freedom
Residual deviance: 23.990  on 58  degrees of freedom

> X2(mod_gamma.2)
[1] 20.04278
```

Im Gegensatz zum ACFM-Messgerät scheint beim MAS100 auch die Luftfeuchtigkeit in den Weinkellern einen signifikanten Einfluss auf die Sporenkonzentration zu haben. Die Modellgüte ist mit einer skalierten Deviance von 69.42 bei 58 Freiheitsgraden und Pearson's X^2 von 20.04 nicht perfekt, aber in Ordnung. In Abbildung 5.28 sind die zugehörigen standardisierten Residuen

dargestellt, wobei die Deviance-Residuen im Bereich sehr hoher Konzentrationen etwas auszureißen scheinen. Um diese Varianzannahme und damit das Modell zu bestätigen, wird auch noch die EQL-Funktion betrachtet.

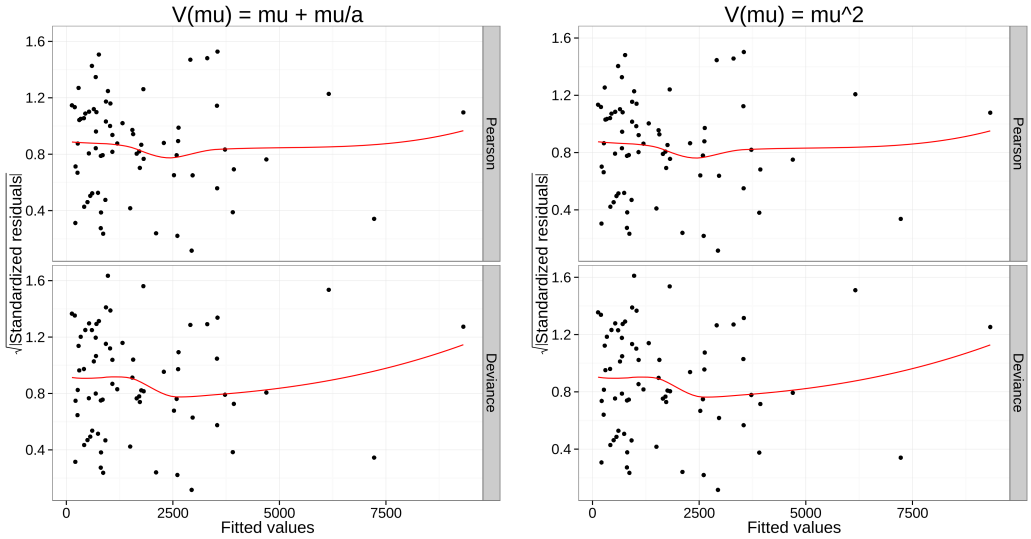

Abb. 5.28: Residuenplots des negativ-binomialen und quadratischen Modells
Die Residuen verhalten sich, wie erwartet bei der gleichen Auswahl an Prädiktoren, sehr ähnlich.

```
EQL-Max: -541.9777 at: theta = 1.907
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -588.7  -563.6  -552.1  -556.0  -544.6  -542.0
```

Das Maximum der EQL-Funktion und das zugehörige 95%-Konfidenzintervall (der Wert 2 liegt innerhalb) bestätigen die quadratische Varianzannahme. Den genauen Zusammenhang bzw. Einfluss der einzelnen Prädiktoren auf die Keimbelastung sieht man am besten in den Abbildungen 5.29 und 5.30. Generell zeigen auch mit dieser Messkonstellation Keller mit starkem sichtbaren Schimmelbefall, die vor 1980 gebaut wurden, die höchste Sporenkonzentration in der Weinkellerluft. Vorallem die Kombination hohe Außenfeuchte mit hoher Innenluftfeuchtigkeit zeigt bei steigender Außentemperatur gravierende Auswirkungen auf die Konzentration der Schimmelpilze.

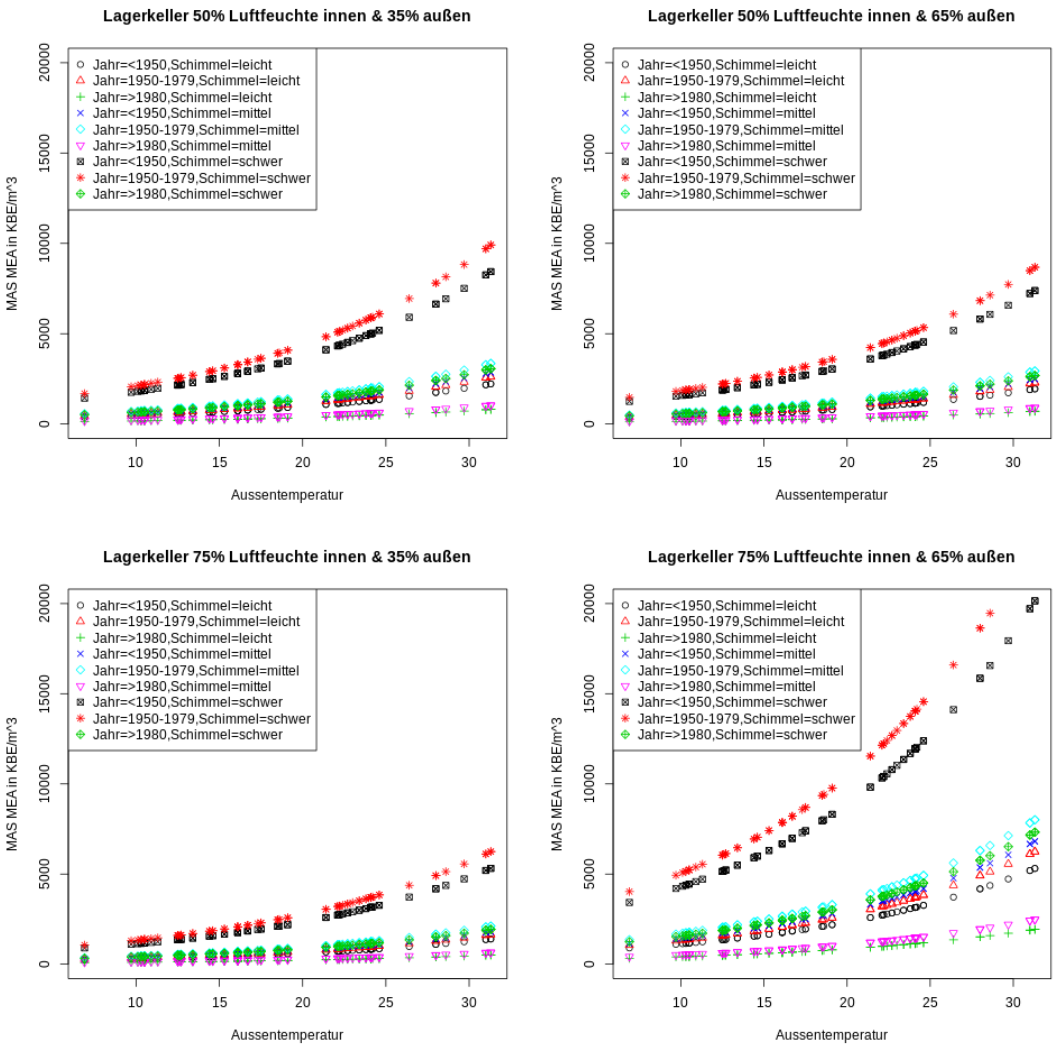

Abb. 5.29: Erwartete Anzahl KBE/m^3 bei MAS100 mit MEA als Nährboden in Lagerkellern bei verschiedenen Luftfeuchtigkeiten.
Unabhängig von jeglichem anderen Faktor steigt die Sporenkonzentration proportional zur Außentemperatur. Zusätzlich zeigt sich, dass trotz schwerem Schimmelbefall die Keimbelastung in Weinkellern, die nach 1980 erbaut wurden, wesentlich geringer ist, als bei älteren Kellern.

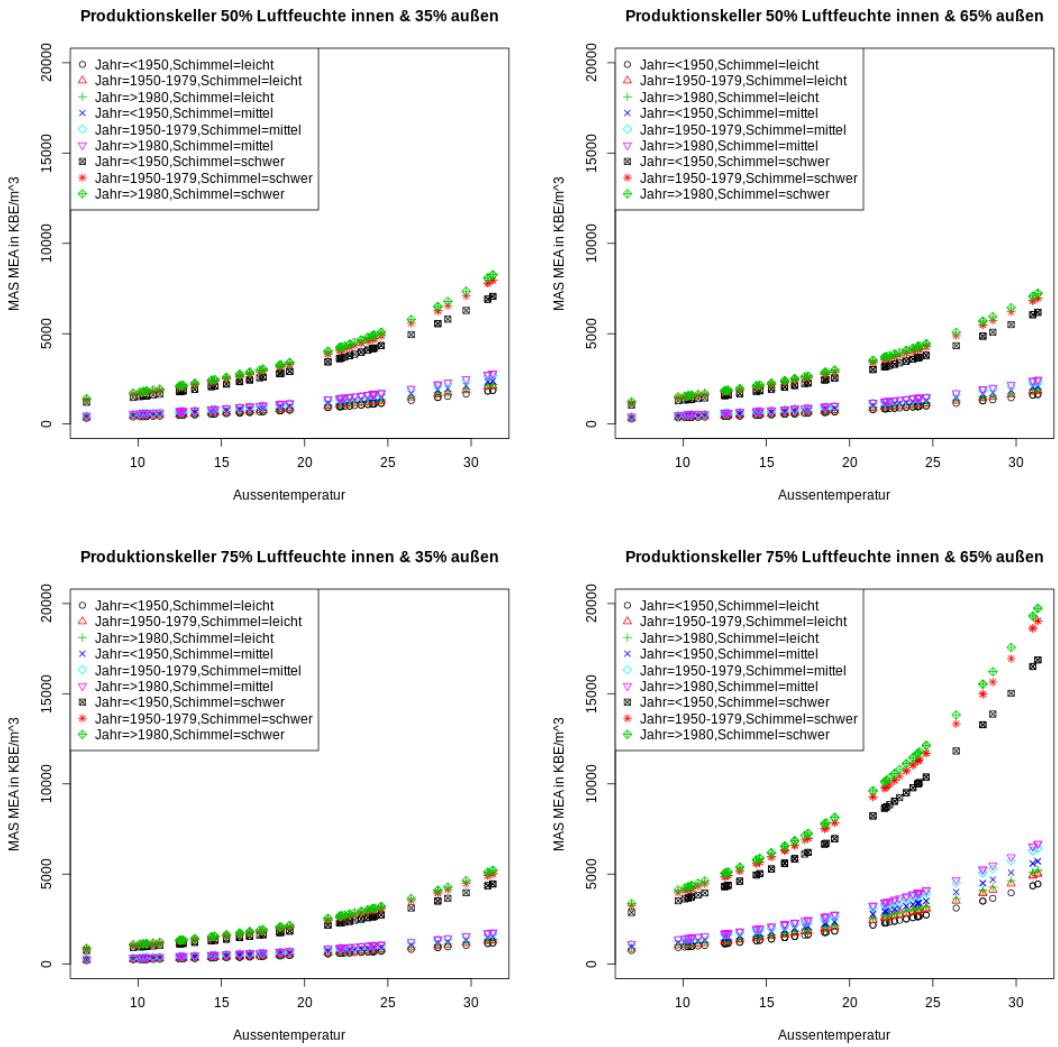

Abb. 5.30: Erwartete Anzahl KBE/m^3 bei MAS100 mit MEA als Nährboden in Produktionskellern bei verschiedenen Luftfeuchtigkeiten.
Wie bei Lagerkellern steigt auch bei Produktionskellern die Sporenkonzentration proportional zur Außentemperatur. Hier ist das Baujahr nicht ausschlaggebend, wenn schwerer Schimmelbefall vorliegt. Die Keimbelastung liegt bei diesem Szenario in Produktionskellern immer auf vergleichsweise höchstem Niveau.

Bei Begutachtung aller Modelle fällt auf, dass bei den Messungen mit dem ACFM Kaskaden-Impaktor nur beim Nährboden DG18 generell eine höhere Anzahl an Schimmelpilzen gefunden wurde, obwohl es durch den Aufbau des Messgerätes wesentlich mehr Partikelgrößen verarbeiten kann als der MAS100. Die Haupteinflussfaktoren sind unabhängig von Messtool

und Nährboden Außentemperatur, Baujahr und Schimmelbefall. Auffällig ist auch, dass die Messungen des MAS100 in beiden Fällen von mehr relevanten Parametern abhängen, als die mit ACFM durchgeführten.

5.2 Sporenkonzentrationen in Wohnräumen

Die zweite Studie, welche im Rahmen dieser Arbeit behandelt wird, befasst sich mit der Schimmelpilzbelastung in Wohnräumen (vgl. Haas et al., 2013). Die über ein Jahr hinweg erhobenen Daten beinhalten unter anderem Messungen der Außen- und Innenluft, vorherrschende Temperatur, sowie Luftfeuchte in- und außerhalb von Wohnungen im Grazer Stadtgebiet. Weiters wurde der Messzeitpunkt in Saisonen eingeteilt: Frühling (März, April, Mai), Sommer (Juni, Juli, August), Herbst (September, Oktober, November) und Winter (Dezember, Januar, Februar). Die Sporenkonzentrationen wurden mit dem MAS100 Air Sampler mit den zwei bereits vorgestellten Nährböden Dichloran Glycerol Agar (DG18) und Malzextrakt Agar (MEA) gemessen. Ziel der nachfolgenden Analysen ist es, den Zusammenhang zwischen Schimmelpilzkonzentrationen in der Innenluft der Wohnungen und den vorliegenden Konzentrationen in der Außenluft zu modellieren. Hierbei werden nur Wohnungen herangezogen, die noch keinen offensichtlichen Schimmelbefall aufweisen.

Der Datensatz beinhaltet 54 Messungen, wobei vorliegende Nullmessungen der KBE/m^3 Luft auf den Wert 1 gesetzt wurden, um die Schätzung unter einem log-linearen Modell gewährleisten zu können. Im Gegensatz zur Weinkellerstudie werden hier die Sporenkonzentrationen nicht getrennt nach Nährböden betrachtet, sondern ein neuer Faktor *naehrboden* eingeführt, der mit „MEA“ und „DG“ kodiert ist. Das bedeutet, die 54 Datenzeilen setzen sich eigentlich aus nur 27 unterschiedlichen Szenarien zusammen, jeweils gemessen mit DG18 und MEA. Betrachtet wird nicht nur die Gesamtanzahl der koloniebildenden Einheiten pro m^3 Luft, sondern im Speziellen auch die Konzentrationen der Gattungen Cladosporium, Penicillium und Aspergillus (allesamt zu xerophilen Schimmelpilzarten zählend).

5.2.1 Explorative Datenanalyse

Um sich einen Überblick zu verschaffen, werden im Folgenden einige Charakteristika der vorliegenden Daten dargestellt.

	innen/aussen	*NB*	*Mean*	*Max*	*Min*
Cladosporium	*innen*	*MEA*	142.8	1600	1
Cladosporium	*innen*	*DG*	211.6	1420	1
Cladosporium	*aussen*	*MEA*	310.4	1600	1
Cladosporium	*aussen*	*DG*	481.2	2620	1
Penicillium	*innen*	*MEA*	117.8	1240	1
Penicillium	*innen*	*DG*	231.6	1340	1
Penicillium	*aussen*	*MEA*	51.7	220	1
Penicillium	*aussen*	*DG*	65.5	280	1
Aspergillus	*innen*	*MEA*	121.7	2000	1
Aspergillus	*innen*	*DG*	126.2	1160	1
Aspergillus	*aussen*	*MEA*	26.2	200	1
Aspergillus	*aussen*	*DG*	51.2	280	1
Gesamt	*innen*	*MEA*	469	2860	20
Gesamt	*innen*	*DG*	665.2	3020	10
Gesamt	*aussen*	*MEA*	581.4	1920	60
Gesamt	*aussen*	*DG*	743.3	3040	20

In der Tabelle ist klar ersichtlich, dass sich grundsätzlich mehr Cladosporium in der Außenluft als in der Innenluft befindet, wohingegen Schimmelpilze der Gattungen Penicillium und Aspergillus vermehrt in den Wohnräumen auftreten. Betrachtet man die Gesamtbelastung an Schimmelpilzarten, so wurden im Mittel in der Außenluft höhere Konzentrationen gemessen. In Abbildung 5.31 werden Median und Quantile der einerseits einzelnen Innen- und Außenmessungen und anderseits Quotienten $\mathrm{KBE}_{innen}/\mathrm{KBE}_{aussen}$ mittels Boxplots veranschaulicht. Zu bemerken ist, dass die Mediane der einzelnen Gattungen innen wie außen weitaus kleiner sind als deren Mittelwerte.

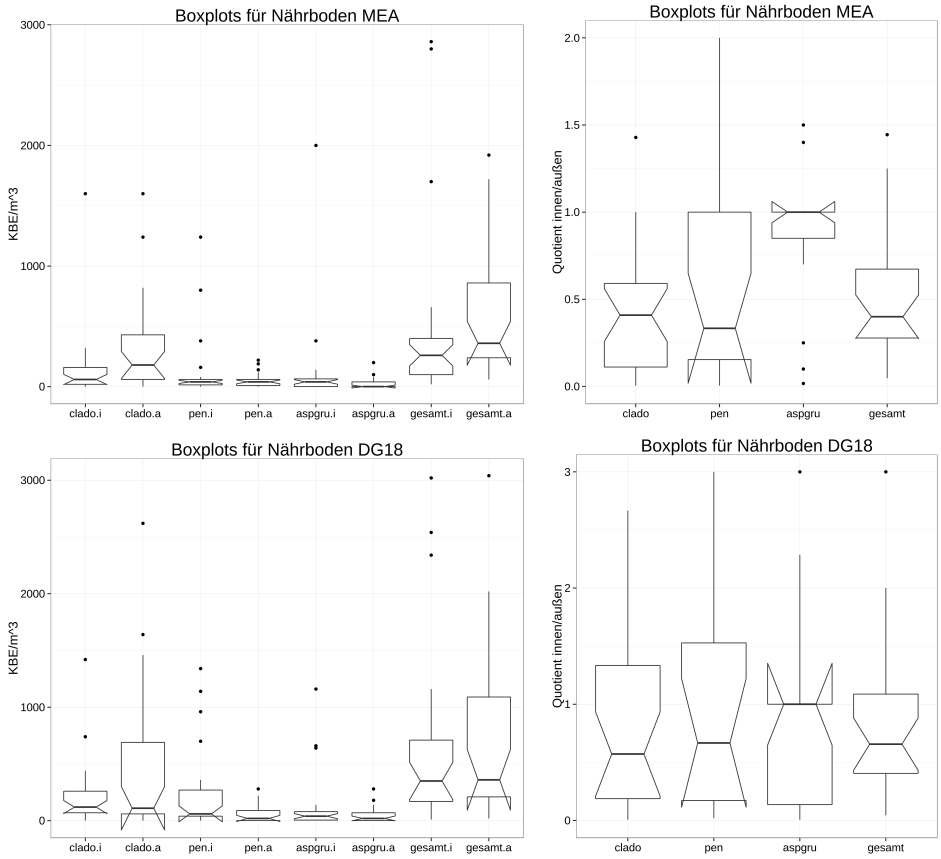

Abb. 5.31: *links:* Darstellung der einzelnen Gattungen getrennt nach Innen- und Außenmessungen.
rechts: Darstellung der Quotienten Innenmessungen zu Außenmessungen gruppiert nach Gattungen (eingeschränkte Y-Skala zur besseren Übersicht).

5.2.2 Saisonale Betrachtungen der Sporenkonzentrationen

Im Folgenden wird nun der Zusammenhang zwischen den Sporenkonzentrationen im Wohnungsinneren in Abhängigkeit von den Sporenkonzentrationen in der Außenluft, dem Nährboden und der vorliegenden Saison mittels loglinearen Modellen untersucht. Der Prädiktor „Konzentration der jeweiligen Pilzgattung in der Außenluft“ fließt logarithmisch als sogenannter *offset* in die Modelle ein und bekommt keinen zu schätzenden Parameter β. So gelingt die Modellierung des Quotienten $\text{KBE}_{innen}/\text{KBE}_{aussen}$. Es werden also Modelle der Form

$$\log\left(\frac{\mathbb{E}(\text{KBE}_{innen})}{\text{KBE}_{aussen}}\right) = \beta_0 + \beta_{naehrboden} + \beta_{saison} + \beta_{naehrboden:saison}$$

geschätzt. Um die passende Varianzannahme zu treffen, wird wieder die Theorie der EQL-Modelle herangezogen.

Regressionsmodell für Cladosporium

Schon in den Boxplots der einzelnen Saisonen (Abb. 5.32) erahnt man die Notwendigkeit der Interaktion von Nährboden und Saison, da sich vorallem bei DG18 der Winter stark von den anderen Saisonen unterscheidet.

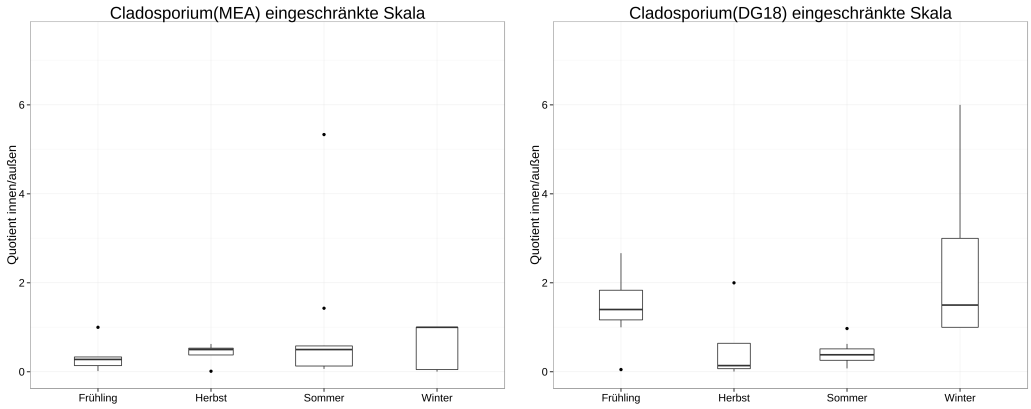

Abb. 5.32: Boxplots des Quotienten der Cladosporium-Konzentrationen über die Saisonen aufgeteilt nach Nährboden. Die Breite der Boxen wird durch die Anzahl der Messwerte bestimmt.

Der Aufruf der Funktion *eql*() und der dazugehörige Profilplot (Abb. 5.33) liefern einen starken Hinweis auf die Adäquatheit einer quadratischen Varianzannahme.

```
EQL-Max: -314.606 at: theta = 2.020
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -380.8  -356.5  -344.9  -345.5  -336.5  -314.6
```

Also wird im nächsten Schritt ein Gamma-Modell gefittet und auf Signifikanz der einzelnen Parameter getestet. Es resultiert das folgende, unter der quadratischen Varianzannahme, optimale Modell.

```
glm(formula = clado.i ~ offset(log(clado.a)) + saison * naehrboden,
    family = Gamma(link = "log"), subset = complete.cases(daten))
```

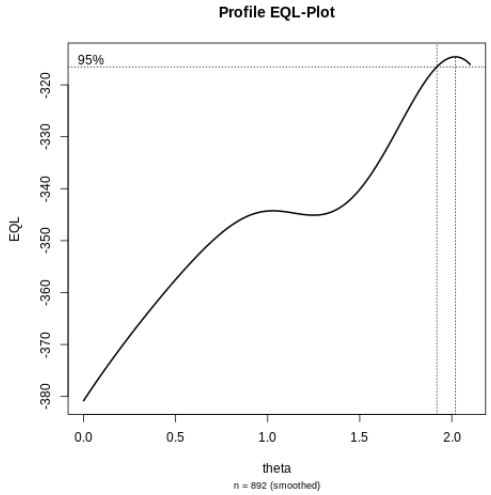

Abb. 5.33: Profilplot des vollen Modells für Cladosporium.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2489  -1.7666  -0.4263   0.3061   2.7783

                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.3683     0.5638   0.653  0.51690
saisonh                -0.9284     0.9350  -0.993  0.32596
saisons                -1.2489     0.7213  -1.732  0.09005 .
saisonw                 2.7228     0.8735   3.117  0.00314 **
naehrbodenMEA           1.8132     0.7974   2.274  0.02768 *
saisonh:naehrbodenMEA  -2.1463     1.3223  -1.623  0.11139
saisons:naehrbodenMEA  -1.0509     1.0200  -1.030  0.30828
saisonw:naehrbodenMEA  -5.3971     1.2353  -4.369 7.04e-05 ***
---
(Dispersion parameter for Gamma family taken to be 2.225322)

    Null deviance: 232.69  on 53  degrees of freedom
Residual deviance: 117.52  on 46  degrees of freedom
AIC: 665.19
```

Das finale Modell spiegelt mit einer skalierten Deviance von 52.8 bei 46 Freiheitsgraden eine gute Anpassung wider. Als Referenzklasse wählt R standardmäßig die Faktorstufe mit dem minimalen alphanumerischen Wert, in diesem Fall den Nährboden DG und die Saison Frühling. Der Intercept beschreibt in Modellen mit offset den Fall der Referenzklasse in Kombination mit $offset = 0$, in diesem Fall also die erwartete Sporenkonzentration ge-

messen mit Nährboden DG18 im Frühling bei einer Außenbelastung von 1 KBE/m^3 Luft. Weiters kann man aus dieser Darstellung schließen, dass die Sporenkonzentration im Winter bei Verwendung von DG18 als Nährboden signifikant höher ist als im Frühling und mit zunehmender Außenbelastung generell ansteigt. Sommer und Herbst unterscheiden sich nicht maßgeblich vom Frühling, im Mittel sind die Konzentrationen jedoch geringer (vorallem mit MEA als Nährboden). Will man die Referenzklasse ändern, so gelingt dies mit dem Befehl *relevel*(). In Abbildung 5.34 wird das finale Modell veranschaulicht und man erkennt, dass der Einfluss des Nährbodens maßgeblich für die erwarteten Messergebnisse verantwortlich ist.

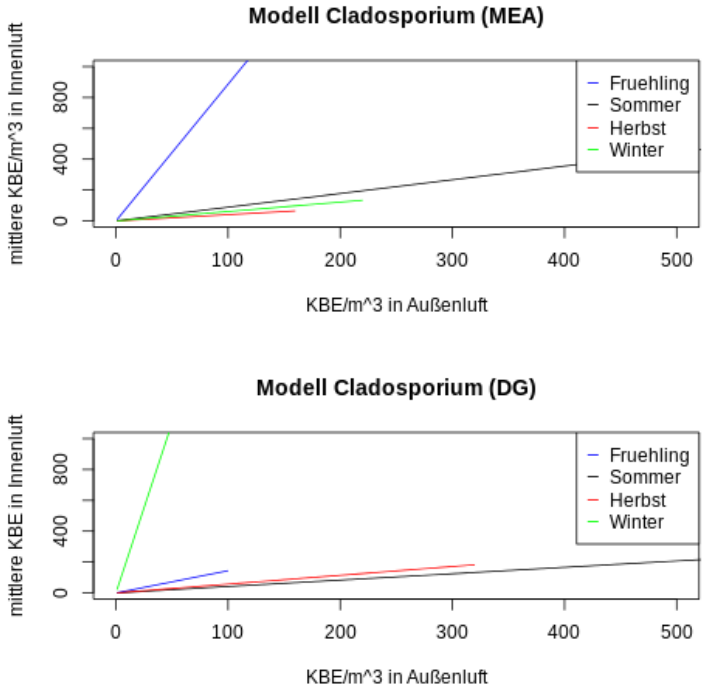

Abb. 5.34: Mittlere Anzahl KBE/m^3 Luft der Gattung Cladosporium in Abhängigkeit von der Außenkonzentration, der Saison und dem Nährboden. Die unterschiedliche Länge der einzelnen Geraden ergibt sich aus den verschieden angenommenen Maximalwerten.

Regressionsmodell für Penicillium

Auch für Penicillium erkennt man schon in den Boxplots der einzelnen Saisonen (Abb. 5.35) die potenzielle Signifikanz der Interaktion von Nährboden und Saison, da sich vorallem bei MEA der Herbst stark von den anderen

Saisonen unterscheidet.

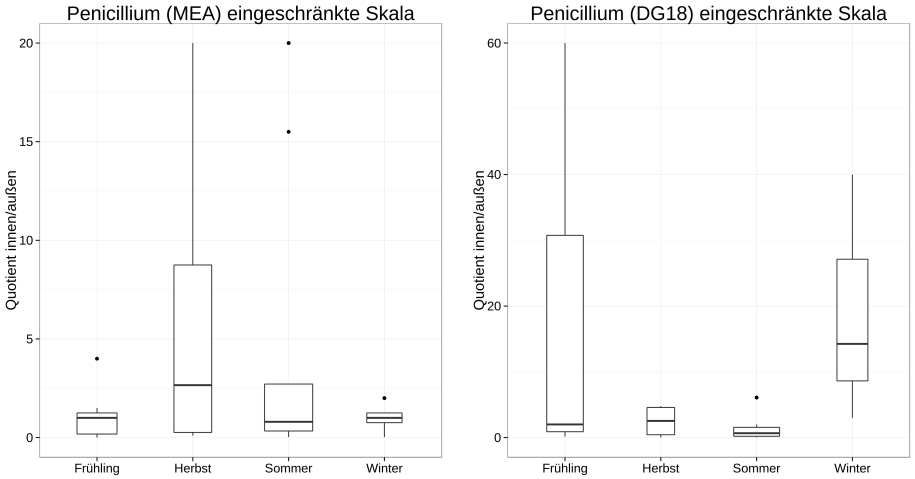

Abb. 5.35: Boxplots des Quotienten der Penicillium-Konzentrationen über die Saisonen aufgeteilt nach Nährboden. Die Breite der Boxen wird durch die Anzahl der Messwerte bestimmt.

Der Aufruf der Funktion *eql*() und der dazugehörige Profilplot (Abb. 5.36) bekräftigen auch hier eine quadratische Varianzannahme.

```
EQL-Max: -311.0581 at: theta = 2.020
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -376.7  -369.1  -363.8  -354.5  -343.2  -311.1
```

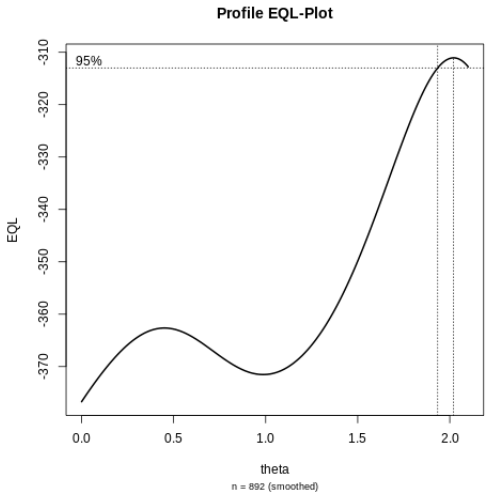

Abb. 5.36: Profilplot des vollen Modells für Penicillium.

Das optimale Gamma-Modell setzt sich aus den folgenden Parametern zusammen und bestätigt die Interpretation der Boxplots.

```
glm(formula = pen.i ~ offset(log(pen.a)) + saison * naehrboden,
    family = Gamma(link = "log"), subset = complete.cases(daten))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3382  -2.4313  -1.6055   0.2062   3.2488

                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.1961     0.6981   6.011 2.79e-07 ***
saisonh                -3.2908     1.1577  -2.842  0.00665 **
saisons                -0.6213     0.8930  -0.696  0.49008
saisonw                 1.1193     1.0815   1.035  0.30612
naehrbodenMEA          -4.0799     0.9873  -4.132  0.00015 ***
saisonh:naehrbodenMEA   5.0236     1.6372   3.068  0.00360 **
saisons:naehrbodenMEA   3.1888     1.2629   2.525  0.01508 *
saisonw:naehrbodenMEA   0.9398     1.5295   0.614  0.54195
---
(Dispersion parameter for Gamma family taken to be 3.411618)

    Null deviance: 323.31  on 53  degrees of freedom
Residual deviance: 221.78  on 46  degrees of freedom
AIC: 669.39
```

Die Anpassung dieses Modells ist mit einer Deviance von 65 bei 46 Freiheitsgraden und einer Pearson X^2-Statistik von 156.9 nicht ganz zufriedenstellend, jedoch liegt der Grund dafür wahrscheinlich in der geringen Datenmenge. Zur Erleichterung der Interpretation dieses Modells dient Abbildung 5.37. Die Signifikanz des Nährbodens und dessen unterschiedliche Abhängigkeiten von den saisonalen Bedingungen ist klar ersichtlich.

Regressionsmodell für Aspergillus

Auch für Aspergillus kann man in den Boxplots der einzelnen Saisonen (Abb. 5.38) eine potenzielle Signifikanz der Interaktion von Nährboden und Saison erahnen, da sich vorallem bei MEA der Winter stark von den anderen Saisonen unterscheidet.

Die *eql*()-Funktion und der dazugehörige Profilplot (Abb. 5.39) liefern ein 95%-Konfidenzintervall, das den Wert 2 noch knapp enthält, somit kann

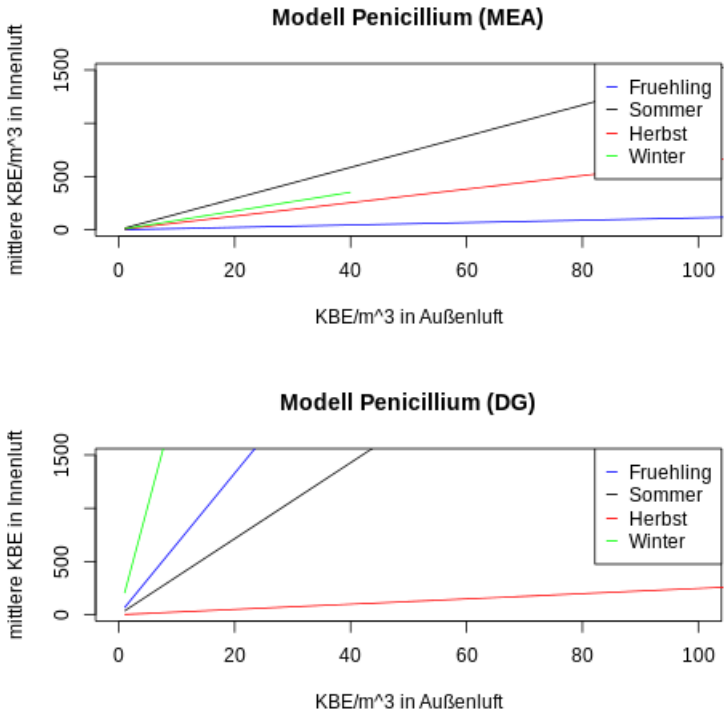

Abb. 5.37: Mittlere Anzahl KBE/m^3 Luft der Gattung Penicillium in Abhängigkeit von der Außenkonzentration, der Saison und dem Nährboden. Die verschiedenen Längen der einzelnen Geraden ergibt sich aus den unterschiedlich angenommenen Maximalwerten.

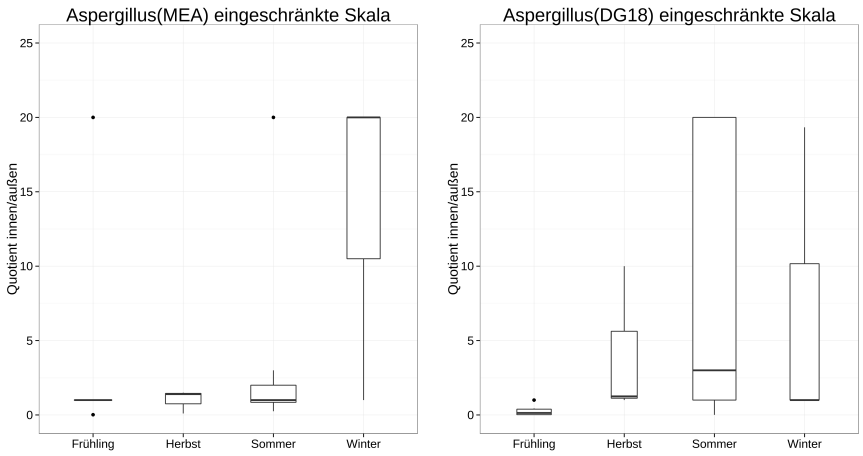

Abb. 5.38: Boxplots des Quotienten der Aspergillus-Konzentrationen über die Saisonen aufgeteilt nach Nährboden. Die Breite der Boxen wird durch die Anzahl der Messwerte bestimmt.

auch hier eine quadratische Varianzfunktion für das beschreibende Modell herangezogen werden.

```
EQL-Max: -275.3454 at: theta = 2.109
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -354.2  -346.7  -331.9  -326.0  -310.1  -275.3
```

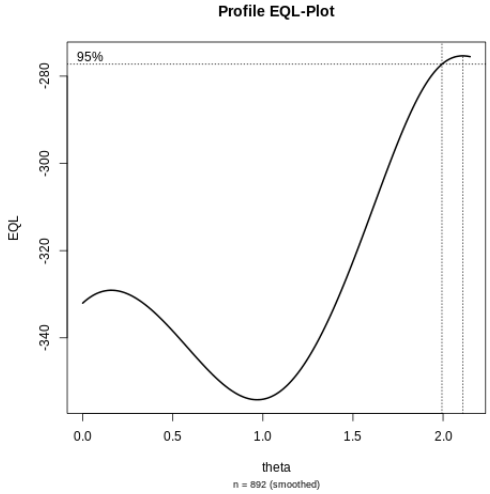

Abb. 5.39: Profilplot des vollen Modells für Aspergillus.

Beim ersten Rechnen des vollen Gamma-Modells fällt auf, dass in der Varianzzerlegung die Interaktion zwischen Nährboden und Saison keine Signifikanz zeigt.

```
                  Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                                 53     269.25
saison             3   41.859        50     227.40 3.6266 0.01971 *
naehrboden         1    3.534        49     223.86 0.9186 0.34286
saison:naehrboden  3    7.215        46     216.65 0.6251 0.60244
```

Auch der Nährboden fällt mit einem p-Wert von 0.34 nicht unter das vorgegebene Signifikanzniveau von $\alpha = 0.05$ und ist somit nicht maßgeblich für die Sporenkonzentration in der Raumluft. Es resultiert das folgende endgültige Modell.

```
glm(formula = aspgru.i ~ saison + offset(log(aspgru.a)),
    family = Gamma(link = "log"),subset = complete.cases(daten))
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9719  -2.4354  -1.4038  -0.2506   4.2920

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8239     0.6353   2.871  0.00599 **
saisonh       0.7530     1.0536   0.715  0.47813
saisons       2.1223     0.8127   2.611  0.01187 *
saisonw       2.4833     0.9842   2.523  0.01486 *
---
(Dispersion parameter for Gamma family taken to be 5.650824)

    Null deviance: 269.26  on 53  degrees of freedom
Residual deviance: 227.40  on 50  degrees of freedom
AIC: 597.28
```

Die Modellgüte ist mit einer skalierten Deviance von 40.2 bei 50 Freiheitsgraden sehr gut, wobei die Deviance-Residuen etwas dezentriert sind. Betrachtet man die Schätzer der einzelnen Faktoren, so fällt schnell auf, dass bei gleichbleibender Außenbelastung die Sporenkonzentration im Wohnraum im Winter am höchsten sind und zwar um einen Faktor von $e^{2.48} = 11.9$ Mal höher als im Frühling bei gleichen Konditionen. Veranschaulicht wird dieses Modell in Abbildung 5.40.

Regressionsmodell für die Gesamtkonzentration

Betrachtet man die Gesamtkonzentration aller Schimmelpilzgattungen, so zeigen die Boxplots über die einzelnen Saisonen (Abb. 5.41) eine sehr große Konzentrationsdifferenz zwischen Außen- und Innenmessungen in den Wintermonaten, falls der Nährboden DG18 zur Züchtung der Proben verwendet wird.

Der Aufruf von *eql*() führt zu folgendem Ergebnis und bekräftigt die Annahme einer quadratischen Varianzfunktion zur Modellschätzung.

```
EQL-Max: -377.9551 at: theta = 2.065
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -425.7  -406.1  -389.3  -393.9  -380.3  -378.0
```

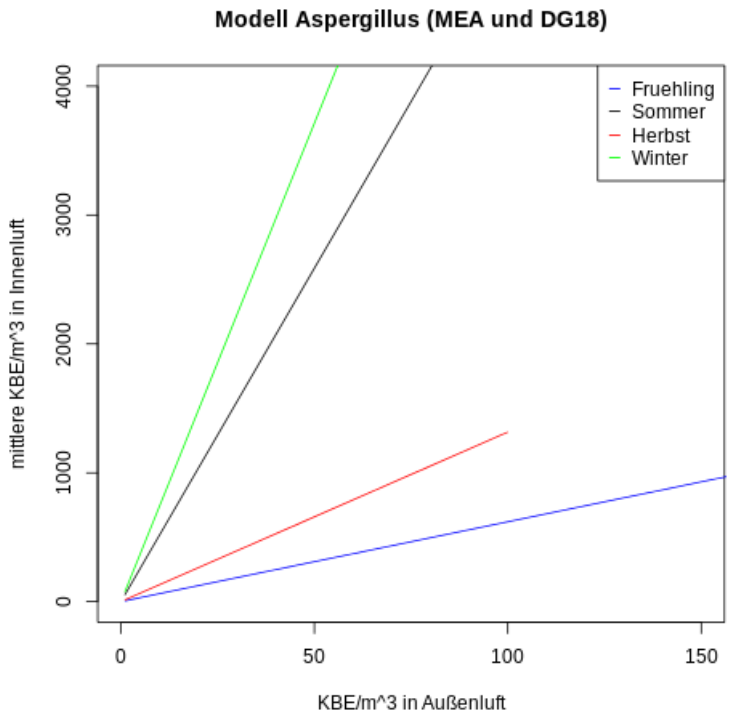

Abb. 5.40: Mittlere Anzahl KBE/m^3 Luft der Gattung Aspergillus in Abhängigkeit von Außenkonzentration und Saison. Der Nährboden ist nicht signifikant, daher gibt es keine Unterscheidung der Messwerte.

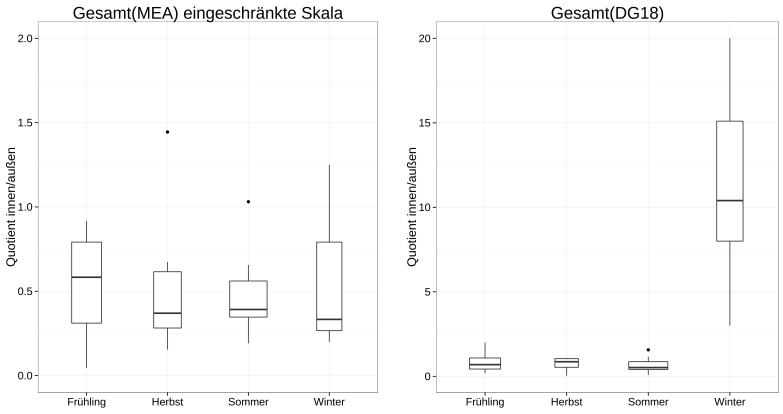

Abb. 5.41: Boxplots der Quotienten der Gesamtkonzentration über die Saisonen aufgeteilt nach Nährböden. Die Breite der Boxen wird durch die Anzahl der Messwerte bestimmt.

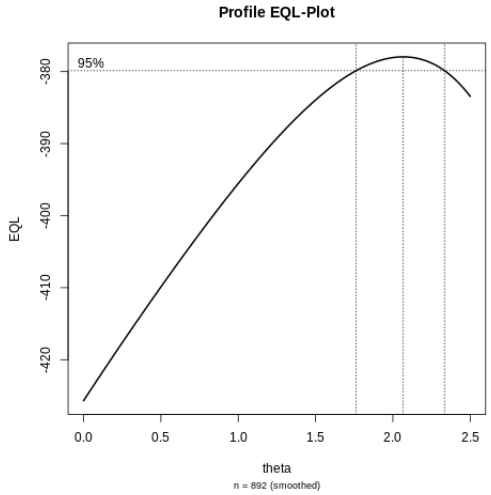

Abb. 5.42: Profilplot des vollen Modells für die Gesamtkonzentration.

Studiert man die Gesamtheit der Sporenkonzentrationen der Wohnräume in Abhängigkeit der Außenkonzentrationen so ist unter der Annahme eines quadratischen Varianzmodells die Interaktion Nährboden mit Saison mit einem p-Wert von 0.46 bzw. der Nährboden selbst mit einem p-Wert von 0.55 als nicht signifikant einzustufen. Es resultiert, wie schon bei der Gattung Aspergillus, ein rein von der Saison abhängiges Regressionsmodell mit einer skalierten Deviance von 63.6 bei 50 Freiheitsgraden.

```
glm(formula = gesamt.i ~ saison + offset(log(gesamt.a)),
    family = Gamma(link = "log"),subset = complete.cases(daten))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3569  -0.7237  -0.2033   0.3255   1.5629

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.37010    0.22598  -1.638    0.108
saisonh     -0.01665    0.37474  -0.044    0.965
saisons      0.10469    0.28907   0.362    0.719
saisonw      2.51461    0.35008   7.183 3.11e-09 ***
---
(Dispersion parameter for Gamma family taken to be 0.7149124)

    Null deviance: 114.64  on 53  degrees of freedom
Residual deviance:  45.47  on 50  degrees of freedom
AIC: 776.2
```

Den größten Einfluss haben, wie schon in den Boxplots ersichtlich, die Wintermonate mit einer Zuwachsrate von $e^{2.51} = 12.3$ im Vergleich zum Frühling bei gleichbleibender Außensporenkonzentration. Abbildung 5.43 skizziert den saisonalen Verlauf der Wohnraumkonzentrationen in Abhängigkeit von der Außenluft und zeigt, wie schon aus den Modellparametern ersichtlich, einen sehr geringen Einfluss der Monate März bis November.

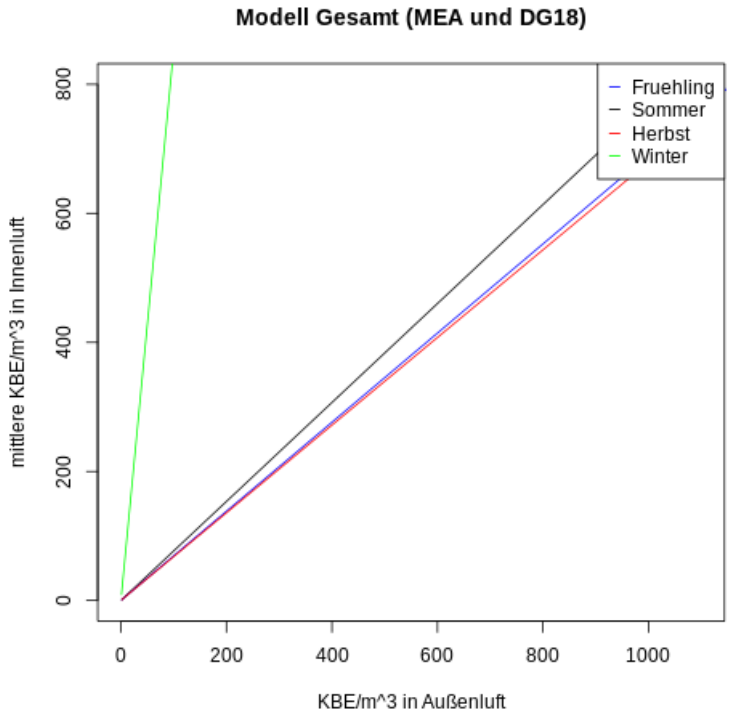

Abb. 5.43: Mittlere Anzahl aller KBE/m^3 Luft in Abhängigkeit von der Außenkonzentration und den Saisonen. Der Nährboden ist nicht signifikant, daher gibt es keine Unterscheidung der Messwerte.

5.2.3 Witterungsabhängige Modellierung der Sporenkonzentrationen

Da die Einteilung eines Jahres in vier Saisonen einer sehr groben Klassifizierung entspricht und auch nicht jeder Monat über die Jahre hinweg immer dieselben Witterungsbedingungen aufzeigt, wird nun versucht, die Sporenkonzentration in Grazer Wohnungen mittels Außentemperatur und Außenluftfeuchte zu beschreiben. Die Herangehensweise ist dieselbe wie schon in Abschnitt 5.2.2 zuvor. Es wird der Quotient $\text{KBE}_{innen}/\text{KBE}_{aussen}$ der drei

verschiedenen Schimmelpilzgattungen Cladosporium, Penicillium und Aspergillus unter Annahme einer quadratischen Varianzfunktion mit log-linearer Linkfunktion in Abhängigkeit von Temperatur, Feuchte und Nährboden modelliert.

Regressionsmodell für Cladosporium

Vor dem Rechnen des ersten Modells werden die Abhängigkeiten der Konzentrationsquotienten explorativ untersucht. In Abbildung 5.44 ist einerseits die Außentemperatur und andererseits die Außenfeuchte gegen den logarithmierten Quotienten aufgetragen. Anhand der Glättungsfunktion ist gut zu erkennen, dass wahrscheinlich keine rein lineare Abhängigkeit von der Luftfeuchte vorliegt. Ein leicht fallender Trend ist bei steigender Temperatur zu erahnen.

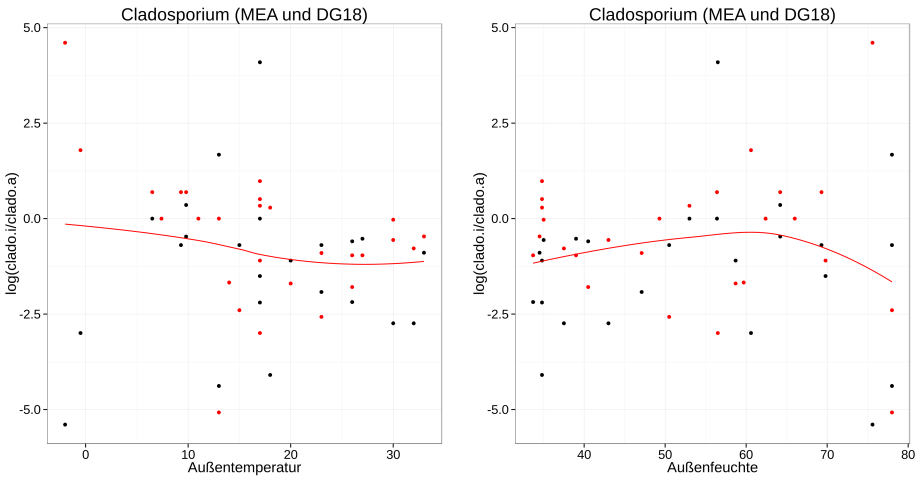

Abb. 5.44: Scatterplot der logarithmierten Quotienten KBE Innenluft/Außenluft von Cladosporium in Abhängigkeit von Außentemperatur (links) und Außenluftfeuchte (rechts). Die rote Linie beschreibt eine loess-Glättungsfunktion und die roten Punkte entsprechen Messungen mit DG18 als Nährboden bzw. ist MEA als Nährboden mit schwarzen Punkten dargestellt.

Die Datenbeschaffenheit macht es leider notwendig die minimalen Werte von 1 KBE/m^3 der Responsevariablen um 5 KBE/m^3 zu erhöhen, um numerische Probleme bei der Berechnung des Modells und somit Divergenz des Algorithmus zu verhindern. Das passende Gamma-Modell wird auch mittels Vorwärtsselektion gesucht, da der Schätzalgorithmus bei einem vollen Modells nicht konvergiert. Gestartet wird mit den Kovariablen Außentem-

peratur, Außenfeuchte und Nährboden ohne Interaktionen oder etwaige Potenzierungen. Die Hinzugabe von Interaktionen oder quadratischer bzw. kubischer Luftfeuchte führt in allen Fällen zu Divergenz. Quadratische bzw. kubische Außentemperatur wird genauso wie der Nährboden mit p-Werten von 0.67, 0.73 bzw. 0.77 als nicht signifikant ausgegeben. Auch die Außenfeuchte scheint keine maßgeblichen Einfluss auf den Quotienten zu haben.

```
Model 1: I(clado.i + 5) ~ temp.a + feuchte.a + offset(log(clado.a))
Model 2: I(clado.i + 5) ~ temp.a + offset(log(clado.a))
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        51     145.91
2        52     147.04 -1  -1.1248 0.0923 0.7625
```

Somit resultiert das folgende minimale Modell mit einer skalierten Deviance von 11.13 bei 52 Freiheitsgraden, jedoch einem Pearson'schen X^2 von 686.92, was einem optimalen Modell widerspricht. Die mangelhafte Datenanpassung ist aber wahrscheinlich auch auf die geringe Zahl an Datensätzen zurückzuführen.

```
glm(formula = I(clado.i + 5) ~ temp.a + offset(log(clado.a)),
    family = Gamma(link = "log"), subset = complete.cases(daten))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3836  -1.5625  -1.0962  -0.3413   6.6559

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.88762    1.06530   2.711  0.00908 **
temp.a      -0.11691    0.05439  -2.150  0.03626 *
---
(Dispersion parameter for Gamma family taken to be 13.21012)

    Null deviance: 209.16  on 53  degrees of freedom
Residual deviance: 147.04  on 52  degrees of freedom
AIC: 704.27
```

Die mittlere Sporenkonzentration von Schimmelpilzen der Gattung Cladosporium hängt also, entgegen der Ergebnisse der explorativen Analyse, bei gleichbleibender Konzentration der Außenluft nur von der Außentemperatur ab. Unter diesem Modell fällt die Konzentration in der Innenluft mit steigenden Temperaturen exponentiell ab (siehe Abb. 5.45).

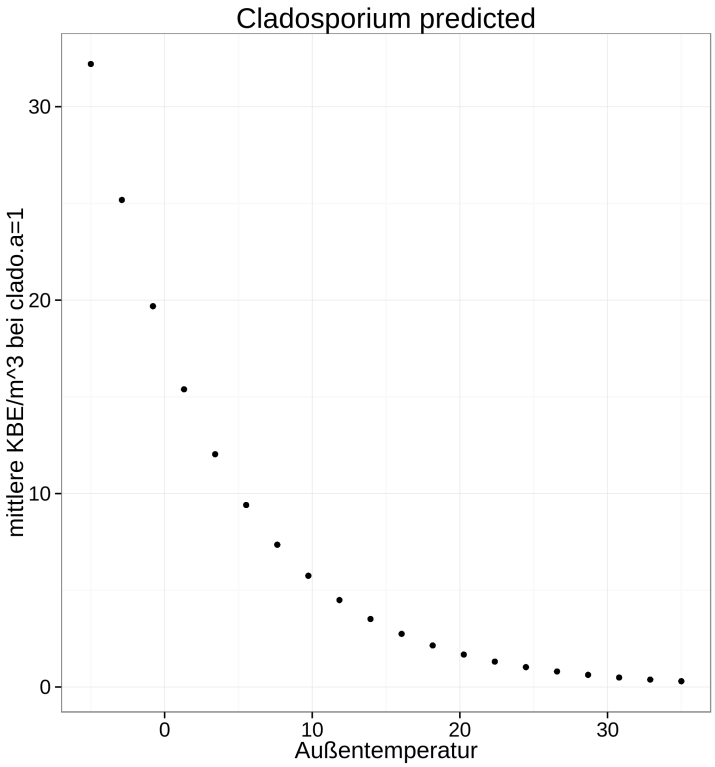

Abb. 5.45: Mittlere Anzahl KBE/m^3 Luft von Pilzen der Gattung Cladosporium in Abhängigkeit der Außentemperatur bei fixiertem Wert der Sporenkonzentration in der Außenluft. Der Nährboden ist nicht signifikant, daher gibt es keine Unterscheidung der Messwerte.

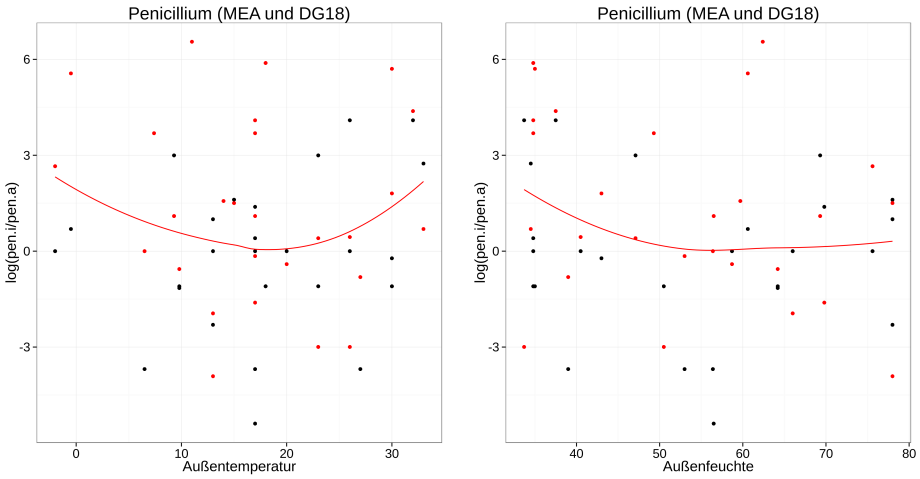

Abb. 5.46: Scatterplot der logarithmierten Quotienten KBE Innenluft/Außenluft von Penicillium in Abhängigkeit von Außentemperatur (links) und Außenluftfeuchte (rechts). Die rote Linie beschreibt eine loess-Glättungsfunktion und die roten Punkte entsprechen Messungen mit DG18 als Nährboden bzw. ist MEA als Nährboden mit schwarzen Punkten dargestellt.

Regressionsmodell für Penicillium

Für Schimmelpilze der Gattung Penicillium wird auch zuerst die Abhängigkeit des logarithmierten Quotienten von der Außentemperatur bzw. der Außenfeuchte betrachtet (siehe Abb. 5.46). Betrachtet man die Glättungsfunktion, so liegt die Vermutung nahe, dass die Außentemperatur quadratisch in das Modell einfließt und mit steigender Luftfeuchtigkeit die Sporenkonzentration abnimmt, sofern die Steigung bei am linken Wertebereich überhaupt eine Signifikanz zeigt.

Auch in diesem Fall scheint der Scatterplot irreführend zu sein, es ergibt sich nämlich ein Modell mit keinen witterungsbedingten Abhängigkeiten.

```
glm(formula = I(pen.i) ~ naehrboden + offset(log(pen.a)),
   family = Gamma(link = "log"), subset = complete.cases(daten))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.784  -2.598  -1.808  -0.512   3.667

             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.2457     0.4094  10.370 2.92e-14 ***
naehrbodenMEA -2.0678     0.5790  -3.571 0.000775 ***
```

```
---
(Dispersion parameter for Gamma family taken to be 4.525727)
    Null deviance: 323.31  on 53  degrees of freedom
Residual deviance: 273.65  on 52  degrees of freedom
AIC: 673.86
```

Abbildung 5.47 illustriert den gefundenen Zusammenhang.

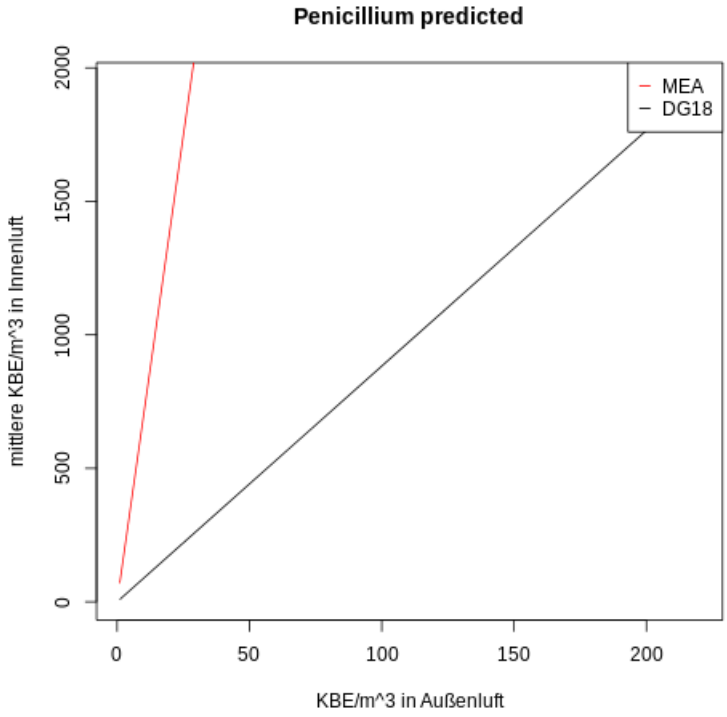

Abb. 5.47: Mittlere Anzahl KBE/m^3 Luft von Pilzen der Gattung Penicillium in Abhängigkeit von den Nährböden.

Regressionsmodell für Aspergillus

Die Gattung Aspergillus zeigt auch wieder eine quadratische Abhängigkeit des logarithmierten Quotienten von der Außentemperatur (siehe Abb. 5.48). Bei Betrachtung des rechten Plots muss eine signifikante Abhängigkeit von der Außenluftfeuchte nicht unbedingt gegeben sein.

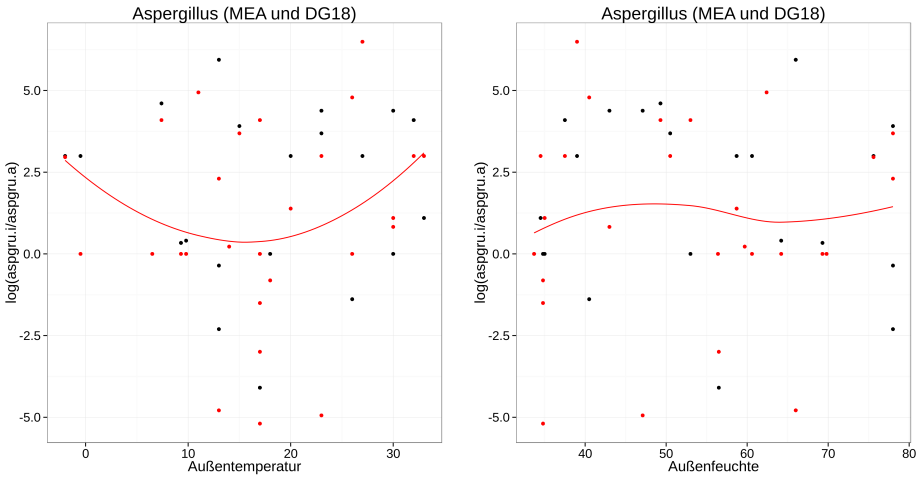

Abb. 5.48: Scatterplot der logarithmierten Quotienten KBE Innenluft/Außenluft von Aspergillus in Abhängigkeit von Außentemperatur (links) und Außenluftfeuchte (rechts). Die rote Linie beschreibt eine loess-Glättungsfunktion und die roten Punkte entsprechen Messungen mit DG18 als Nährboden bzw. ist MEA als Nährboden mit schwarzen Punkten dargestellt.

In diesem Fall scheint der Quotient der Sporenkonzentrationen von Aspergillus keinerlei Abhängigkeiten von Witterung und Nährboden zu haben, es resultiert ein Intercept-only Modell. Anscheinend ist anhand dieser Datenbasis lediglich eine saisonale Komponente ausschlaggebend, wie in Abschnitt 5.2.2 bereits besprochen wurde.

```
glm(aspgru.i ~ offset(log(aspgru.a)), family = Gamma(link = "log"),
    subset = complete.cases(daten))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9600  -2.3124  -2.1514  -0.5881   5.1675

            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.648      0.370   9.859 1.38e-13 ***
---
(Dispersion parameter for Gamma family taken to be 7.391511)

    Null deviance: 269.26  on 53  degrees of freedom
Residual deviance: 269.26  on 53  degrees of freedom
AIC: 604.53
```

Regressionsmodell für Gesamtkonzentration

In der Gesamtheit der Sporenkonzentrationen zeigt sich ein abnehmender Trend der betrachteten Konzentrationen mit steigender Außentemperatur (siehe Abb. 5.49). Die Luftfeuchtigkeit scheint keinen Einfluss zu haben.

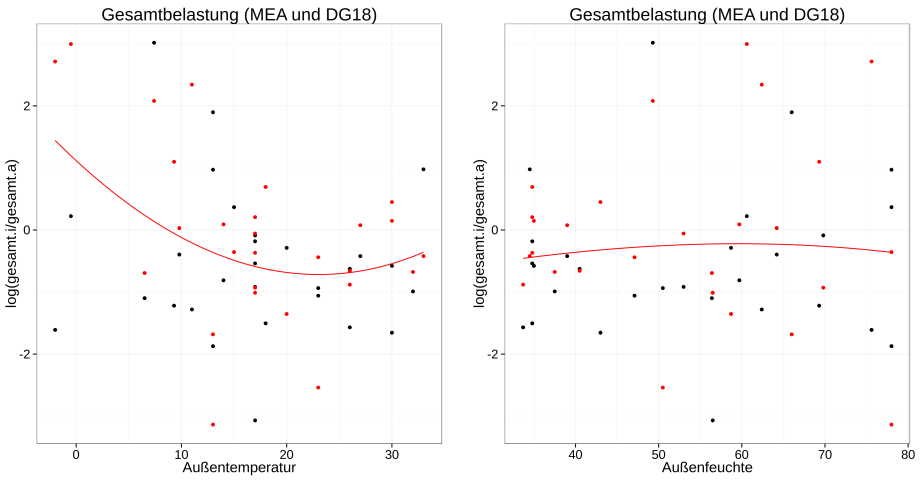

Abb. 5.49: Scatterplot der logarithmierten Quotienten KBE Innenluft/Außenluft ohne Unterteilung in einzelne Gattungen in Abhängigkeit von Außentemperatur (links) und Außenluftfeuchte (rechts). Die rote Linie beschreibt eine loess-Glättungsfunktion und die roten Punkte entsprechen Messungen mit DG18 als Nährboden bzw. ist MEA als Nährboden mit schwarzen Punkten dargestellt.

Das endgültige Regressionsmodell stimmt gut mit den Ergebnissen der explorativen Analyse überein und zeigt die Signifikanz der Außentemperatur und ihren quadratischen Einfluss.

```
glm(gesamt.i ~ temp.a + I(temp.a^2) + offset(log(gesamt.a)),
    family = Gamma(link = "log"), subset = complete.cases(daten))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.78337  -0.97754  -0.41792   0.08054   2.49503


             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.761860   0.492416   5.609 8.35e-07 ***
temp.a      -0.238303   0.059747  -3.989 0.000213 ***
I(temp.a^2)  0.004494   0.001705   2.636 0.011078 *
---
```

```
(Dispersion parameter for Gamma family taken to be 1.548664)

    Null deviance: 114.640  on 53  degrees of freedom
Residual deviance:  68.832  on 51  degrees of freedom
AIC: 800.17
```

Eine skalierte Deviance von 44.45 bei 51 Freiheitsgraden und die Pearson-Statistik $X^2 = 79$ weisen auch auf eine gute Modellanpassung hin. Abbildung 5.50 veranschaulicht das gefundene Modell. Die Gesamtheit der Sporenkonzentrationen nimmt mit zunehmender Außentemperatur exponentiell ab.

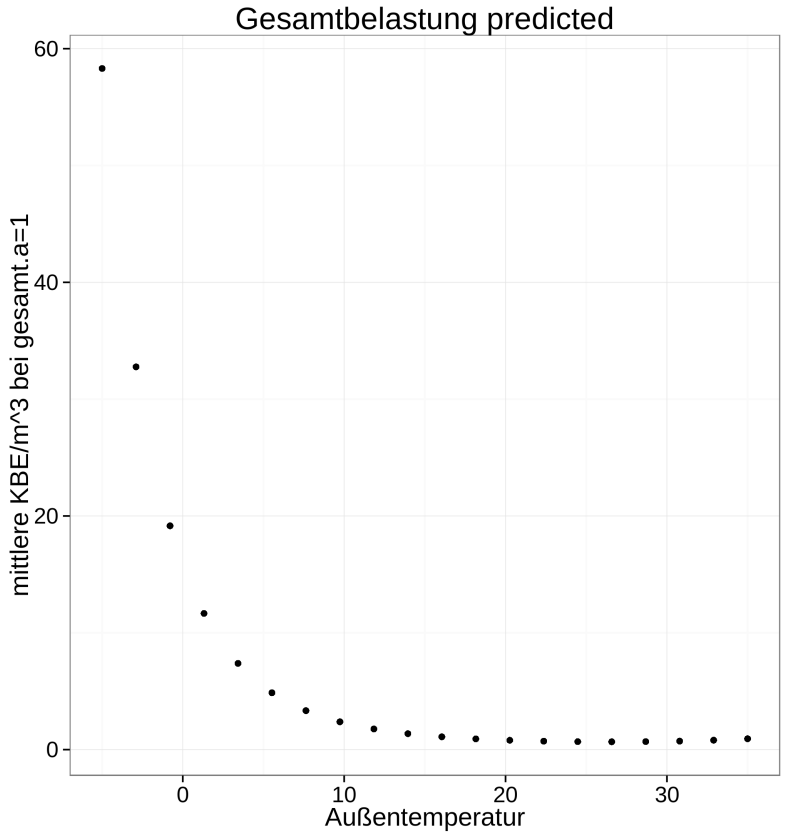

Abb. 5.50: Mittlere Anzahl KBE/m^3 Luft ohne Aufteilung in Untergattungen in Abhängigkeit der Außentemperatur bei fixierter Außenbelastung.

Grundsätzlich ist festzustellen, dass bei allen untersuchten Schimmelpilzgattungen der Nährboden, in dem die Proben kultiviert wurden, ausschlaggebend für die Höhe der Konzentration ist, vorallem in Verbindung mit einer saisonalen Komponente. Entgegen der meteorologischen Theorie überwiegt auch der Einfluss der kalendarischen Monate gegenüber Temperatur

und Luftfeuchte. Dieses Ergebnis könnte aber auch gut mit der geringen Datenmenge erklärt werden. Die berechneten Modelle der Gesamtsporenkonzentration führen eher zu nachvollziehbaren Resultaten.

6 Fazit

In den vorangegangenen Kapiteln wurden nützliche Werkzeuge zur Modellierung von Daten, im Speziellen Anzahlen, detailliert vorgestellt. Vom generalisierten linearen Modell mit zugrunde liegender Exponentialfamilie, über den Quasi-Likelihood-Ansatz bis hin zur Extended Quasi-Likelihood-Funktion, wurde versucht der Problematik der Überdispersion Herr zu werden.

Hauptaugenmerk dieser Arbeit lag vorallem in der Umsetzung der beschriebenen Theorien anhand der zwei mikrobiologischen Datensätze. Das Durchexerzieren aller theoretischen Modelle anhand der Messungen in den Weinkellern diente der Veranschaulichung der potentiell auftretenden Probleme. Eine große Schwierigkeit in der Variablenselektion mittels Varianzanalyse barg das Verlassen der Exponentialfamilie und der Übergang zu Quasi-Verteilungen. Für diesen Fall gab es keine implementierte automatische Funktion, da *R* standardmäßig kein AIC oder ähnliches Informationskriterium berechnen kann. Die Daten der Weinkellerstudie führten uns auch an die theoretischen Grenzen dieser Arbeit und gaben einen kleinen Ausblick in die Welt der exponentiellen Dispersionsmodelle, die hier nicht näher beschrieben wurden.

Kritisch hinterfragt werden sollten die resultierenden Modelle zur Studie der Sporenkonzentration in Grazer Wohnräumen, da sie aufgrund der geringen Anzahl an Datensätzen vorallem bei den Betrachtungen der Untergattungen zu teils unlogischen Ergebnissen führen. Die vielen Nullmessungen und teilweise großen Unterschiede zwischen den Innen- und Außenmessungen führten auch zu numerischen Problemen, weshalb einige Modelle aufgrund von Divergenz des Algorithmus nicht bewertbar waren.

Literatur

Aitkin, M., Francis, B. und Hinde, J. (2009). *Statistical Modelling in R*. Oxford University Press: New York.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov und F. Czaki (Hrsg.), *Proceedings of the Second International Symposium on Information Theory* (S. 267–281).

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, *12*, 162–176.

Box, G. E. P. und Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, *33*, 33–44.

Cameron, A. und Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, *46*, 347–364.

Casella, G. und Berger, R. (2001). *Statistical Inference*. Duxbury Press: Pacific Grove.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer: New York.

Crawley, M. (2007). *The R Book*. Wiley: Chichester.

Fahrmeir, L. und Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, *13*, 342–368.

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*. Springer: Heidelberg.

Fisher, R. A. (1935). The case of zero survivors (Appendix to Bliss, C.I. (1935) 'The calculation of the dosage-mortality curve'). *Annals of Applied Biology*, *22*, 134–167.

Fisher, R. A. (1997). On an absolute criterion for fitting frequency curves. *Statistical Science*, *12*, 39–41.

Galler, H. (2011). *Myzelbildende Pilze und Anisole in der Luft von Weinkellern*. Dissertation, Medizinische Universität Graz.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*, 149–192.

Haas, D., Galler, H., Luxner, J., Zarfel, G., Buzina, W., Friedl, H., ... Reinthaler, F. F. (2013). The concentration of culturable microorganisms in relation to particular matter in urban air. *Atmospheric Environment*, *65*, 215–222.

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press: New York.

Hurvich, C. M. und Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Hurvich, C. M. und Tsai, C. L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, *51*, 1077–1084.

Ismail, N. und Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. *Casualty Actuarial Society Forum*, 103–158.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *49*, 127–162.

Jørgensen, B. (1992). *The theory of exponential dispersion models and analysis of deviance*. Conselho Nacional de Desenvolvimento Científico e Tecnológico, Instituto de Matemática Pura e Aplicada.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, *15*, 209–225.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, *11*, 59–67.

McCullagh, P. und Nelder, J. A. (1989). *Generalized Linear Models (Second edition)*. London: Chapman & Hall.

Millar, R. B. (2011). *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. Wiley: New Jersey.

Myers, R., Montgomery, D., Vining, G. und Robinson, T. (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences*. Wiley: New Jersey.

Nelder, J. A. und Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood : some comparisons. *Journal of the Royal Statistical Society. Series B (Methodological)*, *54*, 273–284.

Nelder, J. A., Lee, Y., Bergman, B., Hynén, A., Huele, A. F. und Engel, J. (1998). Joint modeling of mean and dispersion. *Technometrics*, *40*, 168-175.

Nelder, J. A. und Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*, 221–232.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*, 370–384.

Seber, G. und Lee, A. (2003). *Linear Regression Analysis*. Wiley: New Jersey.

Stigler, S. M. (1981). Gauss and the invention of least squares. *Annals of Statistics*, *9*, 465–474.

Thaler, T. (2009). *Die Extended-Quasi-Likelihood-Funktion in Generalisierten Linearen Modellen.* Masterarbeit, Technische Universität Graz.

Venables, W. und Ripley, B. (2002). *Modern Applied Statistics with S.* Springer: New York.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439–447.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer: New York.

Winkelmann, R. (2008). *Econometric Analysis of Count Data.* Springer: Berlin Heidelberg.

Winkelmann, R. und Zimmermann, K. F. (1991). A new approach for modeling economic count data. *Economics Letters*, *37*, 139–143.