TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

D I P L O M A R B E I T

Absatzschätzung anhand Generalisierter Linearer Modelle

Ausgeführt am Institut für

Wirtschaftsmathematik
der Technischen Universität Wien

unter der Anleitung von

Ao.Univ.Prof. Scherrer Wolfgang

durch

Martin Götzinger
Klosterneuburgergasse 20, 3400 Klosterneuburg

_____ _____
Datum Unterschrift (Student)

_____
Unterschrift (Betreuer)

Einleitung

Für jedes Handelsunternehmen stellt sich die Frage nach einer optimalen Warenzuteilung um Kosten durch Warenüberschuss oder Lagerengpässe zu minimieren. Die vorliegende Diplomarbeit soll zu diesem Zweck Methoden zur Schätzung der Nachfrage anhand historischer Verkaufs- und Lagerdaten liefern.

Bei der Modellierung der Nachfrage kommen Generalisierte Lineare Modelle zur Anwendung, welche eine universell einsetzbare Erweiterung der klassischen linearen Modelle darstellen. Das Konzept dieser Modelle wird in einem theoretischem Teil behandelt und später anhand realer Daten angewandt.

Für den praktischen Teil dieser Arbeit werden Originaldatensätze eines österreichischen Sporthandelskonzerns verwendet. Die benutzten Daten beschränken sich auf Trainingsanzüge, der betrachtete Zeitraum beträgt 18 Monate. Die Methoden zur Modellierung des Absatzes können trotz der Beschränkung auf Trainingsanzüge auch für andere Warengruppen adaptiert werden.

Im ersten Kapitel werden die theoretischen Grundlagen der angewandten Generalisierten Linearen Modelle behandelt. Das zweite Kapitel erklärt Kriterien anhand derer man die Qualität von geschätzten Modellen beurteilen kann. Diese und weitere Kriterien werden im vierten Kapitel zum Vergleich von konkreten Modellen angewandt.

Kapitel 3 widmet sich der Analyse und Aufbereitung der zur Modellschätzung benutzten empirischen Daten.

Im vierten Kapitel werden verschiedene Modelle zur Schätzung der Nachfrage ausgewertet und verglichen. Ein Schwerpunkt liegt dabei in der Interpretation der geschätzten Modellparameter.

Im letzten Kapitel wird die Nachfrage nach bestimmten Größen von Trainingsanzügen behandelt. Dabei werden Größenläufe, die aus unterschiedlichen Modellen resultieren, gegenübergestellt.

Sämtliche statistischen Analysen und Modellschätzungen wurden mit der Statistik-Software R umgesetzt.

Inhaltsverzeichnis

Kapitel 1

Generalisierte Lineare Modelle (GLM)

1.1 Einleitung

Generalisierte lineare Modelle bilden eine Verallgemeinerung der klassischen linearen Modelle.

Im klassischen linearen Modell ist die Zielvariable normalverteilt mit Erwartungswert $\boldsymbol{x}'\boldsymbol{\beta}$ und konstanter Varianz σ^2. Einen universeller einsetzbaren Ansatz bieten generalisierte lineare Modelle. Die Zielvariable muss nicht bedingt normalverteilt sein. Hingegen liegt die Annahme zugrunde, dass die Zielvariable nach einer Verteilung aus der Klasse der Exponentialverteilungen verteilt ist.

Das verallgemeinerte lineare Modell besteht aus 3 Komponenten (vgl. Winkelmann [1994]):

1. linearer Prädiktor
2. Link Funktion
3. Verteilungsfunktion aus der Exponentialfamilie

1.1.1 Linearer Prädiktor

$$\eta = \boldsymbol{x}'\boldsymbol{\beta} \tag{1.1}$$

η wird als linearer Prädiktor bezeichnet. Im Gegensatz zum klassischen linearen Modell ist das jedoch im Allgemeinen noch nicht direkt der Erwartungswert $\mu = E(Y)$ sondern η hängt über die so genannte Link Funktion mit μ zusammen.

1.1.2 Link Funktion

$$E(Y) = \mu = m(\eta) = m(\boldsymbol{x}'\boldsymbol{\beta}) \tag{1.2}$$

m wird als Responsefunktion und $m^{-1}(\mu) = \eta$ als Link Funktion bezeichnet. Es können verschiedene Link Funktionen verwendet werden. Praktikabel sind Linkfunktionen, deren entsprechende Responsefunktion den selben Bildbereich hat wie der Erwartungswert μ der abhängigen Variable Y.

1.1.3 Verteilungsfunktion aus der Exponentialfamilie

Eine Zufallsvariable besitzt eine Verteilung aus der Familie der Exponentialverteilungen wenn sich ihre Dichte oder Wahrscheinlichkeitsfunktion in folgender Form darstellen lässt (vgl. Fahrmeir and Kneib [2009] oder Hilbe [2011]):

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right) \tag{1.3}$$

Der Parameter θ heißt natürlicher oder kanonischer Parameter. Für $b(\theta)$ muss gelten, dass erste und zweite Ableitung $b'(\theta)$ und $b''(\theta)$ existieren. Eine spezielle Linkfunktion m^{-1}, die kanonische Linkfunkion erhält man, wenn m so gewählt wird, dass $\theta = \eta \Leftrightarrow \mu = b'(\theta) = m(\eta)$.
Erwartungswert und Varianz der Exponentialfamilie erhält man durch: $E(y) = b'(\theta), Var(y) = b''(\theta)\phi$.
ϕ wird als Skalierungs- oder Dispersionsparameter bezeichnet (nicht zu verwechseln mit dem Dispersionsparameter α der als Faktor in der Varianzfunktion der Negativen Binomialverteilung vorkommt).

Beispiel. Normalverteilung
setzt man in 1.3
$\theta = \mu$
$\phi = \sigma^2$
$b(\theta) = \frac{\theta^2}{2}$

$c(y, \phi) = \frac{-y^2}{2\phi} - \log\left(\sqrt{2\pi\phi}\right)$
so erhält man die Dichte der Normalverteilung.

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

und erkennt, dass diese Element der Exponentialfamilie ist.
Erwartungswert und Varianz der Normalverteilung: $E(y) = b'(\theta) = \theta = \mu$, $V(y) = b''(\theta)\phi = \sigma^2$. Mit der Identität als Linkfunktion sieht man, dass die klassische lineare Regression ein Spezialfall der Generalisierten linearen Regression ist.

Eine weitere wichtige Verteilung aus der Exponentialfamilie ist die Poissonverteilung. Insbesondere ist sie eine der wichtigsten Verteilungen zur Modellierung von Zähldaten.

Beispiel. Poissonverteilung
Dass die Poissonverteilung tatsächlich aus der Exponentialfamilie stammt erkennt man in dem man in 1.3
$\theta = \log \mu$
$\phi = 1$
$b(\theta) = \exp(\theta)$
$c(y, \phi) = -\log y!$
setzt und erhält somit

$$f(y) = \frac{\mu^y}{y!} \exp(-\mu),$$

die Wahrscheinlichkeitsfunktion der Poissonverteilung.
Für Mittelwert und Varianz der Poissonverteilung folgt:
$E(y) = b'(\theta) = \exp(\theta) = \mu$
$V(y) = b''(\theta)\phi = \exp(\theta) = \mu.$

Für die Wahrscheinlichkeitsfunktion des Negativen Binomialmodells (siehe Kapitel 1.5) mit Varianzfunktion $\mu + \alpha\mu^2$ ist es etwas komplizierter zu zeigen.

Beispiel. Negatives Binomialmodell
Wahrscheinlichkeitsfunktion des Negativen Binomialmodells NB2:

$$f(y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y} \alpha \geq 0, y = 0, 1, 2, \ldots$$

oder umgeformt:

$$f(y|\mu,\alpha) = \frac{(y+\alpha^{-1})!}{y!(\alpha^{-1}-1)!}\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \alpha \geq 0, y = 0,1,2,\ldots$$

mit $p := \frac{1}{1+\alpha\mu}$ und $1-p = \frac{\alpha\mu}{1+\alpha\mu}$ folgt:
$f(y|\mu,\alpha) = \exp\left(y\log(1-p) + \alpha^{-1}\log p + \log\binom{y+\alpha^{-1}-1}{\alpha^{-1}-1}\right)$ und somit:
$\theta = \log(1-p) \Rightarrow p = 1-\exp\theta$
$b(\theta) = -\alpha^{-1}\log p \Rightarrow b(\theta) = -\alpha^{-1}\log(1-\exp\theta)$
$\phi = 1$
$c(y,\phi) = \log\binom{y+\alpha^{-1}-1}{\alpha^{-1}-1}$.
Somit ist gezeigt, dass die Form der Wahrscheinlichkeitsfunktion jener der Exponentialfamilie entspricht.
für Mittelwert und Varianz folgt:
$E(y) = b'(\theta) = \alpha^{-1}\frac{\exp(\theta)}{1-\exp(\theta)} = \mu$
$V(y) = b''(\theta)\phi = \alpha^{-1}\frac{\exp(\theta)}{(1-\exp(\theta))^2} = \mu(1+\alpha\mu)$

Siehe hierzu auch Hilbe [2011].

1.2 Maximum Likelihood Schätzung

Generalisierte Modelle können mit der Maximum Likelihood Methode geschätzt werden.

Definition 1. Likelihood Funktion (vgl.Cameron and Trivedi [1998])
Seien $Y_1 \ldots Y_n$ stochastisch unabhängige Zufallszahlen, deren Verteilung von einem Parametervektor p abhängt, dh. die gemeinsame Dichte von $Y_1 \ldots Y_n$ ist $f(y_1,\ldots,y_n;p) = \prod_{i=1}^{n} f_i(y_i;p)$. Betrachtet man nun die Realisierungen $(y_1,\ldots y_n)$ als fix und sieht den Parameter p als Variabel an, so ist die Dichte $f(y_1,\ldots,y_n;p)$ eine Funktion, die für die betrachteten Realisierungen nur vom Parameter p abhängt.
$L(p;y_1,\ldots,y_n) = \prod_{i=1}^{n} f_i(y_i;p)$ wird dann als Likelihood Funktion bezeichnet.

Der Maximum Likelihood Schätzer ist nun jener Parameter p beziehungsweise Vektor von Parametern, welcher die Likelihood Funktion maximiert.

Oft wird zur Berechnung der Logarithmus der Likelihood Funktion verwendet,

die sogenannte Log-Likelihood-Funktion:

$$l(p; y_1, \ldots, y_n) = \log L(p; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f_i(y_i; p)$$

Aufgrund der Monotonie des Logarithmus hat die Log-Likelihood-Funktion ihr Maximum an der selben Stelle wie die nicht logarithmierte Likelihood Funktion, ist aber oft leichter zu berechnen.

Für bestimmte Verteilungen ist der Maximum-Likelihood-Schätzer auch konsistent, wenn die Verteilung nur zum Teil korrekt spezifiziert ist. So ist er unter Annahme einer Verteilungen aus der Exponentialfamile konsistent wenn der Erwartungswert korrekt spezifiziert ist. Das gilt auch wenn die tatsächliche Verteilung der Daten nicht der Annahme entspricht. Korrekte Werte für die Standardabweichung des Maximum-Likelihood-Schätzers erhält man wenn neben dem Erwartungswert auch die Varianz korrekt spezifiziert ist, jedoch müssen auch hierfür die Daten nicht der angenommenen Verteilung entsprechen.

Für die Berechnung des Maximum-Likelihood-Schätzers gibt es nur für spezielle Verteilungen geschlossene Formeln. Im Allgemeinen müssen numerische Verfahren angewandt werden wie etwa der Fisher-Scoring Algorithmus (vgl. [Winkelmann, 1994]).

1.3 Maximum Likelihood für GLM

Wie im linearen Modell wird bei GLM versucht die beobachteten abhängigen, endogenen Variablen $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ durch eine Matrix unabhängiger, exogener Variablen

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_n} \end{pmatrix}$$

zu erklären. Die Spalten $1 \ldots k$ stehen für die k verschiedenen Merkmale einer Beobachtung $1 \ldots n$ stehen.

Es wird angenommen, dass die Y_i stochastisch unabhängig sind und die Verteilung von Y_i vom Parametervektor $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{x_i}, \boldsymbol{\Theta})$ abhängt. Der Parameter $\boldsymbol{p}$ hängt wiederum von den Werten der exogenen Variable $\boldsymbol{x_i}$ sowie von den zu schätzenden Parametern $\boldsymbol{\Theta}$ ab. Der Einfachkeit halber wird hier angenommen, dass $\boldsymbol{X}$ nicht stochastisch ist. Andernfalls betrachtet man die bedingte Verteilung von $\boldsymbol{Y}$ gegeben $\boldsymbol{X}$

Beim Poissonmodell bedeutet das, die Y_i sind nach Annahme poissonverteilt mit Parameter $p_i = \mu_i$, $Y_i \sim P(\mu_i)$, wobei $\mu_i = m(\boldsymbol{x_i}\boldsymbol{\beta})$ (m ist hier die Responsefunktion). In diesem Fall ist $\boldsymbol{\Theta} = \boldsymbol{\beta}$. Also $\boldsymbol{\Theta}$ besteht genau aus den Koeffizienten $\boldsymbol{\beta}$ des linearen Prädiktors.
Bei anderen Verteilungen kann $\boldsymbol{\Theta}$ noch andere Parameter, etwa zur Modellierung der Varianz enthalten. Die Likelihood-Funktion ist das Produkt der n Dichten oder Wahrscheinlichkeitsfunktionen für die n Beobachtungen. Der Maximum-Likelihood-Schätzer wird wieder durch Maximieren der Likelihoodfunktion bestimmt.

1.4 Poisson Regression

Eines der am häufigsten verwendeten Modelle zur Modellierung von Zähldaten ist das Poissonmodell. Das Poissonmodell hat den Vorteil, dass es nur einen Parameter μ pro Beobachtung hat und dieser dem Mittelwert und der Varianz entspricht. Gleichzeitig kann das ein großer Nachteil sein wenn die Varianz größer beziehungsweise kleiner als der Mittelwert ist. In diesen Fällen spricht man von Über- bzw. Unterdispersion. Die Poissonregression basiert auf der Poissonverteilung, die Daten müssen aber nicht notwendigerweise poissonverteilt sein um ein sinnvolles Modell schätzen zu können.

Wahrscheinlichkeitsfunktion der Poissonverteilung:

$$P(y) = \frac{\mu^y}{y!} e^{-\mu} \tag{1.4}$$

unter Verwendung des natürlichen Logarithmus als Linkfunktion ergibt sich folgender Ansatz für den Erwartungswert:

$$\mu_i = e^{(\boldsymbol{x_i}'\boldsymbol{\beta})} \tag{1.5}$$

Die Verwendung des Logarithmus als Linkfunktion stellt sicher, dass die geschätzten Erwartungswerte positiv sind.

Die (bedingte) Wahrscheinlichkeit, dass Y_i den Wert y_i annimmt ist unter der Annahme Y_i sei poissonverteilt mit $Y_i \sim P(\mu_i)$:

$$P(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{e^{y_i(\boldsymbol{x_i}'\boldsymbol{\beta})} e^{-e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}}{y_i!} \tag{1.6}$$

als Likelihood-Funktion ergibt sich somit

$$L(\boldsymbol{\beta}|\boldsymbol{x_1},\ldots,\boldsymbol{x_n},\boldsymbol{y}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x_i};\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{y_i(\boldsymbol{x_i}'\boldsymbol{\beta})}e^{-e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}}{y_i!} \tag{1.7}$$

und für die Log-Likelihood-Funktion

$$l(\boldsymbol{\beta}|\boldsymbol{x_1},\ldots,\boldsymbol{x_n},\boldsymbol{y}) = \sum_{i=1}^{n} \left(y_i\boldsymbol{x_i}'\boldsymbol{\beta} - e^{\boldsymbol{x_i}'\boldsymbol{\beta}} - \ln y_i!\right) \tag{1.8}$$

Das Maximum der Log-Likelihood-Funktion und somit der Maximum-Likelihood-Schätzer $\hat{\boldsymbol{\beta}}$ wird nun durch iterative Verfahren wie den Fischer Score Algorithmus bestimmt.

Der Poisson-Maximum-Likelihood-Schätzer ist konsistent wenn der bedingte Erwartungswert korrekt spezifiziert ist. Die Daten müssen nicht notwendigerweise Poisson-verteilt sein. Sollte die Varianz der Daten jedoch kleiner (Unterdispersion) beziehungsweise größer (Überdispersion) sein als der Erwartungswertungswert, so kann das zu fehlerhaften Ergebnissen bei der Berechnung der Standardabweichung der geschätzten Parameter führen. [Cameron and Trivedi, 1998].

1.5 Negatives Binomialmodell

Eine Möglichkeit mit Überdispersion umzugehen ist die Verwendung des Negativen Binomialmodell. Hier ist es möglich die Varianz flexibler zu definieren als beim Poissonmodells. Die Varianz wird in diesem Modell über eine Varianzfunktion $V(\mu)$ definiert. Eine der am häufigsten verwendeten Implementierung des Negativen Binomialmodells ist das NB2 Modell mit Erwartungswert $\mu = \exp(\boldsymbol{x^T\beta})$ und Varianz Funktion $\mu + \alpha\mu^2$. Der Parameter $\alpha \geq 0$ wird als Dispersionsparameter bezeichnet (vgl. [Cameron and Trivedi, 1998] sowie [Hilbe, 2011]).

Die Wahrscheinlichkeitsfunktion für das NB2 Modell:

$$f(y|\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \alpha \geq 0, y = 0,1,2,\ldots \tag{1.9}$$

$\Gamma(.)$ bezeichnet die Gammafunktion. Für $\alpha = 0$ erhält man die Wahrscheinlichkeits-

funktion der Poissonverteilung und μ als Varianz.
Definition 2. Die Gammafunktion $\Gamma(a)$ ist definiert als

$$\Gamma(a) = \int_{\infty}^{0} \exp^{-t} t^{a-1} dt, a > 0. \tag{1.10}$$

Die Log-Likelihood Funktion für das Negative Binomialmodell mit $\mu_i = \exp(\boldsymbol{x_i'}\boldsymbol{\beta})$ und Varianzfunktion $\mu_i + \alpha\mu_i^2$ lautet:

$$l(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \left(\sum_{j=0}^{y_i - 1} \log(j + \alpha^{-1}) \right) - \log y_i! - (y_i + \alpha^{-1}) \log(1 + \alpha \exp(\boldsymbol{x_i}'\boldsymbol{\beta})) + y_i \log \alpha + y_i \boldsymbol{x_i'}\boldsymbol{\beta} \right\} \tag{1.11}$$

In R gibt es in der Library MASS [Venables and Ripley, 2002] eine implementierte Funktion zur Schätzung von Negativen Binomialmodellen der NB2 Form.

Eine allgemeinere Klasse von Negativen Binomialmodellen bietet die Verwendung einer Varianzfunktion der Form $\mu_i + \alpha\mu_i^a$, $a \in \mathbb{N}$. Auf diese Verallgemeinerung wird in dieser Diplomarbeit jedoch nicht eingegangen (siehe Cameron and Trivedi [1998]).

Kapitel 2

Modellvergleiche

2.1 Devianz

Die Devianz (Residual Deviance) $D(y)$ ist definiert als [McCullagh and Nelder, 1989]

$$D(y) = -2(l(\hat{\theta}) - l(\hat{\theta_s})). \tag{2.1}$$

Die Devianz ist die Differenz zwischen der Maximum-Log-Likelihood-Funktion des geschätzten Modells und einem vollständigen Modell multipliziert mit -2. Als vollständig wird in diesem Zusammenhang ein Modell bezeichnet, das für jede Beobachtung einen eigenen Parameter verwendet und somit eine perfekte Anpassung an die Daten ermöglicht.

Betrachtet man die Differenz zwischen den Devianzen zweier Modellen, wenn eines der beiden ein Spezialfall des größeren Modells mit weniger Parametern ist, so fällt die Log-Likelihood-Funktion des vollständigen Modells weg und die Differenz entspricht der Teststatistik eines Likehood-Quotienten-Tests ist somit asymptotisch $\chi^2(m)$ verteilt. m ist hierbei die Anzahl an Parametern, die das größere Modell zusätzlich hat. Durch sequentielles Hinzufügen zusätzlicher Parameter und anschließender Berechnung der Devianzen kann so getestet werden ob Parameter zu einer statistisch signifikanten Verbesserung eines Modells führen. In R existiert die Funktion anova.glm, mit der genau das eben beschriebene Verfahren implementiert ist. Eine aussagekräftigere Analyse erhält man, indem von einem Modell ausgehend die Veränderung der Devianz betrachtet wird, die durch Weglassen jeweils eines einzelnen Parameters entsteht. Dieses Verfahren kann in R durch die Funktion drop1 durchgeführt werden.

Modelle bei denen nicht eines ein kleinerer Spezialfall des größeren Modells ist können im Allgemeinen nicht auf diese Weise verglichen werden.

2.2 Akaike Information Criterion AIC

Das AIC ist ein Maß für die Qualität eines Modells. Beim AIC wird der Wert der Log-Likelihood-Funktion betrachtet und zusätzlich noch ein Strafterm abhängig von der Anzahl der Parameter addiert. Je geringer der Wert des AIC, desto besser.
Definition 3. Der klassische AIC ist definiert als:

$$AIC = 2k - 2\log L \tag{2.2}$$

wobei k die Anzahl der geschätzten Parameter ist und $\log L$ die Likelihood-Funktion des geschätzten Modells. Alternativ kann der AIC auch modifiziert werden in dem der Faktor des Strafterms vergrößert wird um das hinzufügen zusätzlicher Parameter stärker zu bestrafen.

Ersetzt man den Faktor des Strafterms durch $\log n$ so erhält man das Bayessche Informationskriterium, bei dem auch die Anzahl der Beobachtungen berücksichtigt wird. Ist die Anzahl der Beobachtungen $n \geq 8$ so werden im Vergleich zum AIC zusätzliche Parameter stärker bestraft.

$$BIC = k\log n - 2\log L \tag{2.3}$$

Kapitel 3

Daten für praktische Anwendung

Nach der theoretischen Behandlung von Generalisierten Linearen Modellen werden diese nun anhand von Originaldatensätzen eines österreichischen Sporthandelsunternehmens angewandt. Die betrachteten Daten sind auf Trainingsanzüge beschränkt, die benutzten Methoden wären auch auf andere Warengruppen anwendbar. Anhand der Daten wird der Absatz mit verschiedenen Modellen geschätzt.

Ziel der Absatzschätzungen ist eine optimale Warenzuteilung um Kosten durch Warenüberschuss oder Lagerengpässe zu minimieren.

Zunächst werden die Daten aufbereitet, in Kapitel 4 werden unterschiedliche Modelle verglichen und bewertet. Im fünften Kapitel wird die Nachfrage nach bestimmten Größen von Trainingsanzügen behandelt.

3.1 Erklärung der Daten

Grundlage der Analysen ist eine Datei bestehend aus 81.072 Datensätzen die über einen Zeitraum von 18 Monaten Informationen wie Lagerstand, Verkaufspreis Verkaufsmenge etc. für jeden Artikel in jeder Filiale beinhalten. Das Zeitinterval der Datensätze beträgt 1 Monat. Die gelieferten Daten sind so konzipiert, dass ein Datensatz für einen Trainingsanzug und einen bestimmten Monat in einer bestimmten Filiale nur beinhaltet ist, wenn der Artikel in diesem Monat mindestens einmal verkauft wird, oder wenn eine neue Lieferung eintrifft. Für alle anderen Monate ist für den Artikel in der Filiale kein Datensatz vorhanden. Jeder Datensatz setzt sich aus folgenden Datenfeldern zusammen:

Jahr/Monat	Beobachteter Monat
Filiale	2 stellige Zahl die für eine bestimmte Filiale steht.
ARTNR	Artikelnummer. 5 bis 7 stellige Zahl.
ARTBEZ	Artikelbezeichnung
Marke	Marke
Farbe	Farbe
GR	Größe
VKMG	Verkaufsmenge des Trainingsanzuges
VK/STK	Verkaufspreis pro Trainingsanzug
Umsatz	Umsatz im beobachteten Monat. Umsatz = VKMG * VK/STK
PG	Preisgruppe. Die Artikel sind ensprechend ihrem Preis einer Preisgruppe zwische
LagerMg	Lagermenge des Trainingsanzuges am Ende des Monats
WZMg	Wahrenzugangsmenge für diesen Trainingsanzug in diesem Monat
UML Abgeber	Umlagerungen zu anderen Filialen
UML Empfänger	Umlagerungen von anderen Filialen

3.2 Ergänzung fehlender Datensätze

Für die weitere Arbeit mit den Daten müssen diese so modifiziert werden, dass für alle Artikel auch für jene Monate Daten zur Verfügung stehen an denen nichts verkauft wurde und sich auch nichts am Lagerstand verändert hat. Um das zu erreichen werden die Daten für die fehlenden Monate durch Vorwärtsrekursion ergänzt: Für jeden fehlenden Monat werden sämtliche Daten wie Verkaufspreis und Lagerstand vom Vormonat übernommen. Verkaufsmenge und Umsatz werden auf 0 gesetzt.

3.3 Verkaufspreise

In den Originaldaten hat ein Artikel in einem Zeitraum nur einen definierten Preis wenn er in diesem Monat auch mindestens ein Mal verkauft wurde. Wird ein Artikel in einem Zeitraum nicht verkauft, so ist nicht erkenntlich um welchen Preis er angeboten worden ist. Um dieses Problem zu beheben wurde folgendermaßen vorgegangen:

1. Suche Artikel in der selben Filiale zum selben Zeitpunkt in anderer Größe oder Farbe mit definiertem Preis.

2. wenn der Artikel auch in anderen Größen und Farben im gesuchten Zeitraum nicht verkauft wurde, wird ein Preis gewählt zu dem der Artikel zu einem früheren oder späteren Zeitpunkt verkauft wurde. Es wird hier jener Zeitpunkt gewählt, der am nächsten zum aktuell gesuchten Zeitraum liegt.

Das Ergänzen der Preise durch jene von anderen Größen und Farben sollte eine gute Näherung sein, da der Preis eines bestimmten Trainingsanzuges nahezu immer, unabhängig von der Wahl der Farbe und Größe, identisch ist. Eine etwas schlechtere Näherung ist die Ersetzung der Preise durch später realisierte Preise, da etwa früher stattgefundene Preissenkungen nicht berücksichtigt werden können.
Mit dieser Vorgangsweise bleiben von 24.579 Datensätzen mit nicht definierten Preisen 1.786 Datensätze übrig bei denen der Preis nicht definiert ist. Betroffen sind hier Modelle, die in bestimmten Filialen über den Zeitraum der ganzen 18 Monate kein einziges Mal verkauft wurden, dort jedoch irgendwann angeboten wurden. Insgesamt sind 43 Modelle von verschiedenen Marken betroffen.

n	ARTBEZ	Marke
1	Basic Woven Tracksui	Adidas
2	Clima Quarter Packag	Adidas
3	Hr. PES Trainer	Adidas
4	Hr. PES Trainer 3 S	Adidas
5	Hr. WEB Trainer SMU	Adidas
6	PES Suit	Adidas
7	TR Train kn	Adidas
8	TS Young	Adidas
9	Hr. Hoody Trainer	Benger
10	Hr. PES Trainer modi	Benger
11	Hr. Trainer Polytr.	Benger
12	Hr.Trainer Toronto	Erima
13	227651 Hr. Trainer N	Keine
14	329609 Hr. Trainer N	Keine
15	329610 Hr. Trainer N	Keine
16	Atlanta Hr. Trainer	Keine
17	Atlanta Trainer 1024	Keine
18	E16259 Hr. Trainer A	Keine
19	Hr. Trainer Chicago	Keine

```
20   K8520 Trainer Lotto   Keine
21   K8521 Trainer Lotto   Keine
22   K8523 Trainer Lotto   Keine
23    L0621Trainer Lotto   Keine
24   L0622 Trainer Lotto   Keine
25     AD Sprint Warm Up    Nike
26  Classic Poly Warm Up    Nike
28  Hr. PES Trainer Park    Nike
29           Hr. Trainer    Nike
30     Polyester Tafetta    Nike
31  Regional Classic War    Nike
32  Striker Polywarp War    Nike
33           Trainer Hr.    Nike
34         BTS Poly Suit    Puma
35       Esito Poly Suit    Puma
36       Hr. Web Trainer    Puma
37         Poly Suit BTS    Puma
38   Hr. PES Trainer SMU  Reebok
39           Hr. Trainer  Reebok
40  Hr. Web Trainer Core  Reebok
41  Tricot graphic Tracs  Reebok
42       Hr. PES Trainer   Umbro
43   Nacionale Polyanzug   Umbro
```

Genauere Analyse zeigt, dass von diesen 43 Modellen 15 Modelle nur in einem Monat angeboten wurden (Tabelle 3.1). Das können einerseits Modelle sein, die nur zu Beginn des Beobachtungszeitraumes angeboten wurden, andererseits auch Artikel, die erst am Ende des Beobachtungszeitraumes ins Sortiment aufgenommen wurden. Aus den Daten ist nur ersichtlich ob ein Artikel irgendwann in einem Monat angeboten wurde, nicht wie lange er in diesem Monat angeboten wurde. So kann es etwa sein, dass Artikel erst am letzen Tag im letzten Monat, für den wir Daten haben, zum ersten Mal angeboten wird. Das bietet eine Erklärung, warum diese 15 Modelle in manchen Filialen gar nicht verkauft wurden.
Es gibt weitere 4 Modelle, die in nur in 2 Monaten angeboten wurden (Tabelle 3.2).

Es bleiben 26 Modelle übrig die über einen längeren Zeitraum angeboten aber in manchen Filialen nie verkauft wurden.

n	ARTBEZ	Marke
1	Basic Woven Tracksui	Adidas
2	Hr. PES Trainer 3 S	Adidas
3	PES Suit	Adidas
4	TR Train kn	Adidas
5	TS Young	Adidas
7	Hr.Trainer Toronto	Erima
8	Classic Poly Warm Up	Nike
9	Clio Polywarp Entry	Nike
10	Hr. PES Trainer Park	Nike
11	Hr. Trainer	Nike
12	Hr. Web Trainer	Puma
13	Hr. Trainer	Reebok
14	Hr. Web Trainer Core	Reebok
15	Nacionale Polyanzug	Umbro

Tabelle 3.1: nur in einem Monat angebotene Artikel

n	ARTBEZ	Marke
1	Hr. Trainer Polytr.	Benger
2	Regional Classic War	Nike
3	BTS Poly Suit	Puma
4	Poly Suit BTS	Puma

Tabelle 3.2: nur in zwei Monaten angebotene Artikel

Die Datensätze für die insgesamt 43 Artikel werden für die entsprechenden Filialen wo sie nie verkauft wurden für die Absatzschätzung gelöscht. Des weiteren werden Datensätze für Artikel gelöscht, die nur sehr selten verkauft oder angeboten wurden.

Durch die Ergänzung der Historie erhöhte sich die Anzahl der Datensätze von 81.072 auf 166.700. Durch das Löschen von selten oder gar nicht verkauften Artikeln reduziert sich die Gesamtzahl an Datensätzen wieder auf 72.001. Die Gesamtverkaufszahl hat sich von ursprünglich 84.936 Artikeln auf 74.708 reduziert, der Gesamtumsatz von 3.987.406 Euro auf 3.626.729 Euro.

Zusätzlich zum Preis wird noch eine Variable Preisgruppe PG geliefert, die die Artikel nach ihrem Verkaufspreis in folgende Gruppen einteilt:

- Preisgruppe 1: VP $>$ 120 EUR
- Preisgruppe 2: 100 EUR $<$ VP $\leq$ 120 EUR
- Preisgruppe 3: 80 EUR $<$ VP $\leq$ 100 EUR
- Preisgruppe 4: 65 EUR $<$ VP $\leq$ 80 EUR
- Preisgruppe 5: 45 EUR $<$ VP $\leq$ 65 EUR
- Preisgruppe 6: 0 EUR $<$ VP $\leq$ 45 EUR
- Preisgruppe 0: Preis undefiniert

Diese Gruppierung der Daten hat sich für die Schätzung des Absatzes als sehr nützlich herausgestellt. Nach dem Ergänzen der Preise müssen die Preisgruppen nun noch so angepasst werden, dass jenen Artikeln, die zuvor undefinierten Preis hatten (Preisgruppe 0) korrekt in eine der Preisgruppen 1 bis 6 zugeteilt wird. Es wird sich später zeigen, dass die Verwendung der Preisgruppen eine bessere Schätzung des Absatzes ermöglicht als bei Anwendung der kontinuierlichen Variable Verkaufspreis.
Abbildung 3.1 zeigt die Einteilung nach dem Verkaufspreises in diese Preisgruppen. Preisgruppe 2 fehlt in dieser Abbildung, da im Beobachtungszeitraum keine Artikel dieser Preisgruppe angeboten wurden.
Abbildung 3.2 zeigt die Anzahl an Datensätzen pro Preisgruppe,
Abbildung 3.3 die Gesamtverkaufsmenge pro Preisgruppe und
Abbildung 3.4 den Gesamtumsatz pro Preisgruppe.
Abbildung 3.5 zeigt eine Kennzahl für das Angebot in den verschiedenen Preisgruppen: Für jede Preisgruppe wird berechnet wie viele unterschiedliche Artikel pro Filiale und Monat angeboten werden. Nach Aufsummieren über alle Monate und Filialen sowie

anschließender Normierung (Summe über alle Preisgruppen gleich 1) erhält man das Angebot in Abbildung 3.5.

Bei der Betrachtung der Verkaufsmengen und Umsätze wird klar, dass eigentlich nur Artikel in den Preisgruppen 4,5 und 6 relevant sind. In Preisgruppe 2 werden keine Artikel angeboten, daher werden für die Modelle auch keine entsprechenden Koeffizienten geschätzt.

Abbildung 3.1: Boxplot: Artikel in Preisgruppen

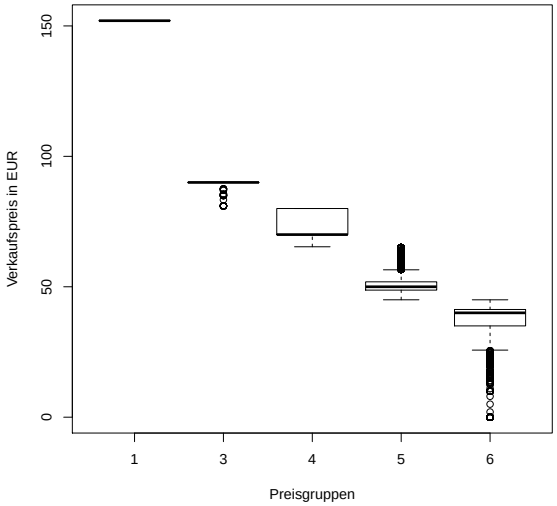


Abbildung 3.2: Datensätze pro Preisgruppe

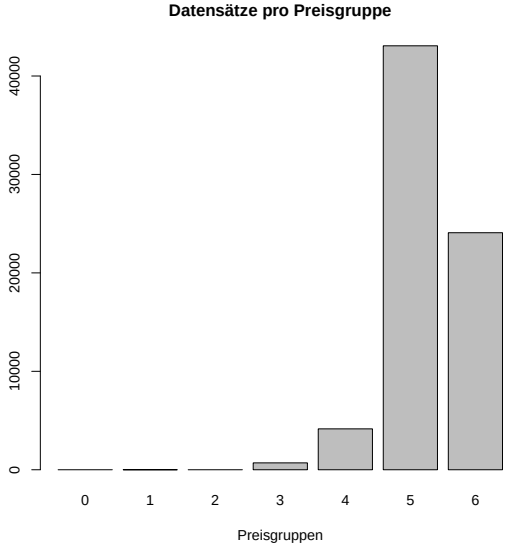


Abbildung 3.3: Verkaufsmenge pro Preisgruppe

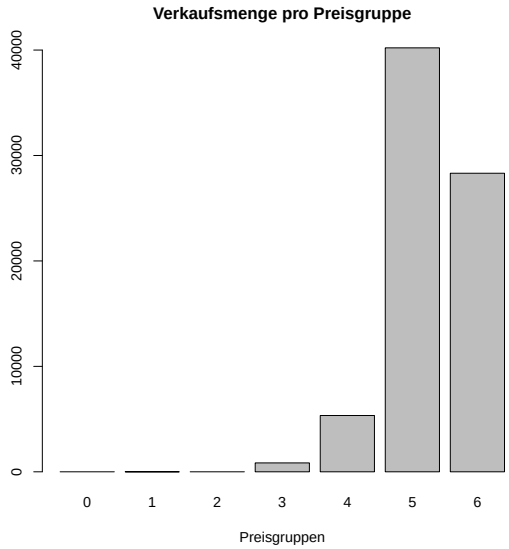


Abbildung 3.4: Umsatz pro Preisgruppe

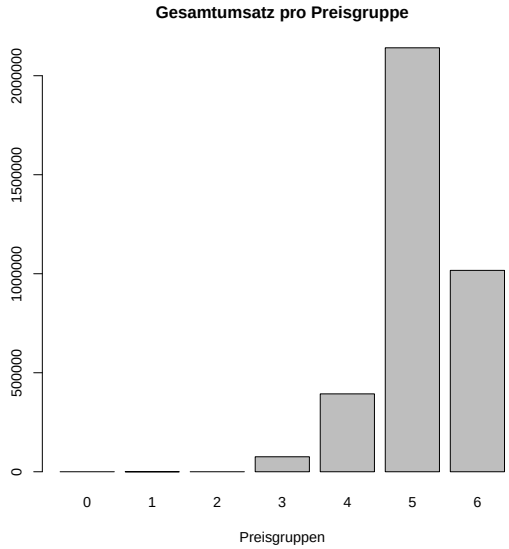
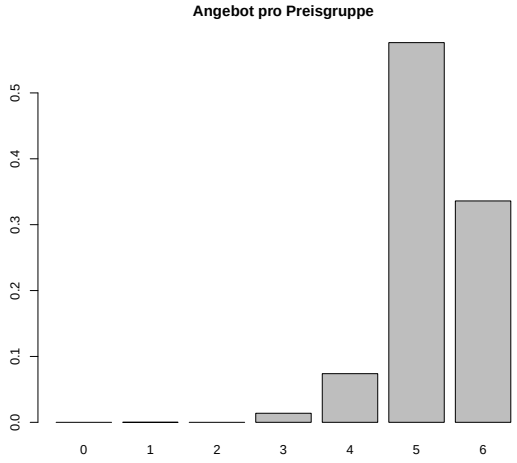


Abbildung 3.5: Angebot pro Preisgruppe



Kapitel 4

Schätzung des Absatzes

4.1 Poisson Modell

Zur Schätzung des Absatzes werden die bereinigten Daten verwendet. Des weiteren werden Datensätze ausgenommen, bei denen der Lagerstand am Ende des Monats auf 0 ist, da bei den betroffenen Artikeln davon auszugehen ist, dass die Nachfrage aufgrund des zu knappen Angebotes unterschätzt werden würde. Später wird sich auch zeigen, dass es großen Einfluss auf das Modell hat, was man als Mindestlagerstand für die Daten vorgibt.

Zunächst wird folgendes einfaches Poissonmodell betrachtet:

```
VKMG ~ 1 + VP + GR + Filiale + Marke
```

Die Verkaufsmenge (VKMG) wird durch die Variablen Verkaufspreis (VP), Größe (GR), Filiale und Marke erklärt. Später werden die erklärenden Variablen schrittweise angepasst und auch andere Ansätze wie das Negative Binomialmodell zur Schätzung der Verkaufsmenge verwendet. In diesem Modell ist die einzige kontinuierlich Variable der Verkaufspreis VP. Die Variablen GR, Filiale und Marke sind Faktorvariablen. Faktorvariablen sind Variablen, die nicht als numerische Werte interpretiert werden. Statt dessen wird für jeden möglichen Zustand der Variable ein eigener Koeffizient geschätzt und dieser mit 1 multipliziert wenn der entsprechende Zustand auftritt, sonst mit 0 multipliziert. Realisiert wird das durch Dummy Variablen die nur den Wert 1 oder 0 annehmen.

In R können solche generalisierten linearen Modelle mit der Funktion `glm` geschätzt werden. Im Folgenden ist der Output aus R mit den geschätzten Koeffizienten zu sehen. Bei der Interpretation der Koeffizienten ist zu beachten, dass das Intercept den Koeffizienten für eine Referenzkategorie (hier: GR4-Filiale2-MarkeAdidas) darstellt und alle anderen zu Faktorvariablen gehörenden Koeffizienten die Differenz zum Intercept darstellen. Die Wahl der Referenzkategorie hat keine Auswirkungen auf das Modell.

```
Call:  glm(formula = VKMG ~ 1 + VP + GR + Filiale + Marke,
    family = poisson(), data = data)

Coefficients:
(Intercept)           VP          GR5          GR6          GR7
  -0.112758    -0.001255     0.573590     0.769686     0.632677
        GR8          GR9         GRXS          GRS          GRM
   0.075768    -0.780117    -1.719891     0.191258     0.887991
        GRL         GRXL        GRXXL     Filiale3     Filiale4
   0.901797     0.329329    -0.531937    -0.885453     0.309307
   Filiale5     Filiale6     Filiale7     Filiale8     Filiale9
  -0.243435    -0.466358     0.159489    -0.747963    -0.429967
  Filiale10    Filiale11    Filiale12    Filiale13    Filiale14
  -0.741295    -0.097141    -0.785291    -0.702720    -0.547291
  Filiale15    Filiale16    Filiale17    Filiale18    Filiale19
  -1.059510    -0.247155    -0.286346    -0.356948    -1.016901
  Filiale20    Filiale21    Filiale22    Filiale23    Filiale24
  -0.614998     0.057826    -0.481281    -0.436597    -0.682869
  Filiale25    Filiale26    Filiale27    Filiale28    Filiale29
  -1.171990    -0.538599    -0.237492     0.059893    -0.450297
  Filiale30    Filiale31    Filiale32    Filiale34    Filiale35
  -0.759939    -0.926411    -0.849905    -0.281905    -0.936149
  Filiale36    Filiale37    Filiale38    Filiale39    Filiale40
  -0.832426    -1.722580    -0.429039    -0.840863    -0.592716
  Filiale41    Filiale42    Filiale43    Filiale44    Filiale45
  -0.488182    -0.797997    -0.749259    -0.761386    -1.044596
  Filiale46    Filiale47    Filiale48    Filiale49    Filiale50
  -0.911886    -0.870902    -0.600016    -0.106424    -0.599628
  Filiale51    Filiale52    Filiale53    Filiale54    Filiale55
```

```
  -0.328234    -0.949521    -0.679050    -0.891379    -0.208523
  Filiale56    Filiale57    Filiale58    Filiale59    Filiale60
  -0.903785    -0.222732    -0.633486    -0.525161    -0.433699
  Filiale61    Filiale62    Filiale63    Filiale64    Filiale65
  -0.838337    -1.938849    -0.529639    -1.063692    -0.868880
  Filiale66    Filiale67    Filiale68    Filiale69    Filiale71
  -0.495926    -0.608296    -0.542949    -0.782201    -0.773504
  Filiale72    Filiale73    Filiale74  MarkeBenger   MarkeErima
  -1.448061    -0.502671    -0.301238    -0.541915    -6.131206
  MarkeNike    MarkePuma  MarkeReebok   MarkeUmbro
  -0.616722    -0.021115    -0.113674    -0.316925

Degrees of Freedom: 47680 Total (i.e. Null);  47592 Residual
Null Deviance:      63030
Residual Deviance: 50890        AIC: 107400
```

Der Output enthält auch bereits Informationen über die Qualität des Models, die jedoch erst beim Vergleich mit anderen Modellen aussagekräftig sind. Die Residual Deviance $D(y)$ ist wie in Kapitel 2.1 definiert als [McCullagh and Nelder, 1989]

$$D(y) = -2(l(\hat{\theta}) - l(\hat{\theta_s})). \tag{4.1}$$

Die Devianz ist Differenz zwischen der Maximum-Log-Likelihood-Funktion des geschätzten Modells und einem vollständigen Modell multipliziert mit -2. Das vollständige Modell hat für jede Beobachtung einen Koeffizienten beziehungsweise so viele Koeffizienten, um die Daten perfekt zu erklären.
Die Null Deviance, die im R Output auch angegeben ist, ist die Devianz jenes Modells, das nur aus dem Intercept als Koeffizienten besteht. Des weiteren wird der AIC Wert des Modells ausgegeben.

4.2 Anpassung des Modells

Nach der Definition des ersten Poissonmodells wird nun versucht das Modell durch Betrachtung der Devianzen schrittweise zu verbessern, indem erklärende Variablen hinzugefügt oder entfernt werden.

Mit der Funktion *anova.glm* erhält man analog zu der Varianz-Analysetabelle bei linearen Modellen eine Devianz-Analysetabelle. Hier wird die Devianz statt der Sum-of-Squares ausgegeben. *Anova* liefert für ein GLM Modell zunächst eine Aussage, wie stark sich die Devianz verringert wenn sequentiell Parameter zum Modell hinzugefügt werden. Gestartet wird beim NULL-Modell, das nur ein Intercept als einzigen Koeffizienten hat. Die Reihenfolge in der die Parameter hinzugefügt werden entspricht der Reihenfolge in der die Parameter bei der Schätzung des Modells angegeben werden. Zusätzlich kann ein statistischer Test angegeben werden mit dem die hinzugefügten Parameter auf Relevanz getestet werden. Hierfür bietet sich für GLMs ein Chiquadrat Test an, da die Differenz der Devianzen asymptotisch χ^2 verteilt ist:

```
Analysis of Deviance Table
Model: poisson, link: log
Response: VKMG

Terms added sequentially (first to last)

        Df  Deviance Resid. Df  Resid. Dev  Pr(>Chi)
NULL                    55014      73875
VP       1    162.9     55013      73712 < 2.2e-16 ***
Marke    6    841.1     55007      72871 < 2.2e-16 ***
GR      12   2361.9     54995      70509 < 2.2e-16 ***
Filiale 70   5118.8     54925      65390 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Eine geschicktere Anwendung der Devianzanalyse bietet die Funktion *drop*1. Statt die Parameter sequentiell hinzuzufügen wird vom Modell, das alle Parameter beinhaltet ausgegangen und jeweils ein Parameter entfernt und die dadurch entstandene Verringerung der Devianz beurteilt.

```
Single term deletions

Model:
VKMG ~ 1 + VP + GR + Filiale + Marke
        Df Deviance    AIC    LRT Pr(>Chi)
<none>        50893 107365
```

```
VP        1    50898 107368    5.0  0.02587 *
GR       11    56412 112862 5519.7  < 2e-16 ***
Filiale 70    55700 112032 4807.6  < 2e-16 ***
Marke     6    53315 109775 2422.0  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Hier zeigt sich bereits ein großer Unterschied zu dem vorigen Verfahren. Die Variable VP scheint das Modell nicht signifikant zu verbessern. Der p-Wert des Chiquadrat Tests würde auch eher dafür sprechen, dass die Variable VP keinen großen Einfluss hat.

Eine Grund dafür könnte die Form sein in der die Abhängigkeit der Nachfrage vom Verkaufspreis modelliert wird. Mit dem bisherigen Ansatz beschreibt folgende Funktion die Abhängigkeit der Nachfrage von Preisänderungen:

$\delta = \exp(\alpha i \mathrm{VP}) - 1$
$\delta \ldots$ relative Änderung der Nachfrage
$i \ldots$ relative Änderung des Preises
$\alpha \ldots$ Koeffizient für die Variable Verkaufspreis VP
VP $\ldots$ ursprünglicher Preis

Eine Verbesserung könnte eine Adaptierung des Modells bringen in dem der logarithmierte Verkaufspreis $\log(VP)$ als Variable verwendet wird. In dieser Form wird die relative Änderung der Nachfrage auf eine Preisänderung i wie folgt modelliert:

$\delta = (1 + i)^{\alpha} - 1$
$\alpha \ldots$ Koeffizient für die Variable Größe log(VP)

Abbildung 4.1 zeigt die Preiselastizitäten die aus den beiden unterschiedlichen Modellen resultieren. Nach dem ursprünglichen Modell hängt die Preiselastizität vom absoluten Preis ab (in Abbildung 4.1 farblich dargestellt: von 10 Euro (rot) bis 70 Euro (gelb), in 10 Euro Schritten). Je höher der ursprüngliche Preis desto stärker sinkt die Nachfrage bei einer relativen Preisänderung.
Aus dem adaptierten Modell resultiert eine Preiselastizität, die vom absoluten Preis unabhängig ist.

Für den Ansatz `VKMG ~ 1+ log(VP) + GR + Filiale + Marke` erhält man folgendes Modell (R Output verkürzt dargestellt):

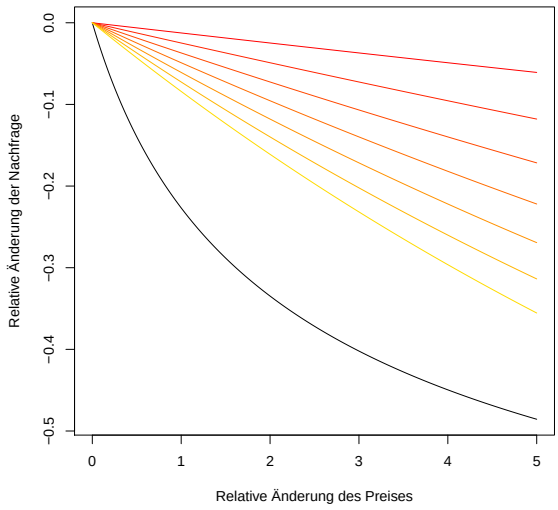


Abbildung 4.1: Preiselastizität unter Verwendung von VP (schwarz) sowie log(VP) (rot:VP=10, gelb:VP=70)

```
Call:  glm(formula = VKMG ~ 1 + log(VP) + GR + Filiale + Marke,
    family = poisson(), data = data)

Coefficients:
(Intercept)      log(VP)          GR5          GR6          GR7
    1.28312     -0.37174      0.58283      0.78370      0.64636
        GR8          GR9         GRXS          GRS          GRM
    0.07623     -0.79056     -1.75072      0.20294      0.89948
        GRL          GRXL         GRXXL     Filiale3     Filiale4
    0.91325      0.34171     -0.52529     -0.90043      0.28321
   Filiale5     Filiale6     Filiale7     Filiale8     Filiale9
      .            .            .            .            .
      .            .            .            .            .
      .            .            .            .            .

Degrees of Freedom: 47680 Total (i.e. Null);  47592 Residual
Null Deviance:      63030
Residual Deviance: 50730        AIC: 107200
```

Der Koeffizient für den Verkaufspreis hat sich nun im Vergleich zum vorherigen Modell in absoluten Werten deutlich von -0.012045 auf -0.37174 erhöht. Die Devianz ist von

50890 auf 50730 gesunken und auch der Wert des AIC hat sich verringert.

Erwartungsgemäß sieht nun auch die Devianzanlyse mittels der Funktion *drop*1 anders aus:

```
Single term deletions

Model:
VKMG ~ 1 + log(VP) + GR + Filiale + Marke
        Df Deviance    AIC    LRT  Pr(>Chi)
<none>        50725 107197
log(VP)  1    50898 107368  172.7 < 2.2e-16 ***
GR      11    56321 112771 5596.3 < 2.2e-16 ***
Filiale 70    55523 111855 4797.8 < 2.2e-16 ***
Marke    6    53407 109866 2681.4 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Der Verkaufspreis scheint zwar nach wie vor die unwichtigste erklärende Variable zu sein, hat aber durch die Bildung des Logarithmus an Relevanz zugenommen.

Eine zusätzliche Möglichkeit den Verkaufspreis einfließen zu lassen bietet die Definition von Preisgruppen um den Preis als Faktorvariable statt als kontinuierliche Größe zu verwenden. Hierfür wird die angepasste Preisgruppeneinteilung verwendet (siehe Kapitel 3.2).

Durch diese Modifizierung erhält man folgendes Modell:

```
Call:  glm(formula = VKMG ~ 1 + PG + GR + Filiale + Marke,
    family = poisson(), data = data)

Coefficients:
(Intercept)          PG3          PG4          PG5          PG6
  -10.69724     10.65207     10.71894     10.43034     10.85401
        GR5          GR6          GR7          GR8          GR9
    0.56399      0.76000      0.61240      0.06794     -0.78635
       GRXS          GRS          GRM          GRL         GRXL
   -1.85276      0.21022      0.88473      0.89883      0.34500
      GRXXL     Filiale3     Filiale4     Filiale5     Filiale6
```

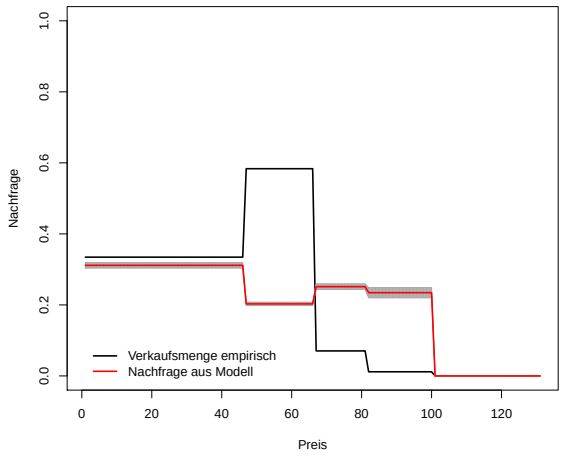


Abbildung 4.2: Abhängigkeit der Nachfrage vom Verkaufspreis bei Schätzung mit Preisgruppen. Mit 95% KI.

```
   -0.49770     -0.92788      0.22414     -0.23645     -0.43374
   Filiale7     Filiale8     Filiale9    Filiale10    Filiale11
    0.10030     -0.76797     -0.40788     -0.85377     -0.18557
          .            .            .            .            .
          .            .            .            .            .
          .            .            .            .            .

Degrees of Freedom: 47680 Total (i.e. Null);  47589 Residual
Null Deviance:      63030
Residual Deviance: 49750        AIC: 106200
```

Abbildung 4.2 zeigt die Abhängigkeit der Nachfrage vom Verkaufspreis bei Verwendung der Preisgruppen. Die Grafik ist so normiert, dass jeweils die Summe über alle Preisgruppen 1 ergibt. Die Berechnung des Konfidenzintervalls, das in der Grafik eingezeichnet ist, wird später in Kapitel 5.3 erklärt.
Die Diskrepanz zwischen der empirischen und geschätzten Preisabhängigkeit lässt sich dadurch erklären, dass die Preisgruppen 3 und 4 nur wenig angeboten werden, im Verhältnis zum Angebot jedoch recht gut verkauft werden. Aus dem Modell resultiert daher eine relativ hohe Nachfrage.

Man sieht eine Verbesserung der Devianz auf 49750 und eine Verringerung des AIC auf 106200.

Die Veränderung der Devianz des Modells durch Weglassen einzelner Variablen zeigt

wieder die Funktion *drop*1:

```
Single term deletions

Model:
VKMG ~ 1 + PG + GR + Filiale + Marke
        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       49750 106228
PG       4    50898 107368 1147.8 < 2.2e-16 ***
GR      11    54985 111441 5235.5 < 2.2e-16 ***
Filiale 70    54384 110722 4634.5 < 2.2e-16 ***
Marke    6    52849 109315 3099.0 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Nach wie vor ist der Verkaufspreis die unwichtigste Variable. Die Verwendung der Preisgruppen scheint aber die erfolgversprechenste Form zu sein den Verkaufspreis einfließen zu lassen.

Das Modell hat aufgrund der vielen Faktorvariablen sehr viele Koeffizienten. Besonders die Variable *Filiale* trägt mit ihren 71 verschiedenen Levels viel dazu bei. Daher scheint es sinnvoll zu untersuchen ob man diese nicht durch andere, kontinuierliche Größen ersetzen kann um das Modell zu vereinfachen ohne die Qualität des Modells gravierend zu verschlechtern. Neben den Daten über die Verkaufsmengen von einzelnen Trainingsanzügen existiert noch auch eine Datei, die Gesamtumsätze für Sportbekleidung aufgeschlüsselt nach Filialen und Zeitraum auflistet. Als kontinuierliche Größe würde sich daher der Gesamtumsatz der einzelnen Filialen anbieten um die Faktorvariable *Filiale* zu ersetzen.

Die Variable $FilialeUmsatz$ beinhaltet den Gesamtumsatz der Filialen in der Warengruppe *Sportbekleidung*. Da der Logarithmus als Linkfunktion im Modell verwendet wird, kommt man wieder zu einem besseren Ergebnis wenn der Logarithmus des Umsatzes verwendet wird.

Mit dieser Vereinfachung wird das Modell um 69 Koeffizienten verringert:

```
Call:  glm(formula = VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke,
    family = poisson(), data = data)
```

```
Coefficients:
       (Intercept)                 PG3                 PG4
         -18.86176            10.57193            10.63933
               PG5                 PG6                 GR5
          10.35501            10.76455             0.56461
               GR6                 GR7                 GR8
           0.76050             0.61266             0.06533
               GR9                 GRXS                GRS
          -0.78659            -1.86593             0.20711
               GRM                 GRL                 GRXL
           0.88228             0.89381             0.33802
             GRXXL  log(FilialeUmsatz)        MarkeBenger
          -0.50421             0.65152            -0.71508
       MarkeErima           MarkeNike           MarkePuma
          -6.05600            -0.63724            -0.01803
      MarkeReebok          MarkeUmbro
          -0.19656            -0.50765

Degrees of Freedom: 47680 Total (i.e. Null);  47658 Residual
Null Deviance:      63030
Residual Deviance: 50110        AIC: 106500
```

Die Devianz des Modells hat sich dadurch von 49750 auf 50110 erhöht und gleichzeitig der AIC von 16200 auf 16500. Das Modell wurde somit stark vereinfacht auf Kosten einer leichten Verschlechterung des Modells.

Bei den bisherigen Modellen wurde der Verkaufszeitpunkt noch überhaupt nicht in Betracht gezogen. Da zu erwarten ist, dass die Verkaufsmenge von Trainingsanzügen nicht über das ganze Jahr konstant ist könnte die Einbeziehung des Verkaufsmonats auch eine Verbesserung des Modells bringen.

Abbildung 4.3 zeigt die Gesamtverkaufsmenge von Trainingsanzügen pro Monat über den ganzen Beobachtungszeitraum von 18 Monaten. Trotz des relativ kurzen Beobachtungszeitraumes ist klar erkenntlich, dass es von Frühjahr bis Sommer einen Einbruch bei den Verkäufen gibt.

Nun wird der Angebotszeitpunkt in das Modell übernommen und die Variable *Monat* hinzugefügt. Die Variable *Monat* ist eine Faktorvariable mit 12 Levels für die 12 Monate im Jahr.

Dadurch wird das Modell um 11 Koeffizienten vergrößert:

```
Call:  glm(formula = VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke +
    Monat, family = poisson(), data = data1)

Coefficients:
       (Intercept)                PG3                PG4
         -18.48730           10.52491           10.65411
               PG5                PG6                GR5
          10.48217           10.90418            0.55376
               GR6                GR7                GR8
           0.75028            0.61631            0.07361
               GR9                GRXS               GRS
          -0.80565           -1.77023            0.24931
               GRM                GRL                GRXL
           0.88801            0.90408            0.38364
             GRXXL  log(FilialeUmsatz)        MarkeBenger
          -0.41301            0.61773           -0.63395
       MarkeErima          MarkeNike          MarkePuma
          -5.97214           -0.66883           -0.04418
      MarkeReebok          MarkeUmbro             Monat2
          -0.33834           -0.46929           -0.17947
            Monat3             Monat4             Monat5
           0.19707            0.12169           -0.33556
            Monat6             Monat7             Monat8
          -0.38732           -1.10562           -0.81974
            Monat9            Monat10            Monat11
          -0.29119            0.13643            0.04503
           Monat12
           0.33204

Degrees of Freedom: 47680 Total (i.e. Null);  47647 Residual
Null Deviance:      63030
Residual Deviance: 45810        AIC: 102200
```

Das Modell hat sich durch die Einbeziehung des Monats tatsächlich deutlich verbessert. Die Devianz hat sich auf 45810 verringert. Abbildung 4.2 zeigt die Verkaufsmengen

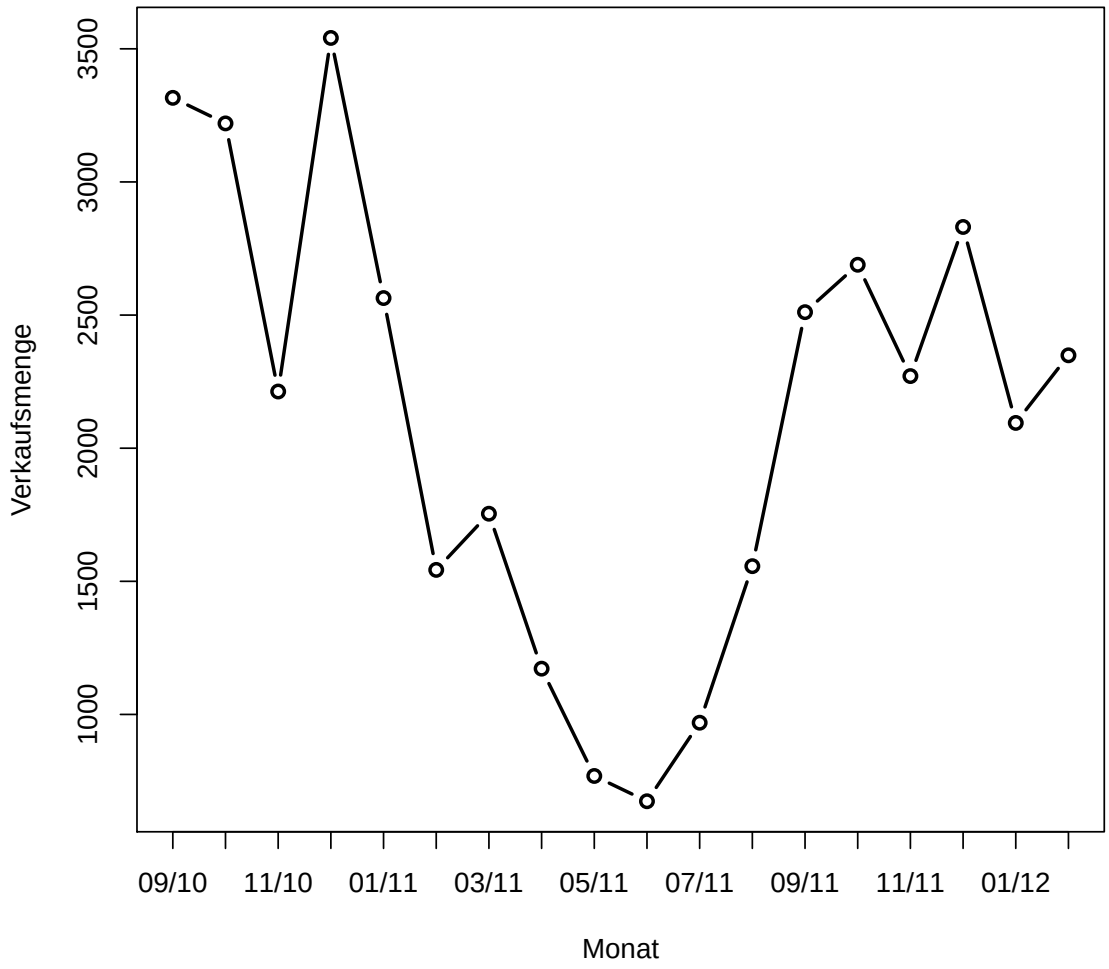


Abbildung 4.3: Gesamtverkaufsmenge nach Monaten

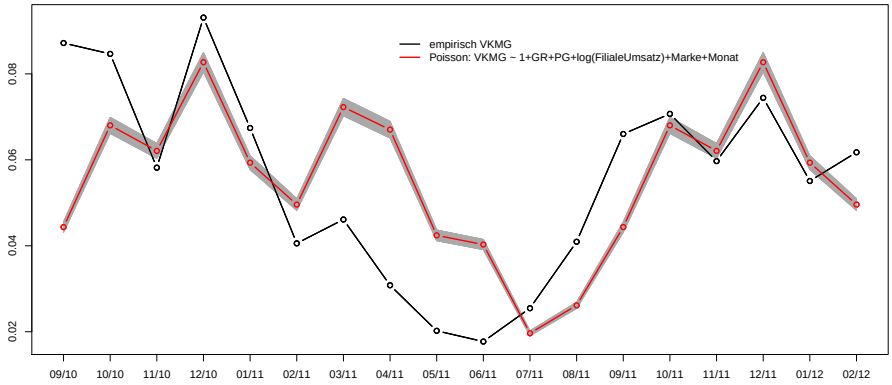


Abbildung 4.4: Gesamtverkaufsmenge nach Monaten: empirisch und nach Poissonmodell

nach dem Poissonmodell sowie ein entsprechendes 95% Konfidenzintervall (Berechnung analog Kapitel 5.3). Aus dem Modell resultiert vor allem für die Monate Februar bis September ein größerer Unterschied zu den empirischen Daten.

4.3 Test auf Überdispersion

Bisher wurden nur Poissonmodelle betrachtet. Durch Testen auf Unter- Überdispersion kann eine Aussage darüber getroffen werden ob andere Modelle geeigneter wären die Daten zu modellieren. Unter- beziehungsweise Überdispersion bedeutet, dass die Varianz der empirischen Daten kleiner beziehungsweise größer ist als die Varianz des geschätzten Modells. Bei der Poissonverteilung wird angenommen, dass die Varianz gleich dem Mittelwert ist. Ist diese Annahme grob verletzt so ist das ein Indiz dafür, dass die zugrunde liegenden Daten eher nicht Poisson verteilt sein. Eine Schätzung mit dem Poissonmodell kann trotz verletzter Equidispersionsbedingung sinnvolle Ergebnisse liefern.

Dennoch kann in diesem Fall erörtern werden, ob ein anderes Modell unter Annahme einer alternativen Verteilung besser geeignet ist, die Daten zu erklären. Eine Möglichkeit mit Überdispersion umzugehen ist die Verwendung des Negativen Binomialmodells, das bei nicht vorhandene Überdispersion dem Poissonmodell entspricht. Daher beruhen die vorgestellten Dispersionstest auf dem Vergleich zwischen Poisson Modell und Negativem Binomialmodell.

Einen einfachen Test auf Überdispersion bietet folgendes Verfahren [Cameron and Trivedi, 1998]:

Die Varianz wird über eine Varianzfunktion definiert:

$$\omega_i = V[y_i|x_i] \tag{4.2}$$

$$\omega_i = \mu_i + \alpha\mu_i^p \tag{4.3}$$

$$\tag{4.4}$$

Der Dispersionsparameter α wird aus den Daten geschätzt. Gilt $\alpha = 0$ so hat man den für Verwendung des Poisson Modells gewünschten Fall, dass die Varianz dem Mittelwert entspricht.

für $p = 1$ folgt:

$$\omega_i = \mu_i + \alpha\mu_i$$

und somit

$$\frac{\omega_i - \mu_i}{\mu_i} = \alpha$$

Es folgt:

$$\frac{(y_i - \hat{\mu_i})^2 - y_i}{\hat{\mu_i}} = \alpha + \epsilon_i$$

Nun kann α mit einer OLS Schätzung bestimmt werden. ϵ_i beschreibt den Fehler der OLS Schätzung. Anschließend wird mit einem t-Test auf $H_0 : \alpha = 0$ getestet. Für das Poisson Modell

```
VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke + Monat
```

folgt mit $z := \frac{(y_i - \hat{\mu_i})^2 - y_i}{\hat{\mu_i}}$ und $u := \hat{\mu}$:

```
Call:
lm(formula = z ~ 0 + u)

Residuals:
   Min     1Q Median     3Q    Max
 -1.83  -1.00   0.03   0.24 509.81
```

```
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
1  0.22633    0.02565   8.822   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.435 on 47680 degrees of freedom
Multiple R-squared: 0.00163,    Adjusted R-squared: 0.001609
F-statistic: 77.83 on 1 and 47680 DF,  p-value: < 2.2e-16
```

Als Ergebnis erhält man für den Dispersionsparameter $\alpha = 0.22633$. Der t-Test zeigt, dass α hoch signifikant ungleich 0 ist. Dieser Test legt nahe, dass Überdispersion in den Daten eine Rolle spielt.

Ein anderer Test auf Überdispersion ist in R in der Bibliothek *pscl* enthalten [Jackman, 2007]. Das Verfahren wird von Cameron,Trivedi [Cameron and Trivedi, 1998] erklärt:

Es wird ein Negatives Binomialmodell und ein Poissonmodell (Negatives Binomialmodell mit Dispersionsparameter $\alpha = 0$) mit den selben Daten geschätzt und die zugrundeliegenden Maximum-Likelihood-Funktionen der beiden Modelle herangezogen.

Ein Test auf $H_0 : \alpha = 0$ gegen $\alpha > 0$ wird durch einen Likelihood Ratio Test durchgeführt unter Verwendung von -2* Differenz der Log Likelihood Funktionen der beiden Modelle $d = 2(l_{NB} - l_{Poission})$. Die Teststatistik d ist jedoch nach keiner Standardverteilung verteilt, da der Dispersionsparameter α mit 0 begrenzt ist. Die Teststatistik ist 0 für alle Werte kleiner 0 und $\chi^2(1)$ verteilt für Werte größer 0. Cameron,Trivedi [Cameron and Trivedi, 1998] schlagen daher vor die Nullhypothese zum Signifikanzniveau δ zu verwerfen wenn $d > \chi^2_{1-2\delta}(1)$ statt $d > \chi^2_{1-\delta}(1)$.

Für das Poissonmodell

```
VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke + Monat
```

erhält man mit Aufruf der Funktion
`odTest(glm.nb(VKMG ~ 1+PG+GR+log(FilialeUmsatz)+Marke+Monat,data=data))`
folgenden Output (glm.nb Schätzt ein Negatives Binomialmodell,das entsprechende Poissonmodell wird automatisch geschätzt):

```
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  543.1705 p-value = < 2.2e-16
```

Der Wert der Teststatistik $D = 2(l_{NB} - l_{Poission})$ ist 543.1705, also weit größer als der kritische Wert 2.7055 zum Signifikanzniveau alpha=0.05. Entsprechend ist auch der zugehörige p-Wert sehr klein. Folglich ist die Nullhypothese, dass das Poissonmodell das wahre Modell ist zu verwerfen. Also weist auch dieser Test auf Überdispersion hin.

4.4 Negatives Binomialmodell

Die Ergebnisse der Dispersionstests legen nahe ein alternatives Modell zu testen, das besser in der Lage ist mit Überdispersion umzugehen. Das Negative Binomialmodell (siehe Kapitel 1.5) bewerkstelligt das durch eine Varianzfunktion die vom geschätzten Erwartungswert μ abhängt und somit die Varianz flexibler modellieren kann als das Poissonmodell.

Ausgehend von den bisherigen Anpassungen werden die selben Variablen verwendet die beim Poissonmodell das beste Ergebnis lieferten. Das Negative Binomial Modell ist in R mit Varianzfunktion $V(\mu) = \mu + \alpha\mu^2$ als Funktion glm.nb im Paket MASS implementiert.

```
Call:  glm.nb(formula = VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke +
    Monat, data = data1, init.theta = 9.199574673, link = log)

Coefficients:
       (Intercept)                PG3                PG4
         -29.49566           21.54586           21.67032
               PG5                PG6                GR5
          21.48109           21.91356            0.55654
               GR6                GR7                GR8
           0.75255            0.61409            0.07133
               GR9               GRXS                GRS
```

```
          -0.81323            -1.77714             0.24576
               GRM                 GRL                GRXL
           0.89058             0.90594             0.38415
             GRXXL  log(FilialeUmsatz)         MarkeBenger
          -0.41603             0.61812            -0.64311
       MarkeErima           MarkeNike           MarkePuma
          -5.97126            -0.67801            -0.04758
      MarkeReebok          MarkeUmbro              Monat2
          -0.34589            -0.47994            -0.17656
            Monat3              Monat4              Monat5
           0.20472             0.13842            -0.32794
            Monat6              Monat7              Monat8
          -0.37612            -1.10072            -0.81659
            Monat9             Monat10             Monat11
          -0.28750             0.13798             0.04957
           Monat12
           0.33376

Degrees of Freedom: 47680 Total (i.e. Null);  47647 Residual
Null Deviance:      57370
Residual Deviance: 41500        AIC: 101600
```

Ein Vergleich der Koeffizienten mit jenen des Poissonmodells zeigt auf den ersten Blick große Differenzen bei den Preisgruppen. Bei der Interpretation der Koeffizienten ist jedoch zu beachten, dass das Intercept den Koeffizienten für die Referenzkategorie PG1-MarkeAdidas-Monat1 beinhaltet. Alle anderen zu Faktorvariablen gehörenden Koeffizienten stellen die Differenz zum Intercept dar. Somit hat sich nur der Koeffizient für die Preisgruppe 1 stärker verändert. Die Nachfrage nach der Preisgruppe 1 war jedoch schon im Poissonmodell nahezu 0. Die Nachfrage nach dieser Preisgruppe hat sich nur noch stärker verringert. Bei allen anderen Koeffizienten bewegen sich die Differenzen der beiden Modelle im unteren einstelligen Prozentbereich, also sind die Ergebnisse der beiden Modelle sehr ähnlich.

4.5 Kreuzvalidierung

Eine Verfahren, Modelle zu vergleichen ist die Kreuzvalidierung [Canty and Ripley, 2012]. Hierbei wird eine Gesamtmenge N von Daten in k möglichst gleich große Teilmengen $T_1 \dots T_k$ unterteilt. Anschließend werden k Testläufe durchgeführt, wobei für den i-ten Testlauf die Menge T_i als Testmenge verwendet wird und die restlichen Daten als Trainingsmenge verwendet werden. Das Modell wird unter Verwendung der Trainingsmenge geschätzt. Anschließend wird die Testmenge herangezogen um den Fehler zum geschätzten Modell für diese Daten zu ermitteln (out of sample error). Die Gesamtfehlerquote wird als Durchschnitt aus den k Durchläufen bestimmt. Wählt man $k = N$ so erhält man die Leave-One-Out-Kreuzvalitierung, ist jedoch für große Datenmengen kaum praktikabel.

In R gibt es mit der Funktion *cv.glm* eine Implementierung der Kreuzvalidierung für Generalisierte Lineare Modelle[Canty and Ripley, 2012]. Die Daten werden zufällig in k gleich große Teilmengen eingeteilt. Eine Funktion zur Berechnung des Fehlers kann frei gewählt werden. Als Standard wird das durchschnittliche Fehlerquadrat ausgegeben. Das Verfahren kann nun dazu benutzt werden die bisherigen Modellanpassungen aus Kapitel 4.2 zu verifizieren.

Folgender Output zeigt das Ergebnis der Kreuzvalidierung mit $k = 10$ für die bisher betrachteten Modelle und einem zusätzlichen NULL Modell, dass nur aus dem Intercept besteht (als Fehler wird das durschnittliche Fehlerquadrat ausgegeben):

Modell	Fehler
VKMG ~ 1 (Poisson)	1.386749
VKMG ~ 1+VP+GR+Filiale+Marke (Poisson)	1.190916
VKMG ~ 1+log(VP)+GR+Filiale+Marke (Poisson)	1.188958
VKMG ~ 1+PG+GR+Filiale+Marke (Poisson)	1.185824
VKMG ~ 1+PG+GR+log(FilialeUmsatz)+Marke (Poisson)	1.189872
VKMG ~ 1+PG+GR+log(FilialeUmsatz)+Marke+Monat (Poisson)	1.101195
VKMG ~ 1+PG+GR+log(FilialeUmsatz)+Marke+Monat (Neg Bin)	1.102504

Auch durch die Kreuzvalidierung wird das Ergebnis aus der Devianzanalyse bestätigt, dass das Modell unter Einbeziehung des Monats das beste von den bisher betrachteten ist. Das Negative Binomialmodell ist bei dieser Analyse sogar noch etwas schlechter als das Poissonmodell. Zu beachten ist, dass die Einteilung in die 20 Gruppen zufällig stattfindet, daher können kleine Unterschiede im Ergebnis auch aus der unterschiedlichen

Wahl der Testmengen resultieren.

4.6 Vergleich der Verteilung

Ein Vergleich der empirischen Verteilungsfunktion mit jener theoretischen Verteilungsfunktion, die aus dem Modell resultiert kann ebenfalls dazu dienen die Qualität des Modells zu beurteilen.

Die theoretische Dichte aus dem Poissonmodell erhält man durch Einsetzen der geschätzten Erwartungswerte μ_i in die Wahrscheinlichkeitsfunktion für die Poissonverteilung.

$$p(Y_i = y) = \frac{\mu_i^y}{y!} e^{-\mu_i}$$

Durch Aufsummieren der Wahrscheinlichkeiten und Division durch die Anzahl der Beobachtungen n erhält man die gewünschte Wahrscheinlichkeit, dass ein Wert gleich y ist.

Dichte und Verteilung für das Negative Binomialmodell lassen sich aus der Wahrscheinlichkeitsfunktion des Negativen Binomialmodells berechnen.

$$f(y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y} \alpha \geq 0, y = 0, 1, 2, \ldots$$

Die empirische Dichte ist die relative Häufigkeit der beobachteten Verkaufsmengen aus den Daten.

Abbildung 4.5 zeigt die aus Poissonmodell und Negativem Binomialmodell resultierenden Dichten und Verteilungsfunktionen und stellt sie der empirischen Dichte und Verteilungsfunktion gegenüber. Das Poissonmodell scheint die empirische Verteilung etwas besser abzubilden.

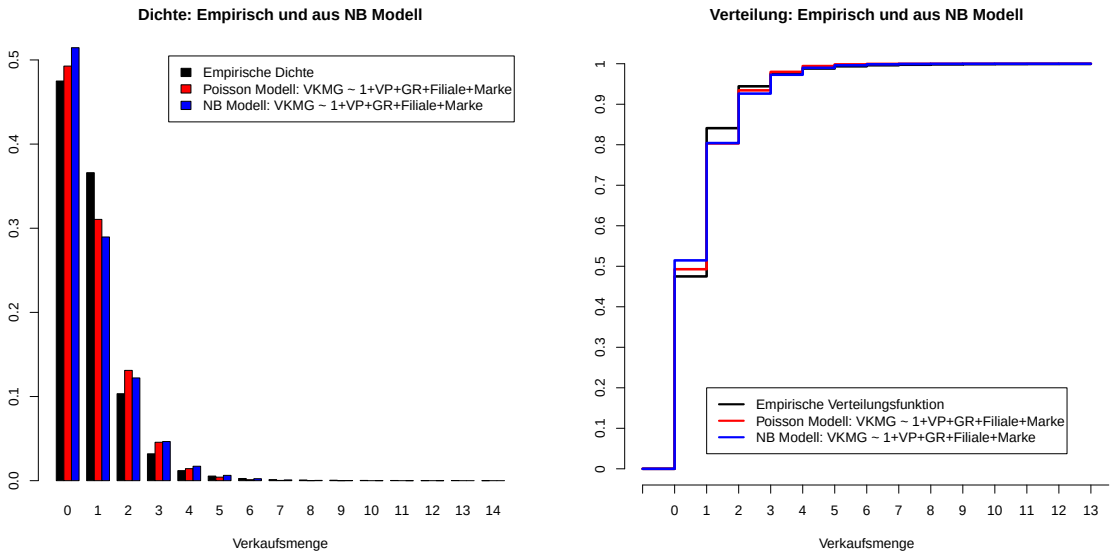


Abbildung 4.5: Dichte und Verteilungsfunktion: Empirisch, Negatives Binomialmodell, Poisson Modell

Kapitel 5

Schätzung des Größenlaufes

5.1 Berechnung des Größenlaufes aus dem Modell

Bei der Aufteilung der Waren ist eine sinnvolle Schätzung des Größenlaufes wichtig um Engpässe zu vermeiden und ein Überangebot an wenig nachgefragten Größen zu vermeiden.

Zunächst ist anzumerken, dass in den Datensätzen zwei verschiedene Größenskalen verwendet werden.

- XS. . . XXL
- 0. . . 10

Jedem Trainingsanzug ist eine Größe aus einem der beiden Größenschemen zugeordnet. Bei der Berechnung der Größenläufe aus den geschätzten Modellen werden diese getrennt, da die unterschiedlichen Gesamtverkaufsmengen zwischen den Größenschemen den Größenlauf verzerren würden.

Zunächst ist die Frage interessant, ob ein geschätzter Größenlauf überhaupt allgemein für verschiedene Artikel angewendet werden kann. Hierfür wird der empirische Größenlauf von verschiedenen Artikeln berechnet und grafisch dargestellt. Der empirische Größenlauf entspricht der Verkaufsmenge der einzelnen Größen über den gesamten Beobachtungszeitraum. Der Größenlauf wird so normiert, dass die Gesamtverkaufsmenge über alle Größen des entsprechenden Größenschemas gleich 1 ist. Abbildung 5.1 zeigt wie stark die Größenläufe von unterschiedlichen Artikeln über alle Marken variieren. Abbildung 5.2 zeigt die unterschiedlichen Größenläufe von verschiedenen Marken. Aus

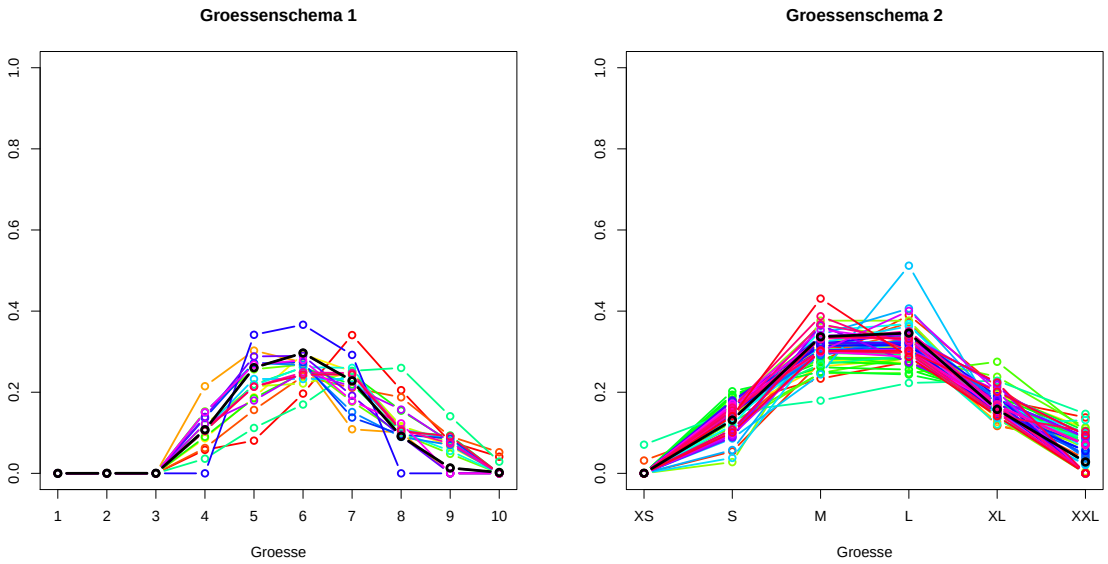


Abbildung 5.1: Empirischer Größenlauf für verschiedene Artikel

den Abbildungen ist erkennbar, dass das erste Größenschema nur von Adidas verwendet wird. Weiters sieht man, dass die Größenläufe für einzelne Artikel stärker variieren. Große Ausreißer bei den Größenläufen resultieren aus niedrigeren Stückzahlen der betroffenen Artikel. Die Größenläufe für die verschiedenen Marken sind wieder sehr ähnlich.

Die Größenläufe werden mit verschiedenen Modellen geschätzt und jeweils auch mit dem empirischen Größenlauf verglichen. Den empirischen Größenlauf über alle Filialen und Trainingsanzüge erhält man indem man die Verkaufsmengen einzeln für jede Größe aufsummiert und dann durch die Gesamtanzahl der verkauften Artikel dividiert.

Den geschätzten Größenlauf aus dem Poisson bzw Negativem Binomialmodell erhält man folgendermaßen:

Man betrachte die geschätzten Erwartungswerte:

$$\boldsymbol{Y} = \exp \boldsymbol{X'\beta} \tag{5.1}$$

Den Größenlauf erhält man nun, indem in 5.1 $\boldsymbol{X}$ durch eine Matrix $\boldsymbol{Z}$ ersetzt wird, bei der jede Zeile eine Größe repräsentiert. Die Zeilen der Matrix $\boldsymbol{Z}$ unterscheiden sich nur durch die Einträge, die die Größe repräsentieren. Die anderen Einträge werden für alle Zeilen auf 0 gesetzt. Für das Modell (mit Intercept geschätzt) erhält man somit

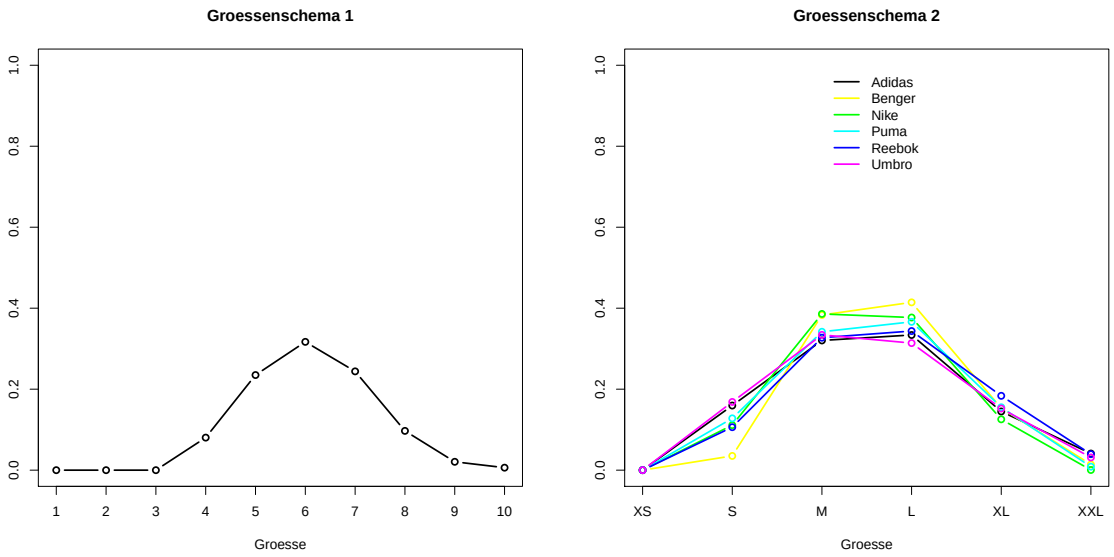


Abbildung 5.2: Groeßenlauf für verschiedene Marken

für die Matrix $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

Wenn das Modell mit Intercept geschätzt wird, so beinhaltet das Intercept den Koeffizienten für die erste Größe, sowie für jeweils die erste Kategorie jeder weiteren Faktorvariable. Die anderen Koeffizienten geben nur die Differenz zu dieser Referenzkategorie an.

Bei der Multiplikation von $\boldsymbol{Y}$ mit $\boldsymbol{\beta}$ resultiert ein Vektor $\tilde{\boldsymbol{g}}$ der genau aus den unterschiedlichen Koeffizienten von $\boldsymbol{\beta}$ für die Verschiedenen Größen plus Intercept (β_1) besteht. Nach Komponentenweiser Anwendung der Exponentialfunktion und Normierung hat man den Größenlauf $\boldsymbol{g}$ aus dem Modell.

$$\text{Referenzgröße:} \qquad \tilde{g}_1 = \exp \beta_1 \tag{5.2}$$

$$\text{die restlichen Größen:} \qquad \tilde{g}_i = \exp\left(\beta_1 + \beta_{j_i}\right), \qquad i = 2 \ldots m \tag{5.3}$$

$$\text{Normierung:} \qquad g_i = \frac{\tilde{g}_i}{\sum_{i=1}^{m} \tilde{g}_i}, \qquad i = 1 \ldots m \tag{5.4}$$

Alternativ kann das Modell auch ohne Intercept geschätzt werden, was das selbe Ergebnis liefert, nur ist die Darstellung der Koeffizienten anders. In diesem Fall fließt eine der Faktorvariablen, hier die Variable Größe, nicht in die Referenzkategorie ein. Alle anderen Faktorvariablen geben wieder die Differenz zur Referenzkategorie an.

Bei der Schätzung ohne Intercept sieht der Größenlauf so aus (β_i sind die zu den Größen gehörenden Koeffizienten):

$$g_i = \frac{\exp(\beta_i)}{\sum_{i=1}^{m} \exp(\beta_i)}, \qquad i = 1 \dots m \tag{5.5}$$

5.2 Analyse des Größenlaufes

Abbildung 5.3 stellt den empirischen Größenlauf und den aus dem Poissionmodell `VKMG ~ 1 + GR + PG log(FilialeUmsatz) + Marke + Monat` stammenden Größenlauf gegenüber, aufgeteilt auf die beiden unterschiedlichen Größenschemen. Für den empirischen Größenlauf werden auch Datensätze verwendet bei denen die Lagermenge am Monatsende gleich 0 ist. Es ist zu erkennen, dass aus dem Modell eine viel höhere Nachfrage nach der Größe 10 resultiert und dadurch der gesamte Größenlauf flacher ist als der empirische.

Betrachtet man die Verkaufszahlen für diese Größe, so sieht man, dass die Größe 10 nur für sehr wenige Artikel überhaupt angeboten wird. Folgende Tabelle zeigt die Anzahl der über den Beobachtungszeitraum verkauften Exemplare für Artikel, die auch in Größe 10 angeboten werden:

ARTNR	Marke	ARTBEZ	Verkaufszahlen nach Groesse									
			1	2	3	4	5	6	7	8	9	10
1260621	Adidas	TS BTS Logo ch 3S	0	0	0	20	22	27	31	21	2	5
1190456	Adidas	Hr. PES Trainer CL	0	0	0	41	59	141	248	145	55	29
1201400	Adidas	Hr. PES Trainer	0	0	0	6	27	46	37	20	3	7
1326065	Adidas	Tracksuit BTS Logo	0	0	0	24	60	92	81	72	35	20
1341758	Adidas	TS BTS	0	0	0	164	353	554	474	215	125	3
1192584	Adidas	Hr. WEB Trainer SMU	0	0	0	7	23	44	46	35	22	11
1149779	Adidas	Hr. WEB Trainer SMU	0	0	0	2	6	23	43	31	15	4
1341765	Adidas	TS BTS	0	0	0	10	31	47	70	72	39	8

Insgesamt wurden also in der Größe 10 nur 87 Exemplare verkauft.

Da der Größenlauf aufgrund des geringen Angebots deutlich überschätzt wird, werden alle Datensätze mit Größe 10 für die weitere Schätzung des Größenlaufes entfernt. Durch diese Bereinigung der Daten erhält man mit erneut geschätztem Poissonmodell den Größenlauf in Abbildung 5.4.

In Abbildung 5.5 ist nun auch der Größenlauf anhand des Negativen Binomialmodells `VKMG ~ 1 + PG + GR + log(FilialeUmsatz) + Marke + Monat` zu sehen. Die Größenläufe dieser beiden Modelle sind sehr ähnlich. Zusätzlich ist in der Abbildung auch der Größenlauf eines stark reduzierten Poissonmodells dargestellt. Das reduzierte Modell besteht nur aus einem Intercept und der Faktorvariable für die Größe als erklärende Variable `VKMG ~ 1 + GR`. Die Gegenüberstellung von Poisson und Negativem Binomialmodell sowie der Vergleich mit dem sehr simplen Modell lässt vermuten, dass es für die Schätzung des Größenlaufes einerseits irrelevant ist ob man sich für Poisson oder Negatives Binomialmodell entscheidet. Andererseits zeigt es, dass qualitativ das selbe Ergebnis mit sehr viel einfacheren Modellen erzielbar ist.

5.3 Konfidenzintervall für Größenlauf

Die geschätzten Koeffizienten des Modells sind asymptotisch normalverteilt (vgl. Cameron and Trivedi [1998]):

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, V[\hat{\boldsymbol{\beta}}])$$

Für eine Funktion $f(\boldsymbol{\beta})$ gilt:

$$f(\hat{\boldsymbol{\beta}}) \overset{a}{\sim} N(f(\boldsymbol{\beta}), V[f(\hat{\boldsymbol{\beta}})])$$

wobei

$$V[f(\hat{\boldsymbol{\beta}})] = \frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} V[\hat{\boldsymbol{\beta}}] \frac{\partial f(\boldsymbol{\beta})'}{\partial \boldsymbol{\beta}}$$

Als Konfidenzintervall für den Größenlauf anhand des Poissonmodells
`VKMG ~ GR+PG+log(FilialeUmsatz)+Marke+Monat` resultiert somit für die i-te Größe:

$$g_i(\boldsymbol{\beta}) = \frac{\exp(\beta_i)}{\sum_{j=1}^{m}[\exp(\beta_j)]} \tag{5.6}$$

$$\frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\exp(\beta_i)(\sum_{j=1}^{m}[\exp(\beta_j)] - \exp(\beta_i))}{\left(\sum_{j=1}^{m}[\exp(\beta_j)]\right)^2} \tag{5.7}$$

bzw. (5.8)

$$\frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_l} = \frac{-\exp(\beta_i + \beta_l)}{\left(\sum_{j=1}^{m}[\exp(\beta_j)]\right)^2} \text{ für } i \neq l \tag{5.9}$$

$$V[g_i(\hat{\boldsymbol{\beta}})] = \frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) V[\hat{\boldsymbol{\beta}}] \frac{\partial g_i(\boldsymbol{\beta})'}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) \tag{5.10}$$

und somit das 95% Konfidenzintervall: (5.11)

$$\frac{\exp(\hat{\beta}_i)}{\sum_{j=1}^{m}[\exp(\hat{\beta}_j)]} \pm 1.96\sqrt{V[g_i(\hat{\boldsymbol{\beta}})]} \tag{5.12}$$

Die für die Berechnung benötigte geschätzte Varianz-Kovarianz-Matrix $V[\hat{\boldsymbol{\beta}}]$ kann in R mit der Funktion `vcov` ausgegeben werden. Abbildung 5.6 zeigt den Größenlauf mit 95% Konfidenzintervall für das Poissonmodell. Für das Negative Binomialmodell erhält man leicht schmälere Konfidenzintervalle. Der Unterschied ist jedoch so gering, dass er in einer Grafik nicht erkennbar wäre.

Das Modell wurde hierfür ohne Intercept geschätzt. Als Referenzkategorie für die Faktorvariable Preisgruppe(PG) wurde `PG5` gewählt, als Marke `Adidas` und `Jänner` für die Faktorvariable `Monat`. Bei veränderter Wahl der Referenzkategorien ändert sich der relative Größenlauf und das entsprechende Konfidenzintervall nicht, jedoch die Werte der Koeffizienten. Nachfolgend sind die geschätzten Koeffizienten für die Größe mit den entsprechenden Standardabweichungen aufgelistet.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
GRXS              -9.77537    0.59037 -16.558  < 2e-16 ***
GRS               -7.75583    0.12468 -62.205  < 2e-16 ***
GRM               -7.11713    0.12351 -57.622  < 2e-16 ***
GRL               -7.10106    0.12340 -57.543  < 2e-16 ***
GRXL              -7.62149    0.12438 -61.274  < 2e-16 ***
GRXXL             -8.41815    0.12827 -65.629  < 2e-16 ***
```

```
GR4                -8.00514    0.12847 -62.312  < 2e-16 ***
GR5                -7.45138    0.12485 -59.681  < 2e-16 ***
GR6                -7.25486    0.12435 -58.342  < 2e-16 ***
GR7                -7.38882    0.12481 -59.198  < 2e-16 ***
GR8                -7.93153    0.12742 -62.248  < 2e-16 ***
GR9                -8.81079    0.14174 -62.160  < 2e-16 ***
```

5.4 Variation der Lagermenge

Bei den bisherigen Schätzungen wurden nur Datensätze betrachtet, bei denen der Lagerstand am Ende des Monats größer 0 ist um die Nachfrage aufgrund fehlenden Angebots nicht zu unterschätzen. Im Folgenden wird nun der Einfluss auf den geschätzten Größenlauf untersucht wenn diese vorgegebene minimale Lagermenge variiert wird. Es gibt Argumente, die für eine Erhöhung dieser Grenze sprechen. Kleinere Restbestände werden in den Geschäften nicht so einladend platziert wie größere Bestände, außerdem könnten kleinere Bestände aufgrund von Fehlbuchungen ein Angebot suggerieren, dass nicht existiert. Gegen eine Erhöhung dieser Grenze spricht, dass dadurch die Zahl der betrachteten Datensätze für die Schätzung reduziert wird und dadurch weniger aussagekräftig ist.

Bei Verwendung des Reduzierten Datensatzen bleiben 72801 Datensätze wenn auch Datensätze mit Lagermenge 0 am Ende des Monats berücksichtigt werden. Durch die bisher angewandte Einschränkung auf minimale Lagermenge 1 bleiben 47789 Datensätze übrig. Folgende Tabelle zeigt, wie sich die Anzahl der Datensätze durch weiteres Einschränken verringert.

min. Lagermenge	Anz. Datensätze	Verkaufsmenge	Umsatz
0	72.801	74.708	3.626.729
1	47.789	37.701	1.861.527
2	30.129	21.355	1.051.243
3	17.496	12.120	592.457
4	8.824	6.927	334.717
5	4.470	4.101	196.874
6	2.437	2.496	117.915
7	1.232	1.529	70.836

Die Anzahl der Datensätze verringert sich stark durch Erhöhung der minimalen Lagermenge. Der Umsatz der betrachteten Datensätze fällt um nahezu die Hälfte wenn die minimale Lagermenge um 1 erhöht wird. Abbildung 5.7 zeigt wie sich der Größenlauf, geschätzt durch das Poissonmodell `VKMG ~ 1+PG+GR+log(FilialeUmsatz)+Marke+Monat`, verändert wenn die minimale Lagermenge variiert wird. Der schwarze Größenlauf entspricht in der Grafik dem empirischen Größenlauf anhand der Verkaufsmenge wenn alle Datensätze verwendet werden. Die farbigen Kurven stellen den Größenlauf für die geschätzten Größenläufe dar. Es ist zu erkennen, dass bereits eine geringe Erhöhung zu einer verstärkten Nachfrage nach Randgrößen führt.

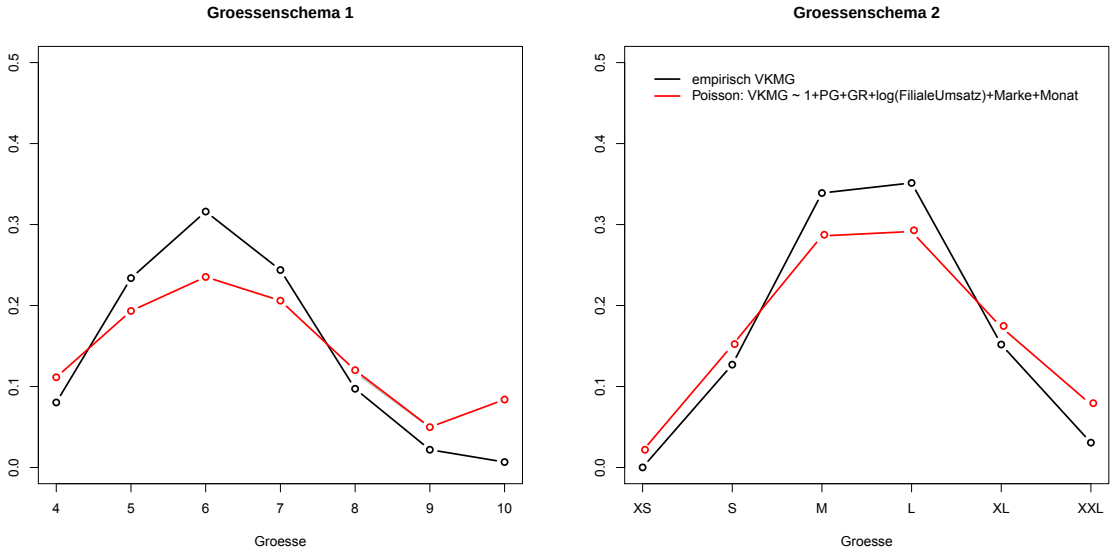


Abbildung 5.3: Größenlauf, empirisch und Poissonmodell

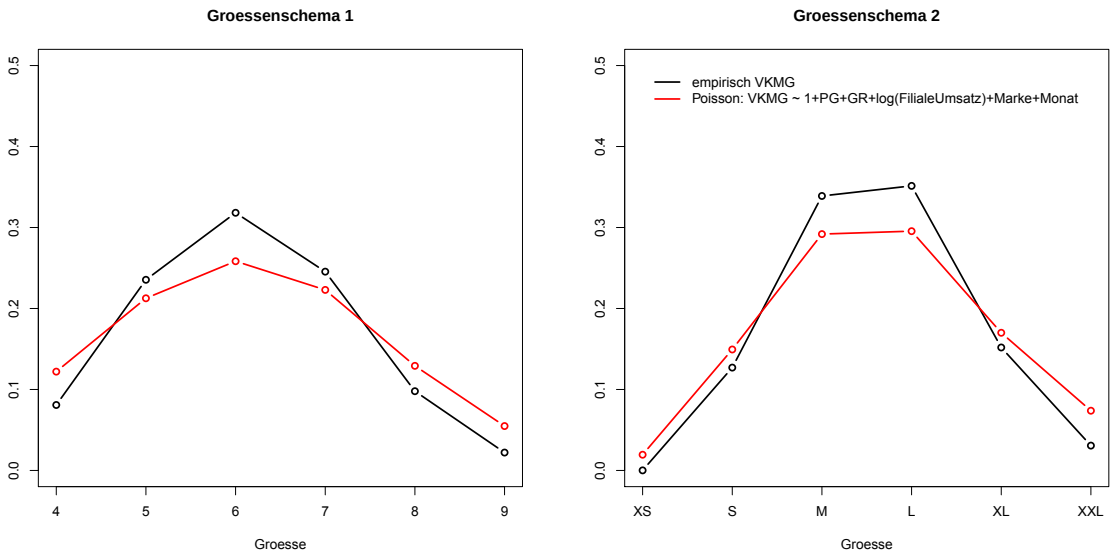


Abbildung 5.4: Größenlauf Poissonmodell ohne Größe 10

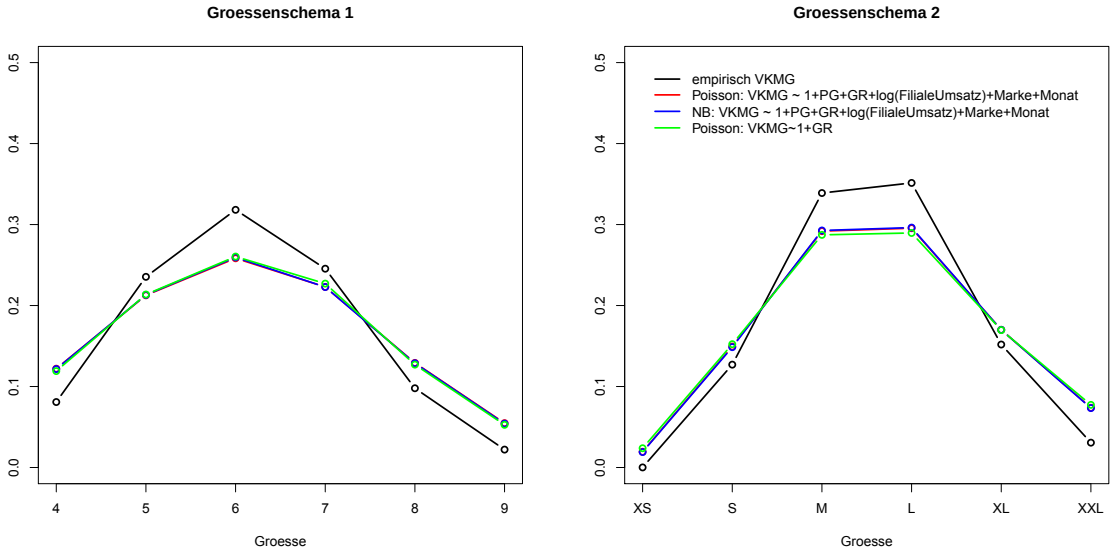


Abbildung 5.5: Größenlauf Poissonmodell und Negatives Binomialmodell

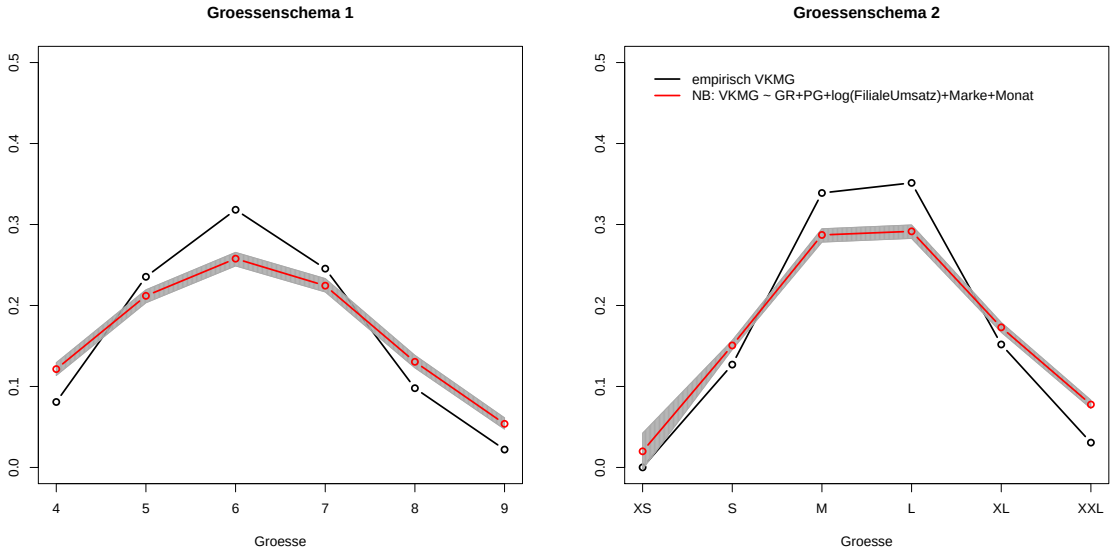


Abbildung 5.6: Groeßenlauf Poissonmodell mit Konfidenzintervall

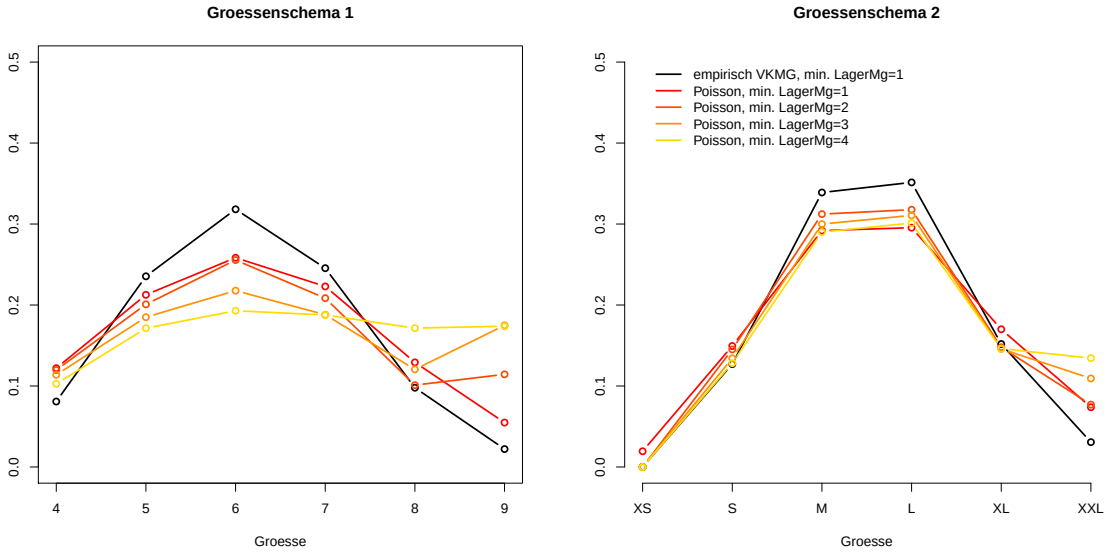


Abbildung 5.7: Größenlauf, Variation der minimalen Lagermenge für Poissonmodell

Kapitel 6

Fazit, Zusammenfassung

In dieser Arbeit wurde ein Überblick über die theoretischen Grundlagen von Generalisierten Linearisierten Modellen gegeben. Die theoretischen Ausführungen wurden so verfasst, dass die für eine praktische Anwendung benötigten Informationen auch ohne Vorkenntnisse verständlich sein sollten.

Im praktischen Teil wurde die Anwendung von verschiedenen Modellen zur Absatzschätzung anhand von Lager- und Verkaufsdaten eines österreichischen Sporthandelskonzerns demonstriert und Methoden zur Bewertung der Qualität von geschätzten Modellen vorgestellt. Ein wichtiges Resultat ist dabei der geschätzte Größenlauf für Trainingsanzüge. Die Schätzung der Nachfrage nach bestimmten Größen eines Artikels ist ein Problem mit dem jedes Handelsunternehmen konfrontiert ist.

Die aus Modellen resultierenden Größenläufe lassen auf eine stärkere Nachfrage nach Randgrößen schließen, als sie bei den betrachteten historischen Bestellungen angenommen wurde. Eine Empfehlung aus den Resultaten wäre daher die Bestellungen entsprechend anzupassen um Kosten durch Überangebot oder Engpässe zu verringern. Es hat sich gezeigt, dass für die Schätzung des Größenlaufes bereits sehr einfache Modelle Ergebnisse liefern, die qualitativ mit komplexeren Modellen vergleichbar sind, was eine technische Implemtierung vereinfachen würde.

Konkrete Empfehlungen für Bestellmengen bestimmter Artikel können aus den geschätzten Modellen nicht geliefert werden. Ein Problem stellt die Arbeit mit zensierten Daten dar. Aus den Daten ist nicht erkennbar wann innerhalb eines Monats Warenlieferungen oder Verkäufe stattgefunden haben. Man weiß daher nicht ob ein Artikel über den ganzen Monat angeboten oder etwa erst am letzten Tag des Monats geliefert wurde. Ein weiteres Problem ist die Tatsache, dass bestimmte Artikel

und Größen nur sehr selten angeboten werden und dann im Verhältnis zum Angebot auch gut verkaufen, was zu einer Überschätzung der Nachfrage dieser Artikel führen kann. Weiters ist anzunehmen, dass die Nachfrage nach bestimmten Artikeln nicht nur aus historischen Absatz- und Lagerdaten ermittelt werden kann. Vielmehr hängt die Nachfrage auch von anderen Einflüssen wie etwa Werbeausgaben oder der Art der Produktpräsentation in den Filialen ab. Für Konkrete Bestellempfehlungen müssten neben der Nachfrage auch finanzielle Kosten und Nutzen von verschiedenen Bestellstrategien berücksichtigt werden.
Trotz dieser Grenzen kann anhand der Modelle der Einfluss von bestimmten Größen abgeschätzt werden. So wurde in dieser Arbeit analysiert, wie sich die Nachfrage nach Trainingsanzügen über die Jahreszeit verändert oder wie sich Preisänderungen auswirken.

Literatur

A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, 1998. URL `http://statwww.epfl.ch/davison/BMA/`. ISBN 978-0-521-63201-0.

Angelo Canty and B. D. Ripley. *Bootstrap R (S-Plus) Functions*, 2012. R package version 1.3-7.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. URL `http://statwww.epfl.ch/davison/BMA/`. ISBN 0-521-57391-2.

A.C. Davison and A. Gigli. Deviance residuals and normal scores plots. Biometrika 76, 211-221 (1989)., 1989.

Ludwig Fahrmeir and Thomas Kneib. *Regression. Modelle, Methoden und Anwendungen*. Springer, 2009. ISBN 978-3-642-01836-7.

J. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2011. ISBN 9780521198158. URL `http://books.google.at/books?id=DDxEGQuqkJoC`.

Simon Jackman. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Stanford, California, 2007. URL `http://pscl.stanford.edu/`.

Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989. ISBN 0-412-31760-5.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

Rainer Winkelmann. *Count Data Models, Econometric Theory and an Application to Labor Mobility*. Springer, 1994. ISBN 3-540-57828-5.