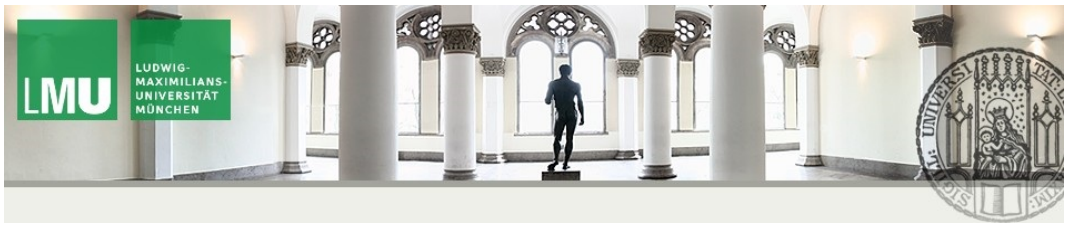

Masterarbeit

Affiliate Marketing: Analyse zeitlicher Aspekte im Online-Shopping

Maximilian Meingast
Institut für Statistik
Ludwig-Maximilians-Universität München
Kontakt: maximilian.meingast@googlemail.com

Betreuer:
Prof. Dr. Göran Kauermann

4. September 2013

Erklärung

Hiermit versichere ich, dass ich diese Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 4. September 2013 .....

Abstract

Affiliate-Marketing ist eine verbreitete Form von Online Marketing und heutzutage aus dem Werbemix vieler Unternehmen nicht mehr wegzudenken. Im Kern funktioniert Affiliate-Marketing über die Vermittlung zwischen werbetreibenden und werbetragenden Websites, den sogenannten Advertisern und Publishern. Durch einen Klick auf ein verlinktes Werbemittel auf der Website des Publishers gelangt der Nutzer in den Online-Shop des Advertisers. Wie im Performance-Marketing üblich, findet eine erfolgsbasierte Vergütung der Werbeträger statt. Der Erfolg von Werbemittelschaltungen bemisst sich an den Conversion Rates. Diese beschreiben das Verhältnis zwischen Ausbringung des Online-Werbemittels und den Klicks oder Bestellungen durch den Kunden.
In der vorliegenden Arbeit geht es darum, zeitliche Einflüsse auf die Erfolgsraten im Affiliate Marketing zu identifizieren. Zunächst wird eine deskriptive Analyse der umfangreichen Daten eines Affiliate-Netzwerkes durchgeführt. Untersuchungsobjekte sind dabei Advertiser aus dem Bereich Online-Retail. In einem kurzen Methodenteil wird ein Überblick über die statistischen Grundlagen für die Modellierungen gegeben. Die zeitlichen Einflüsse auf die Conversions werden in erster Linie anhand von Generalisierten Additiven Modellen (GAMs) sowie gemischter Modelle untersucht. Dabei werden mehrere zeitliche Ebenen betrachtet, im einzelnen Tages-, Wochen-, Monats- und Jahresverlauf. Des weiteren werden die Verweildauern zwischen Klicks und Orders im Affiliate Marketing mittels Methoden der Lebensdaueranalyse beleuchtet. Zudem findet eine Analyse des Einflusses von zeitlichen Komponenten auf die Höhe des mittleren Warenkorbwertes bei Online-Bestellungen statt.

Inhaltsverzeichnis

Kapitel 1

Motivation

1.1 Einführung

Online-Marketing hat in den vergangenen Jahren extrem an Bedeutung gewonnen und ist heutzutage aus dem Marketing-Mix von großen Unternehmen nicht mehr wegzudenken. Aufgrund der inzwischen flächendeckenden Internetnutzung kann mit Online-Kampagnen eine große Zielgruppe erreicht werden. Nach aktuellen Erhebungen macht der weiteste Nutzerkreis des Internets über 70% der deutschen Bevölkerung aus (BVDW, 2013). Die große Erreichbarkeit von Konsumenten, die vom Internet ausgeht, macht Online-Marketing für Unternehmen besonders attraktiv. Für das Jahr 2012 wurde der Anteil von Online-Werbung an den gesamten Werbeausgaben auf 21,8% geschätzt. Für Werbung im Internet werden in Deutschland nach der TV-Werbung die zweithöchsten Investitionen getätigt (BVDW, 2013). In jüngster Vergangenheit zeichnete sich ein fortwährender Zuwachs der Spendings für Online-Marketing ab.
Eine Form von Online-Marketing ist das sogenannte Affiliate Marketing (engl. „affiliate“ = der Partner). Dabei schließen Werbetreibende (Advertiser) und Werbeträger (Publisher) eine Partnerschaft, deren Ziel der Vertrieb von Produkten oder Dienstleistungen ist. Bei einer solchen Partnerschaft wird das Produkt oder die Marke des Advertisers auf der Website des Publishers beworben. Häufig geschaltete Werbemittel sind hier zum Beispiel Banner oder Textlinks, über welche der User auf die Website des Advertisers gelangt. Eine Impression entsteht, wenn dem Internetnutzer ein Werbemittel auf der Website des Publishers gezeigt wird. Sie ist nur dann als messbarer Erfolg zu werten, wenn sie zu einem Klick auf das Werbemittel führt. Entsprechend findet im Affiliate Marketing beispielsweise eine Vergütung auf Klick-Basis statt. In diesem Fall vergütet der Advertiser den Publisher für die Anzahl generierter Klicks auf sein Werbemittel. Er erhält im Gegenzug die Wahrnehmung seines Produkts bzw. seiner Dienstleistung. Im Idealfall wird vom User eine Order (Bestellung) generiert. Orders lassen sich weiter in Sales und Leads differenzieren. Während es sich beim Sale um einen Online-Verkauf handelt, können Leads ganz unterschiedliche Formen annehmen und führen nicht unmittelbar zu einer monetären Transaktion zwischen Konsument und Advertiser. Ein Lead liegt zum Beispiel vor, wenn sich der Internetnutzer nach erfolgtem Klick auf ein Werbemittel zu einem Newsletter des Advertisers anmeldet. Auch Orders können zu einer Vergütung des

Werbeträgers führen. Wann es im Einzelnen zur Vergütung kommt, hängt von der vertraglichen Vereinbarung der Partner ab. Es wird zwischen Pay per Click-, Pay per Lead- und Pay per Sale-Vergütung unterschieden. Aufgrund dieser leistungsorientierten Vergütungsstruktur zählt Affiliate Marketing zum sogenannten Performance Marketing. Der Werbetreibende zahlt ausschließlich für messbaren Werbeerfolg.
Um die Messbarkeit des Erfolgs und auch eine genaue Abrechnung zwischen den Partnern zu gewährleisten, müssen Orders, Klicks und teilweise auch Impressions dokumentiert werden. Technisch wird das unter anderem mittels Tracking-Pixeln oder Einsatz von Cookies umgesetzt. Aufgrund der immensen Datenfülle laufen die meisten Dokumentationsprozesse automatisch ab.
Der Weg des Kunden im Affiliate Marketing, als Teil der „Customer Journey“, führt von der Generierung einer Impression beim Internetnutzer über den Klick auf ein Werbemittel zur finalen Order (siehe Abbildung 1.1). Wie in der Graphik schematisch dargestellt, unterscheiden sich die drei Stufen zahlenmäßig recht stark. Aus einer Vielzahl von Impressions entstehen nur wenige Klicks. Lediglich ein Bruchteil dieser Klicks führt anschließend zur Order. Das Verhältnis von Klicks zu vorangegangenen Impressions wird als Click-through Rate (CTR) oder Klickrate bezeichnet. Die Relation zwischen Orders und Klicks wird durch die Conversion Rate (CR) beschrieben. Die durchschnittliche Click-through Rate für Standardbanner liegt in Deutschland bei ca. 0.10% (Internet World Business, 2010), d.h. pro 1000 Impressions wird im Schnitt nur einmal auf das Werbemittel geklickt.

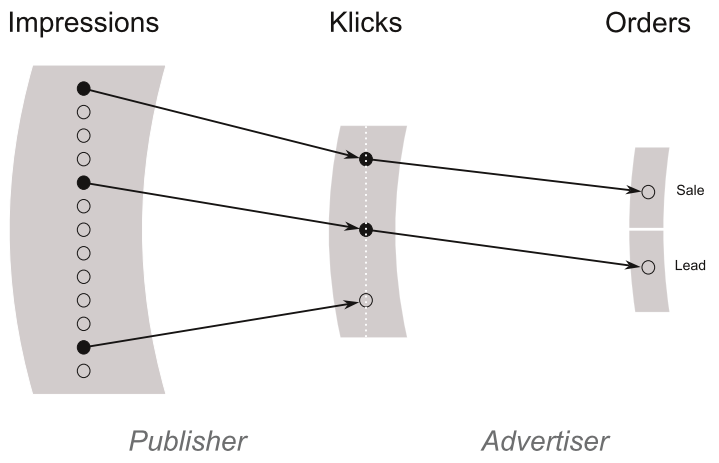

Abbildung 1.1: Schema Customer Journey im Affiliate Marketing

Im Affiliate Marketing wird der Kontakt zwischen Advertisern und Publishern in der Regel durch Affiliate Netzwerke hergestellt. Im Allgemeinen gibt es eine unüberschaubare Anzahl an Publishern, die Werbeschaltungen auf ihren Websites anbieten. Die Vermittlung durch Affiliate Netzwerke hat für Advertiser den Vorteil, dass sie ihre Werbung einem Pool aus geprüften Publishern zur Verfügung stellen können. Die Vergütung zwischen Advertiser und Publisher erfolgt in diesem Fall über das Netzwerk, das selbst eine Gebühr vom Advertiser erhält. Diese bemisst sich beispielsweise

prozentual zum Warenkorbwert der Bestellung.

1.2 Umbruch im Online-Marketing

Seit einigen Jahren stellt sich im Online-Marketing eine Inflation an Werbemittel-Schaltungen ein. Diese hat dazu geführt, dass der Tausender-Kontakt-Preis (TKP) in den letzten Jahren massiv zurückgegangen ist. Der Preisverfall hat weitreichende Ursachen. In den frühen Jahren des Online-Marketings waren die Preise für Werbemittel-Platzierungen deutlich überbewertet. Gerade die sich einstellende „Werbeblindheit“ hat dazu geführt, dass der Erfolg von Online-Werbemaßnahmen hinter den Erwartungen zurückblieb. Außerdem stieg die Anzahl von Werbemittel-Platzierungen im Zeitverlauf an, was unter anderem mit neuen Platzierungsstrategien zusammenhängt.
Ein Umbruch im Markt hat auch dahingehend stattgefunden, dass sich fortwährend neue Formen des Online-Marketings durchsetzen. So hat in neuester Zeit das sogenannte Retargeting im Affiliate-Marketing an Bedeutung gewonnen. Im Retargeting werden bei Besuchern von Online-Shops Cookies gesetzt und gespeichert. Beim späteren Besuch von Websites werden den potentiellen Kunden gezielt individuelle Werbemittel präsentiert. Dabei schließen Advertiser Werbeverträge mit großen Netzwerken. Ein Pool mit zahlreichen Publishern ist hier nicht erforderlich.
Um die Position des Affiliate Marketings im sich verändernden Online-Werbemarkt zu stärken, müssen neue Zielsetzungen erfolgen und bestehende Konzepte weiter verbessert werden. Besonders im Long Tail Bereich, in dem niedriger Umsatz generiert wird, ist ein Umdenken notwendig. Als mittelfristiges Ziel sollen statische Konzepte hin zum intelligenten und dynamischen Pooling modifiziert werden. Es ist daher zweckmäßig, die Prozesse im heutigen Affiliate Marketing zu verstehen und die Kenngrößen zu identifizieren. Die Kenngrößen des Marketing-Erfolgs sind vielfältig und umfassen zum Beispiel zeitliche, lokale und demographische Aspekte. Um die Wirkungsweise dieser Faktoren zu verstehen, ist es notwendig, empirische Daten auf verschiedene Fragestellungen hin auszuwerten. Die gewonnenen Informationen können dann eingesetzt werden, um die Marketing-Konzepte zu verbessern und um letztendlich bessere Conversions zu schaffen. Diese Arbeit befasst sich in erster Linie mit der Analyse von zeitlichen Aspekten im Affiliate Marketing. Dabei sollen Einflüsse von Tageszeit sowie Zeitpunkten im Monats- und Jahresverlauf auf die Conversion Rates untersucht werden. Gleichzeitig soll die Wirkungsweise einiger anderer Einflussgrößen und deren Wechselwirkungen mit den zeitlichen Komponenten identifiziert werden. Die Resultate sollen eine Hilfe zur Gestaltung eines dynamischeren Poolings durch zeitliche Differenzierungen bieten und zur Verbesserung der Conversions durch gezielte zeitliche Aussteuerung beitragen.

Kapitel 2

Datengrundlage

Die Datengrundlage für die folgenden Analysen bilden Datensätze eines Affiliate Netzwerkes. Das Unternehmen dokumentiert jede Transaktion zu Abrechnungszwecken. Die Daten wurden nicht grundsätzlich zum Zweck der Datenanalyse erhoben, es handelt sich um Sekundärdaten. Diese Tatsache muss in den nachfolgenden Analysen berücksichtigt werden und führt dazu, dass vereinfachende Annahmen getroffen werden müssen.
Aus der Vielzahl der vorliegenden Daten wurden zur Analyse Stichproben herangezogen. Der Umfang der Daten schließt eine Vollerhebung aus. Der zeitliche Rahmen der Stichproben variiert je nach Teildatensatz und Analysezweck. Einen groben Zeitrahmen bilden die Jahre 2012 und 2013, aus denen alle betrachteten Daten stammen. Alle Websites in den vorliegenden Datensätzen sind in Deutschland oder dem deutschsprachigem Raum ansässig.
Die Datensätze beschränken sich auf Advertiser aus dem Bereich „Online Retail“, die eine hochrelevante Gruppe im Online-Marketing darstellen. Innerhalb der Kategorie „Online Retail“ wurde wiederum eine Stichprobe von Advertisern extrahiert. Diese Auswahl der Advertiser erfolgte anhand einer Selektion, nicht durch einen Zufallsprozess. Die Selektion wurde vom Affiliate Netzwerk für vorangegangene Analysen vorgenommen. Sie erfolgte anhand verschiedener Kriterien. Es wurden Advertiser ausgewählt, welche seit einem längeren Zeitraum im Netzwerk aktiv sind, großen Umsatz generieren und deren Daten-Dokumentation ausreichend bekannt ist. Für die Analysen wurden sämtliche Affiliates pseudonymisiert. Anonymität wird auch dadurch gewährleistet, dass nur ein kleiner Anteil von Werbetreibenden aus dem Pool des Netzwerkes selektiert wurde. Nach Bereinigung liegen für die empirische Analyse Daten von insgesamt 75 Advertisern vor.
Für die Analysen liegen verschiedene Arten von Daten vor. Dabei handelt es sich um Impression-, Klick-, Order- und aggregierte Daten. In Kapitel 2.2 bis 2.5 werden diese genauer charakterisiert. Um die Datensätze zueinander in Beziehung setzen zu können, war es notwendig, einige Annahmen zu treffen. Problematisch ist, dass ein User in den Teildatensätzen keine identische ID-Nummer aufweist. Somit lässt er sich nicht ein-eindeutig über die Teildatensätze hinweg identifizieren. Daher ist unbekannt, an welchem Punkt ein bestimmter Nutzer verloren geht und wie sich seine Customer Journey im Affiliate Marketing genau gestaltet.

2.1 Variablen

In den verschiedenen Datensätzen liegt eine Vielzahl von Variablen vor, von denen viele zeitlicher Natur sind. So wird zum Beispiel das Klick-Datum zu Analysezwecken in verschiedene Komponenten aufgeteilt. Darüber hinaus gibt es andere Einflussfaktoren, von denen einige in allen Teildatensätzen auftauchen. Tabelle 2.1 gibt einen Überblick der wichtigsten dieser Variablen und deren Bedeutung.
Nicht alle Kovariablen sind selbsterklärend. Daher sollen einige wichtige Variablen im Folgenden genauer erläutert werden. Die Variable „Branche“ bezeichnet die Branche, in der ein Advertiser tätig ist. Hier werden im Weiteren die Ausprägungen Elektro, Fashion, Home & Accessoires, Kinder, Vollversender und Sonstige betrachtet. Bei der Branche „Kinder“ handelt es sich um Retailer, die Produkte für Kinder, wie zum Beispiel Spielzeuge vertreiben. Vollversender sind Advertiser, die eine große Produktbreite aufweisen, sodass sie nahezu jeder Branche zugeordnet werden könnten. Bei den Vollversendern handelt es sich also um große Online-Warenhäuser. Fashion-Advertiser vertreiben Bekleidung und Modeartikel. Bei der Branche Elektronik handelt es sich um Online-Shops für Elektro-Artikel und PC-Zubehör.
Die Variable „Business Model“ schafft eine Publisher-seitige Differenzierung. Sie gibt das Geschäftsmodell der Website eines Werbeträgers an. Die Websites der Publisher lassen sich in folgende Business Models einteilen:

- **Cash Back**: Online-Auftritte von Bonusprogrammen (Cash Back Programme)
- **Coupon**: Websites, die Rabatt-Coupons für Produkte anbieten
- **Media**: Online-Medien, zum Beispiel News-Websites
- **Portal & Communities**: Soziale Netzwerke und Foren
- **Preisvergleich**: Preisvergleichsportale (z.B. von Konsumgütern)
- **Topic Website**: Websites mit Themenschwerpunkt (z.B. Internetnutzung über DSL-Anschluss)

Ein in Deutschland noch recht unbekanntes Business Model ist das der Coupon Websites. Im europäischen Vergleich ist das Einlösen von Rabatt-Coupons hierzulande relativ unpopulär (Internet World Business, 2013). In den Rohdatensätzen gibt es noch weitere Geschäftsmodelle, die jedoch nur wenige Orders generieren und deshalb in den weiteren Analysen vernachlässigt werden. Im Einzelnen handelt es sich hierbei um Publisher Websites mit den Business Models Email Distributor, Suchmaschinen und Suchmaschinen-Marketing & Pay-per-Click.
Die Kovariable „Short-, Middle-, Long-Tail“ (STMTLT) ordnet die Advertiser in Größenklassen ein, gemessen am Umsatz, den sie im Netzwerk generieren. Long Tail Advertiser sind Werbetreibende mit sehr wenig Umsatz, die wenige Impressions, Klicks und Orders erzeugen. Entsprechend handelt es sich bei Short Tailern um sehr umsatzstarke Advertiser. Hier gilt zu beachten, dass für die Variable nicht der absolute Umsatz beim Advertiser betrachtet werden konnte. Es handelt sich um den Umsatz, den die Werbetreibenden durch ihre Netzwerkgebühr im Affiliate Netzwerk generieren. Bei der Netzwerkgebühr handelt es sich um einen festgeschriebenen Prozentsatz, gemessen an der Vergütung des Publishers, welcher an das vermittelnde

Affiliate Netzwerk abgeführt werden muss. Advertiser entlohnen Publisher für einen getätigten Sale mit einem vereinbarten Anteil am Warenkorbwert. Bei Leads (kein Warenkorbwert) handelt es sich in der Regel um eine fixe Vergütung. Da sowohl der Vergütungssatz nach Werbemittel, Advertiser und Order, als auch die Netzwerkgebühr nach Advertiser variieren können, handelt es sich um eine approximative Einteilung der Advertiser in Größenklassen. Ein Zugang zu den exakten Umsatzzahlen der Advertiser ist schwer, vor allem weil dann nur die Umsätze, welche durch Online Marketing geschaffen werden, extrahiert werden müssten. In der Variable wird also der Umsatz im Netzwerk innerhalb eines festgelegten Zeitraums erfasst. Die Advertiser werden in 7 Kategorien beginnend mit „0-100 EUR“ bis zu „mehr als 10000 EUR Umsatz“ eingeteilt.

Variable	Orders	Klicks	Aggr.	Erläuterung
AdvAccountManager	X	X	X	Hat Advertiser einen Key Account Manager im Netzwerk? (ja/nein)
AdvertiserCategory	X	X	X	Kategorie, in der Advertiser tätig ist
AdvertiserID	X	X	X	Advertiser ID-Nummer
Branche	X	X	X	Branche, in der Advertiser tätig ist
BusinessModel	X	X	X	Business-Modell der Publisher Website
ClickTime	X			Zeitpunkt des letzten Klicks auf Werbemittel vor der Order
ClickTimeOld	X			Zeitpunkt des vorletzten Klick auf Werbemittel vor der Order
CreativeType	X	X		Typ des Werbemittels
OrderPrice	X			Warenkorbwert
OrderType	X			Art der Order (Lead/Sale)
PartnerCategory	X	X	X	Kategorie, in der Publisher tätig ist
PubAccountManager	X	X	X	Hat Publisher einen Key Account Manager im Netzwerk? (ja/nein)
PublisherID	X	X	X	Publisher ID-Nummer
ST.MT.LT	X	X	X	Umsatzkategorie des Advertisers
VoucherCode	X			Wurde bei der Order ein Voucher eingelöst? (ja/nein)

Tabelle 2.1: Variablenübersicht für alle Datensätze

2.2 Klick-Daten

Die Daten zu den Klicks sind sehr umfangreich. Daher beschränken wir uns zunächst auf den Monat Februar 2013 als Stichprobe. Dabei werden analog zu den anderen Datensätzen 75 Advertiser betrachtet. Die Klick-Daten lagen im Zeitformat GMT vor und wurden entsprechend in das Zeitformat MEZ umgerechnet. Nach Bereinigung liegen über 4 Millionen Beobachtungen von 28 Variablen vor. Dabei sind viele Variablen zeitlicher Natur. Der Zeitpunkt eines Klicks ist im Datensatz sekundengenau dokumentiert. Abbildung 2.1 bietet eine Übersicht über die wichtigsten Variablen

im Datensatz.

2.3 Order-Daten

Hat ein Internetnutzer den kompletten Vertriebskanal des Affiliate Marketings durchlaufen, dann kommt es im letzten Schritt zu einer Order. Dabei handelt es sich um einen Sale oder Lead. Jede Order bei einem Advertiser wird dokumentiert. Dabei wird u.a. festgehalten, wann die Order und vorab der zugehörige Klick getätigt wurde, wie hoch der Warenkorbwert ist und von welchem Publisher der Kunde geschickt wurde.
Insgesamt liegen im Orders Datensatz Aufzeichnungen aus acht Wochen in 2013 vor. Es handelt sich um den Zeitraum vom 28. Januar bis 24. März 2013 (KW 5 - 12). Der Zeitraum wurde so gewählt, dass er jeden Wochentag gleich oft enthält.
Eine Herausforderung bei den Order-Daten besteht darin, dass nicht alle Orders automatisiert erfasst werden. In der Praxis finden auch sogenannte Order-Importe statt. Bei diesen Importen wird eine Order nachträglich vom Advertiser gemeldet. Häufig wird dabei der Zeitpunkt nicht übermittelt und vom System automatisch auf „0:00:00 01.01.2001“ gesetzt. Diese Beobachtungen wurden herausgefiltert und entfernt. Weitere Order-Importe konnten mittels einer Variable identifiziert und nachträglich aus dem Datensatz bereinigt werden. Zwei Advertiser mit einem hohen Anteil an Order-Importen wurden gänzlich aus dem Datensatz entfernt. Im Datensatz verbleiben 75 Werbetreibende, welche auch für die anderen Teildatensätze herangezogen werden.
In einigen wenigen Fällen wurde innerhalb einer Order mit unterschiedlichen Systemzeiten dokumentiert. Auch diese Fälle wurden aus dem Datensatz bereinigt. Außerdem wurden aus dem Order-Datensatz Beobachtungen entfernt, bei denen der letzte oder vorletzte Klick vor dem Kauf vor dem 1. Dezember 2012 erfolgt ist. Diese Ausreißer sind aufgrund ihrer geringen Anzahl vernachlässigbar und wurden ebenfalls entfernt.
Nach allen Bereinigungen umfasst der Datensatz rund 400.000 Beobachtungen von 61 Variablen. Viele Variablen sind dabei zeitlicher Natur. Zeit und Datum wurde jeweils in mehrere Variablen gesplittet, was eine Vielzahl neuer Variablen erzeugt.

2.4 Aggregierte Daten

Wird beim Besuch einer Website ein Werbemittel geladen, so wird beim User eine sogenannte Ad-Impression generiert. Diese Ad-Impressions werden im Online Marketing weniger präzise dokumentiert als Klicks oder Orders, weil sie in der Regel nicht zur Vergütung führen. Falls ein Publisher überhaupt Impressions dokumentiert, liegen diese auf Tagesebene aggregiert vor. Im vorliegenden Datensatz finden sich die täglichen Impressions, Klicks und Orders pro Partnerschaft.
Im Original-Datensatz lagen Daten aus dem gleichen Zeitraum wie in den Orders-Daten, d.h. 28.01. - 24.03.2013, vor. Auch an diesem Datensatz wurden Bereinigungen vorgenommen. Analog zu den anderen Datensätzen wurden wieder 75 Advertiser ausgewählt, 3 Advertiser wurden aus dem Datensatz entfernt.
Außerdem wurde der Datensatz auf Tage eingeschränkt, an denen mindestens eine

Impression stattgefunden hat. Alle anderen Fälle sind irrelevant, da ein Werbemittel nur dann zu Klicks und Orders führen kann, wenn es auch tatsächlich Usern gezeigt wurde.
Die Dokumentation der Impressions wird in der Regel von den Publishern vorgenommen. Bei manchen Publishern sind die Impressions für Abrechnungszwecke irrelevant, sodass die Anzahlen im Datensatz nicht vorliegen. Die Views von Textlinks werden generell nicht erfasst. Bannerlinks können durch das Laden eines Banner-Pixels dokumentiert werden.
Bei Publishern aus dem Geschäftsmodell „Media“ basiert die Vergütung bereits auf der Anzahl der generierten Ad-Impressions. Daher stehen in diesen Fällen die Impressions lückenlos zur Verfügung. Für viele Analysen, die die Impressions berücksichtigen, wurde der Datensatz auf die Media-Publisher reduziert. Bei Publishern mit Key Account werden die Anzahlen zudem vom Affiliate Netzwerk mitverwaltet. Der Datensatz wurde in einem weiteren Schritt auf diese Fälle eingeschränkt. Die Analysen der Impressions beschränken sich zwar auf einen Teildatensatz, dieser weist allerdings eine lückenlose Dokumentation aller relevanter Größen auf. Das erzeugte Subsample umfasst rund 4200 Tages-Beobachtungen aus den verschiedenen Partnerschaften.
Außer den Anzahlen von Impressions, Klicks und Orders pro Affiliate-Paar liegen noch einige andere Variablen in den aggregierten Daten vor (siehe Übersicht in Tabelle 2.1).

2.5 Jahresdaten

Der Datensatz zu den Jahresdaten ist im Grunde genommen identisch zu den aggregierten Daten aufgebaut. Statt Daten aus dem bisher betrachteten Acht-Wochen-Zeitraum in 2013, liegen hier tageweise aggregierte Beobachtungen aus dem gesamten Jahr 2012 vor. Die Kovariablen sind identisch mit denen aus den aggregierten Daten (Übersicht in Tabelle 2.1).
Die späteren Analysen der Jahresdaten basieren auf der Anzahl der Impressions. Daher wird der Datensatz wiederum auf Publisher des Business Models Media mit Account Manager gefiltert. Für die Analyse verbleiben dann rund 35.000 tageweise aggregierte Beobachtungen aus dem Jahr 2012.

Kapitel 3

Deskriptive Analysen

Im Folgenden findet eine explorative Analyse der vorliegenden Datensätze statt. Aus den dabei gewonnen Einsichten sollen später Hypothesen gebildet und statistisch getestet werden. Zunächst werden die Datensätze zu den Orders, Klicks und Impressions getrennt untersucht.
In Kapitel 2.1 wurden einige Kovariablen vorgestellt. Für die deskriptiven Analysen ist zunächst interessant, wie sich die 75 Advertiser auf diese Kategorien aufteilen. Bei der Variable Short-Middle-Long Tail Advertiser sind die Klassen mit niedrigem und hohem Umsatz unterbesetzt (siehe Tabelle 3.1). Diese Tatsache kann Implikationen auf die späteren Analysen haben. Günstiger ist die Stichprobe der Advertiser, was die Verteilung auf die Advertiser-Branchen angeht. Die häufigste Branche in der Stichprobe ist Fashion mit 27 Advertisern. Die kleinste Klasse Home & Accessoires umfasst immerhin noch 7 Advertiser. Die meisten betrachteten Advertiser sind Key Account-betreut. Nur 23 der 75 Advertiser haben keinen Account Manager. Für die Advertiser aus jeder Branche umfasst die Stichprobe sowohl Unternehmen mit als auch ohne Key Account (siehe Tabelle 3.2). Da die Datensätze weit über 1000 Publisher enthalten, stellt die Besetzung der Klassen publisher-seitig kein Problem dar.

0 €	0-100 €	100-500 €	500-1000 €
1	4	10	6
1000-2000 €	**2000-5000 €**	**5000-10000 €**	**>10000 €**
22	23	7	2

Tabelle 3.1: Verteilung der Advertiser auf die Umsatzklassen

	Elektro	Fashion	Home&Acc.	Kinder	Sonstige	Vollvers.	
Nein	2	8	3	4	1	5	23
Ja	6	19	4	5	9	9	52
	8	27	7	9	10	14	

Tabelle 3.2: Kreuztabelle Advertiser Account Manager und Branche

3.1 Orders

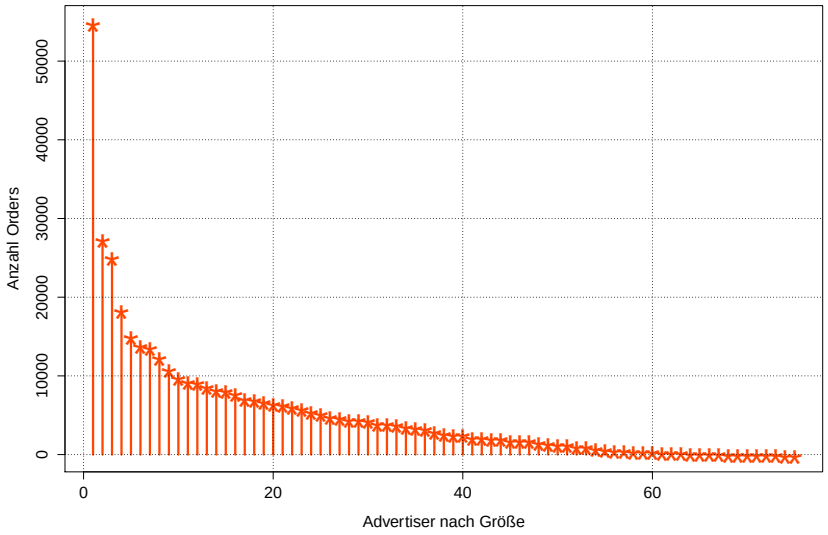

Abbildung 3.1: Advertiser nach Anzahl der Orders

Im Orders-Datensatz sind Sales und Leads, generiert aus Partnerprogrammen von 75 Advertisern mit über 2400 Publishern, dokumentiert. Die Orderzahlen sind heterogen zwischen den einzelnen Advertisern. Es gibt einige große Advertiser, bei denen in den betrachteten acht Wochen mehr als 10.000 Orders vorliegen. Die Mehrzahl der untersuchten Advertiser liegt unter dieser Grenze.
Bei den Orders im Tagesverlauf zeigt sich ein deutlicher Peak zwischen 19.00 und 22.30 Uhr. In den Nachtstunden, vor allem von 2.00 bis 6.00 Uhr, ist die Bestellaktivität am geringsten. Ab 6.00 Uhr findet ein deutlicher Anstieg der Orders statt, der einen vorläufigen Höhepunkt gegen 11.00 Uhr findet. In der Zeit bis 19.00 Uhr schwankt die Orderaktivität auf diesem Niveau mit leichten Peaks gegen 14.00 Uhr und 17.30 Uhr, vergleiche Abbildung 3.2 (oben). In der Analyse der einzelnen Advertiser (Abbildung 3.3) zeigt sich, dass die beiden größten Advertiser in ihren Orders genau diesen charakteristischen Tagesverlauf aufweisen. Um sicherzustellen, dass das Ergebnis nicht durch die beiden großen Advertiser verzerrt wird, betrachtet man in Abbildung 3.2 (unten) den kumulierten Tagesverlauf unter deren Ausschluss. Vergleicht man die zeitliche Verläufe, stellt man fest, dass der Peak zwischen 19.00 und 22.30 Uhr zwar ein wenig abgeschwächt wird, es aber ansonsten keine größeren Abweichungen vom ursprünglichen Verlauf gibt. Eine getrennte Betrachtung von Sales und Leads im Zeitverlauf zeigt keine relevanten Unterschiede.
Eine differenzierte Betrachtung der Orders im Tagesverlauf zeigt Unterschiede an den einzelnen Wochentagen, siehe Abbildung 3.4. Am Wochenende wird der erste Order-Peak erst zwischen 11.00 und 12.00 Uhr erreicht. Morgens ist die Bestellaktivität gegenüber den übrigen Wochentagen vergleichsweise gering. Dafür werden nachmittags verhältnismäßig mehr Orders getätigt, der abendliche Anstieg bis ca. 22.00 Uhr fällt etwas geringer aus. Die Wochentage Montag bis Donnerstag verhalten

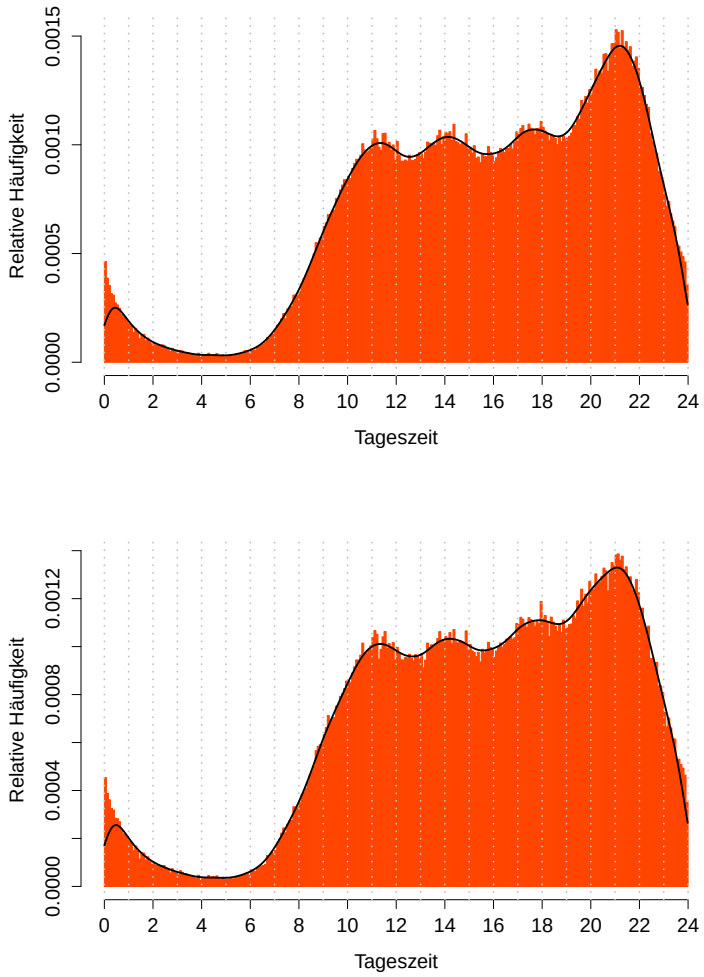

Abbildung 3.2: Orders im Tagesverlauf mit Kerndichteschätzer kumuliert über alle Advertiser (oben) bzw. exklusive der beiden größten Advertiser (unten)

sich nahezu gleichförmig. Auch freitags fällt der abendliche Anstieg geringer aus.
Neben den Tageszeiten könnten auch die Wochentage einen Einfluss auf die Anzahl der getätigten Bestellungen oder Leads haben. Die Analyse des Acht-Wochen-Zeitraums zeigt, dass sontags deutlich mehr Orders getätigt werden, als an allen anderen Wochentagen (vergleiche Tabelle 3.3). Dabei ist h_j die absolute Häufigkeit, d.h. die Anzahl der Orders, die im Zeitraum am jeweiligen Tag gezählt wurden. Die relative Häufigkeit f_j ist der relative Anteil der Orders pro Wochentag an den gesamten Orders. Der erwartete relative Anteil läge bei 14,29%, für den Fall, dass Unabhängigkeit zwischen Orderintensität und Wochentag vorliegt. Die empirische Analyse ergibt für Sonntag einen Anteil von 17,26% der Gesamtorders. Ob es sich dabei um einen signifikanten Effekt handelt, wird in späteren Ausführungen getestet. Die Wochentage Montag bis Donnerstag bewegen sich auf dem Niveau von ca. 14%. Die wenigsten Orders werden freitags und samstags mit einem relativen Anteil von

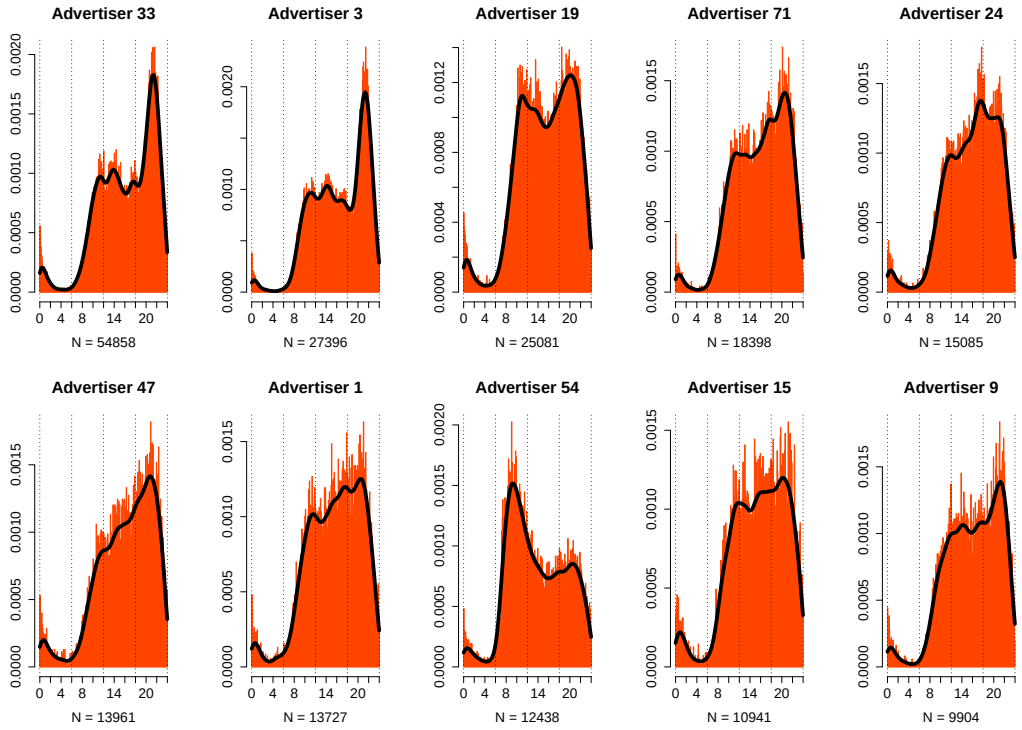

Abbildung 3.3: Tagesverlauf Orders für die 10 bestellstärksten Advertiser

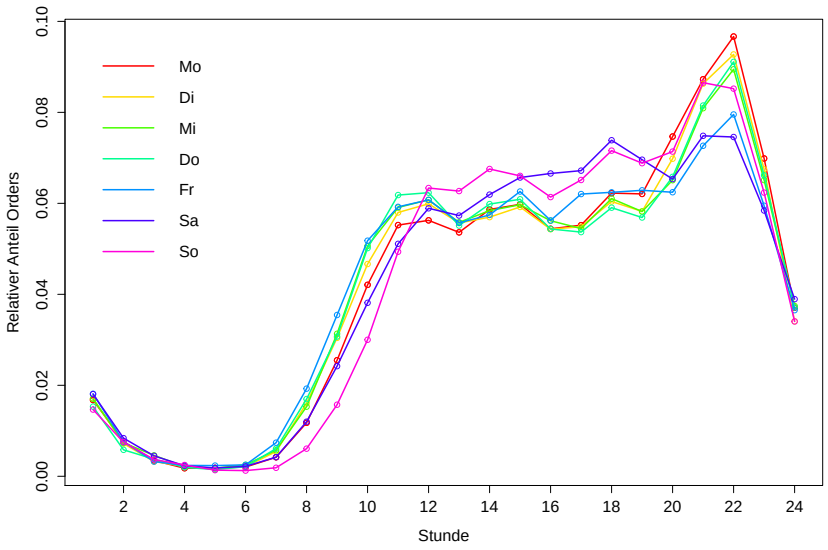

Abbildung 3.4: Relativer Anteil Orders pro Stunde pro Wochentag

jeweils ca. 13% getätigt.
Eine interessante Frage ist, ob die Orderaktivität bei allen Advertisern im Zeitverlauf homogen ist. Dazu wurden die Orders nach den 75 Advertisern getrennt betrachtet.

	Mo.	Di.	Mi.	Do.	Fr.	Sa.	So.
h_j	58496	55875	55769	56711	50682	50177	68372
f_j	14.77%	14.11%	14.08%	14.32%	12.80%	12.67%	17.26%

Tabelle 3.3: Absolute und relative Häufigkeiten Orders pro Wochentag

Die Frage nach der Homogenität im Zeitverlauf muss wohl verneint werden. In der deskriptiven Analyse zeigt sich, dass der Tagesverlauf aus Abbildung 3.2 zwar immer wieder auftaucht, sich die Orderaktivität aber je nach Advertiser unterscheidet. Abbildung 3.3 zeigt exemplarisch den Tagesverlauf der 10 Advertiser mit den meisten Orders. N beschreibt dabei die Anzahl der Orders pro Advertiser. Man sieht, dass zum Bespiel die Advertiser 19 und 54 vom üblichen Verlauf abweichen. Es ist also anzunehmen, dass sich der Tagesverlauf der Orders zwischen den einzelnen Advertisern recht stark unterscheidet.
Neben dem Effekt der Uhrzeit und des Wochentages könnte auch der Monatsverlauf einen Einfluss auf das Userverhalten haben. Die Analyse der täglichen Orders über alle Advertiser lässt sowohl einen Trend als auch eine zyklische Komponente vermuten. Im Acht-Wochen-Zeitraum (28. Januar bis 24. März 2013) findet ein Anstieg der Orderzahlen statt (vgl. Abbildung 3.5). Im Verlauf von Ende Januar bis Mitte März hat sich die Bestellaktivität der User bei den betrachteten Advertisern erhöht. Die Bestellungen am Sonntag (orange Balken) liegen dabei immer über dem Niveau der Vorwoche. Die deskriptive Analyse deutet darauf hin, dass der oben beschriebene Wochenverlauf als zyklische Komponente herangezogen werden kann.

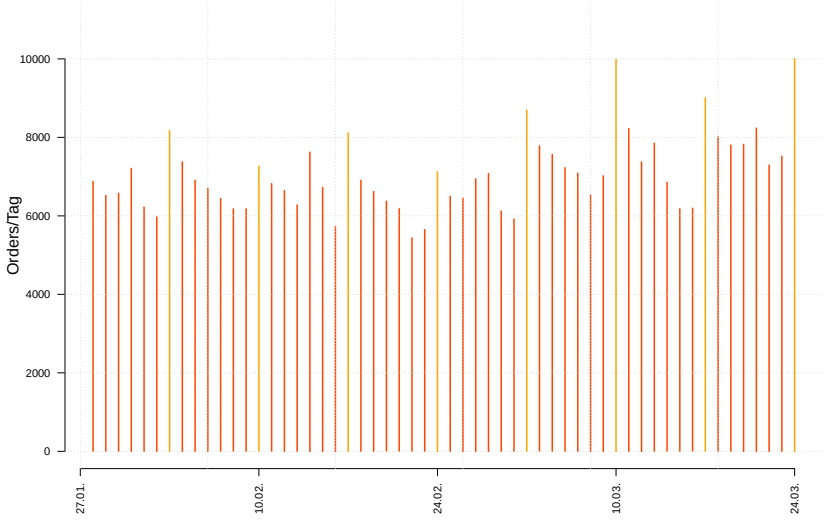

Abbildung 3.5: Orders pro Tag im Zeitverlauf

Es wird vermutet, dass Orders und Klicks zwischen den Affiliates heterogen sind. Aufgrund der Vielzahl von Affiliates können nicht alle Paare verglichen werden. Es ist daher sinnvoll, Advertiser und Publisher in vernünftige Gruppen zusammen-

zufassen. Für die Differenzierung bieten sich einige Variablen aus Tabelle 2.1 an. Die Kovariablen „AdvertiserCategory“ und „PartnerCategory“ teilen Advertiser bzw. Publisher in verschieden Kategorien ein. Allerdings liegen in diesen Kategorien sehr viele Ausprägungen vor, sodass manche Kategorien nur wenige oder gar einen Advertiser/Publisher enthalten. Außerdem teilen sich die Publisher zum Teil selbst in diese Kategorien ein, was nicht selten zu Verzerrungen führt. Eine gröbere Einteilung wird für die Variable „Branche“ für Advertiser bzw. „BusinessModel“ für Publisher getroffen. Hier ergeben sich ausreichend große Gruppen. Die beiden Variablen sind für eine Differenzierung advertiser- bzw. publisherseitig besser geeignet. Die Tabellen 3.4 und 3.5 liefern einen Überblick über die Aufteilung auf die einzelnen Kategorien. Die Advertiser aus den Branchen „Kinder“ und „Fashion“ erzeugen im vorliegenden Datensatz die meisten Orders. Von den 75 Advertisern vertreiben 27 Fashion-Artikel, nur 8 bzw. 7 Advertiser sind in den Branchen „Elektro“ und „Home & Accessoires“ tätig. Die Gruppen sind jedoch groß genug um eine vernünftige Differenzierung zu gewährleisten. Seitens der Publisher ist bei den Orders das Business-Modell „Coupon“ sehr stark vertreten (Tabelle 3.5). Bei den Topic Websites ist die Anzahl der Publisher, die Orders generierten, am höchsten (1007 Publisher). Allerdings ist die Gesamtzahl der Orders eher gering, d.h. pro Publisher wurden durchschnittlich wenige Orders generiert. Im Datensatz liegen noch weitere Business Models vor, die aber aufgrund einer geringen Anzahl von Orders hier nicht weiter berücksichtigt wurden.

Branche Advertiser	Anzahl Orders	Anzahl Advertiser
Kinder	125449	9
Fashion	116586	27
Sonstige	81889	10
Vollversender	36424	14
Elektro	26871	8
Home & Accessoires	8863	7

Tabelle 3.4: Vergleich Orders nach Advertiser-Branche

Business-Modell Publisher	Anzahl Orders	Anzahl Publisher
Coupon	211749	253
Cash back sites	81389	128
Media	35912	72
Topic Website	35280	1007
Price Comparison	18077	208
Portal & Communities	7704	352

Tabelle 3.5: Vergleich Orders nach Business Model

Differenziert man die Tagesverläufe bei den Orders nach dem Geschäftsmodell des Publishers, bei dem der Klick stattgefunden hat, so zeigen sich einige Unterschiede. Bei den Werbeträgern mit Business Model Cash Back oder Coupon ist der Aus-

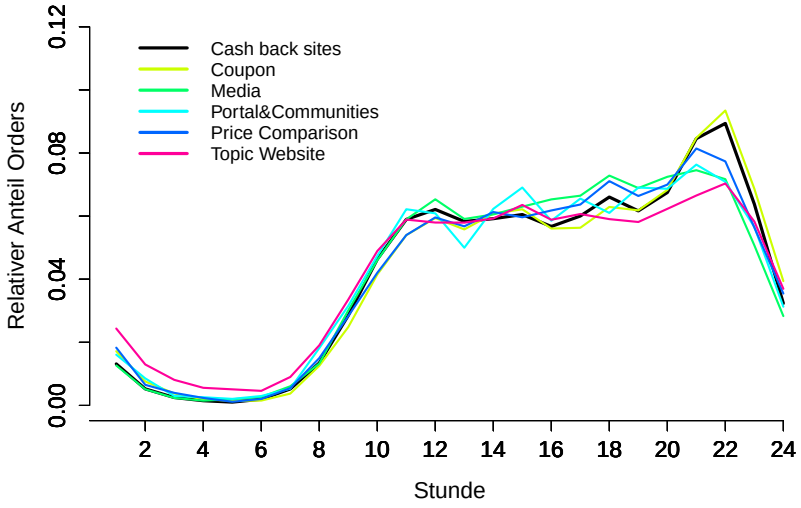

Abbildung 3.6: Relativer Anteil Orders pro Business Model im Tagesverlauf

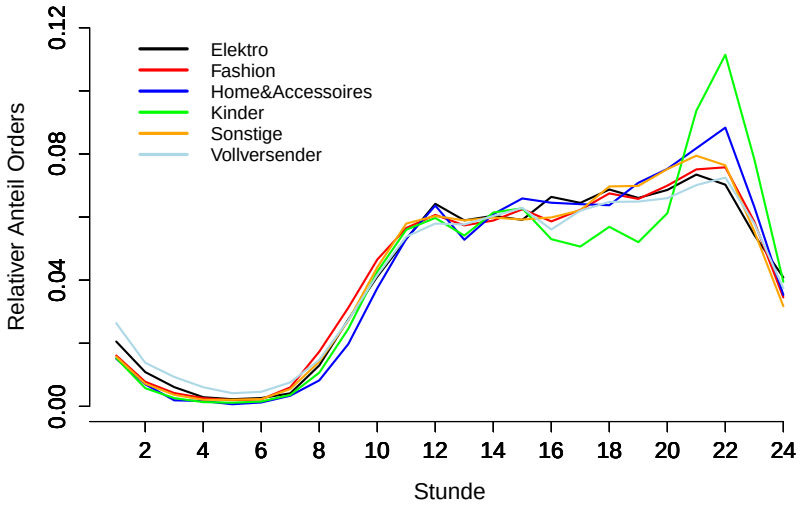

Abbildung 3.7: Relativer Anteil Orders pro Branche im Tagesverlauf

schlag zwischen 19.00 Uhr und 22.30 Uhr am stärksten ausgeprägt. Vergleichsweise wenige Orders zu dieser Uhrzeit generieren die Topic Websites. Diese weisen jedoch in den Nachtstunden eine überdurchschnittliche Orderaktivität auf (vgl. Abbildung 3.6). Bei der Abbildung werden keine Absolutwerte, sondern die stündlichen Anteile innerhalb eines Geschäftsmodells betrachtet, d.h. aufgrund der täglichen Orders im Business Model gewichtet. Das absolute Niveau der Orders pro Business Model ist recht heterogen und kann in der Graphik nicht abgelesen werden. Wie aus Tabelle 3.5 ersichtlich ist, erzeugt das Business Model Coupon mit Abstand am meisten Or-

ders im vorliegenden Datensatz.
Seitens der Advertiser gibt es in der Branche „Kinder“ bei den Orders den stärksten Ausschlag zwischen 19.00 Uhr und 22.30 Uhr. Bei den Produkten für Kinder ist die Orderaktivität nachmittags vergleichsweise niedrig (vgl. Abbildung 3.7). In den Nachtstunden gibt es bei den Vollversendern relativ viele Bestellungen. Ansonsten verhalten sich die Orders in den verschiedenen Branchen der Werbetreibenden annähernd gleichförmig. Auch hier handelt sich um den stündlichen Anteil an den gesamten Orders pro Tag, für jede Branche einzeln berechnet. Keine sichtbaren Unterschiede ergeben sich, wenn man die Orders danach differenziert, ob ein Voucher eingelöst wurde oder nicht.
Für jeden getätigten Sale wird der Warenkorbwert registriert. Abbildung 3.8 zeigt ein Histogramm des Warenkorbwerts der registrierten Bestellungen. Man kann hier von einer stark rechtsschiefen Verteilung ausgehen. Die meisten Sales haben einen Gesamtpreis im Bereich von 0 bis 200 Euro. Rund 78% der Bestellungen haben einen Warenkorbwert von 100 Euro oder weniger. Kleinbestellung mit einem Warenkorbwert von weniger als 20 Euro machen einen Anteil von 17% aus. Aus Illustrationsgründen wurden in Abbildung 3.8 nur Warenkorbwerte bis 1000 Euro dargesellt. Es wird hier nicht ersichtlich, dass es auch vereinzelt sehr hochwertige Einkäufe mit einem Kaufpreis von bis zu ca. 7350 Euro gibt. Durch diese Ausreißer wird der Mittelwert stark verzerrt. Ein besseres Lagemaß ist in solchen Fällen der Median. Er ist robust gegenüber Ausreißern und beziffert den Wert der mittleren Beobachtung. Weitere Quantile lassen sich aus Tabelle 3.6 ablesen. Jeweils q% der Orders im Datensatz haben einen Bestellwert kleiner gleich dem Quantilswert. Zum Beispiel haben 75% aller Bestellungen in den Daten einen Warenkorbwert von 88.45 Euro oder weniger. Hier liegt der Median bei 44.87 Euro, bei einem arithmetischen Mittel von 81.91 Euro. Die mittlere Beobachtung für den Warenkorbwert liegt also immerhin bei 44.87 Euro. Kunden, die durch Affiliate Marketing auf Online-Angebote aufmerksam werden, tätigen also nicht selten hochwertige Anschaffungen.

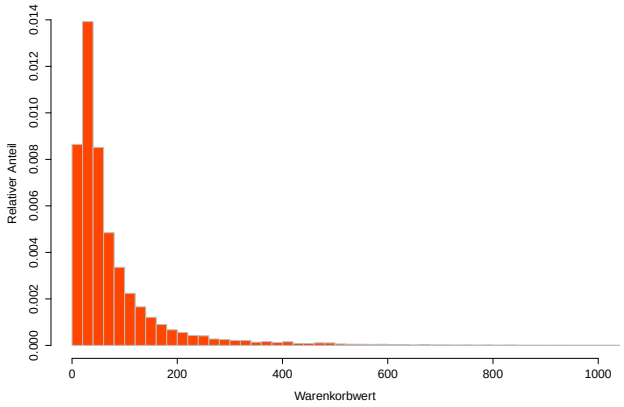

Abbildung 3.8: Histogramm Warenkorbwert (in EUR)

Der mittlere Wert des Warenkorbs ist stark von der Branche des Advertisers abhängig (siehe Abbildung 3.9). Die nach Branchen gruppierten Boxplots zeigen für die Lagemaße deutliche Unterschiede. Die höchsten Warenkorbwerte werden in der

q	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
€	9.34	14.95	18.49	21.76	25.17	28.36	31.85	35.27	39.98	44.78
q	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
€	50.37	57.60	65.98	75.59	88.45	106.20	130.76	172.90	270.90	7367.79

Tabelle 3.6: Quantile Warenkorbwert Sales

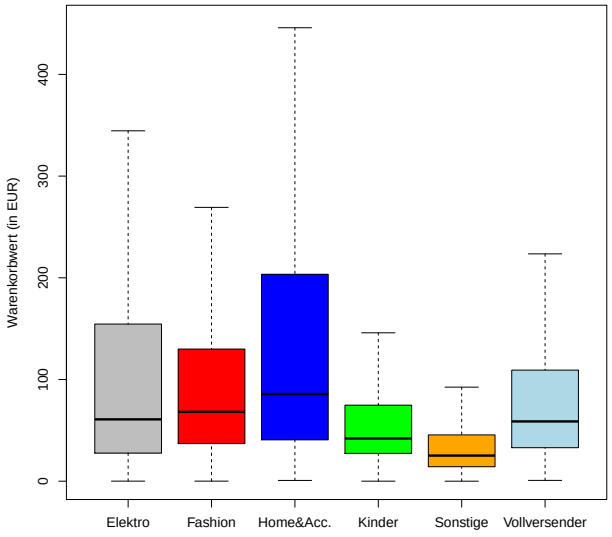

Abbildung 3.9: Boxplots Warenkorbwert vs. Branche (ohne Ausreißer)

Kategorie Home & Accessoires erwartet. Da die Retailer dieser Branche Möbel und Einrichtungsgegenstände vertreiben, sind hohe Warenkorbwerte erwartungsgemäß. Außerdem ist die Streuung der Warenkorbwerte für diese Advertiser am größten. Einkäufe von geringem Wert und mit kleinerer Streuung werden bei den Retailern für Kinderprodukte und den sonstigen Advertisern beobachtet.

3.2 Klicks

Aufgrund der großen Datenmenge wurden für die deskriptive Analyse nur die Daten aus dem Februar 2013 herangezogen. Setzt man die Klickaktivität der User im Tagesverlauf in Vergleich zu den Orders aus dem gleichen Zeitraum, dann fallen deutliche Unterschiede auf (siehe Abbildung 3.10). Die Klickzahlen sind viel gleichmäßiger über den Tag verteilt als die Orderzahlen. Auch bei den Klicks wird gegen 11.00 Uhr ein erster Peak erreicht. Allerdings kommt es fortwährend zu einem sanften Anstieg bis ca. 21.30 Uhr. Einen Ausschlag wie bei den Orders gibt es hier jedoch nicht. Vielmehr kommt es zu einem früheren Abfall der Klickzahlen. Bereits ab 21.30 Uhr nehmen die Klicks auf Werbemittel stark ab und erreichen ein Tief zwischen 2.00 Uhr nachts und 5.30 Uhr morgens. Verglichen zu den Orders ist die Nutzeraktivität bei den Werbemittel-Klicks nachts deutlich höher.

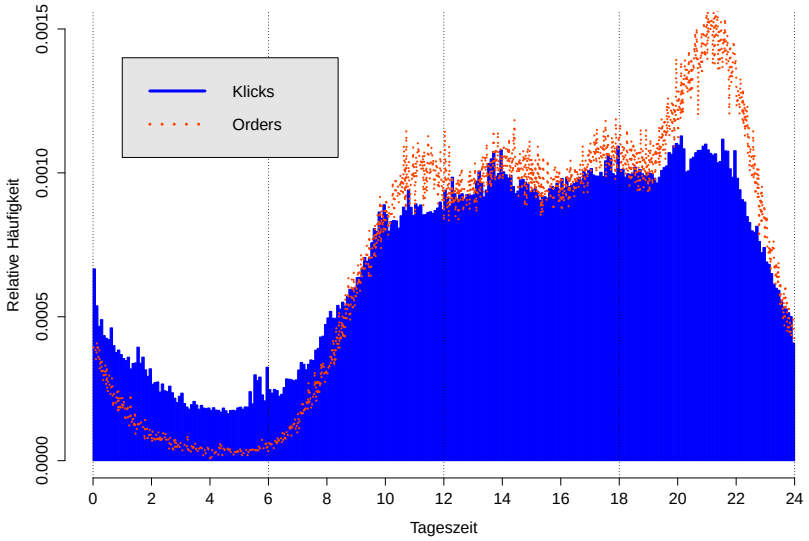

Abbildung 3.10: Klicks im Tagesverlauf (vs. Orders)

	Mo.	Di.	Mi.	Do.	Fr.	Sa.	So.
h_j	647558	626964	602448	606004	593472	569457	650633
f_j	15.07%	14.59%	14.02%	14.10%	13.81%	13.25%	15.14%

Tabelle 3.7: Absolute und relative Häufigkeiten Klicks pro Wochentag

Tabelle 3.7 zeigt die absoluten bzw. relativen Häufigkeiten verteilt auf die Wochentage (h_j bzw. f_j). Vergleicht man die Klicks pro Wochentag mit den Orders, fällt vor allem auf, dass der Ausschlag am Sonntag nicht so stark ist. Sonntags werden zwar immer noch die meisten Klicks generiert (15.14% der gesamten Klicks), allerdings ist der Anteil im Gegensatz zu den Wochentagen Montag und Dienstag nur wenig überhöht. Man kann sagen, dass die Klicks viel gleichmäßiger auf die Wochentage verteilt sind als die Orders (vergleiche Tabelle 3.7 und 3.3).
Der Vergleich von Klicks und Orders im Verlauf des Monats Februar 2013 zeigt einen relativ gleichförmigen Trend. Der gesamte Verlauf ist ähnlich, nur die Ausschläge an einzelnen Wochentagen unterscheiden sich. Abbildung 3.11 zeigt die Monatsverläufe auf unterschiedlichen Skalen.

3.3 Impressions/aggregierte Daten

Aufgrund der Datenmenge werden die Ad-Impressions in der Regel über den Tagesverlauf aggregiert. Außerdem führen sie in den seltensten Fällen zur Vergütung zwischen den Affiliates. Die Datenqualität leidet bei den Impressions im Gegensatz zu den Orders und Klicks unter dieser Aggregation. Die Anzahl der Impressions wird in der Regel von den Publishern übermittelt, allerdings gilt das nicht für alle Views. Die Views von Textlinks werden generell nicht erfasst, Bannerlinks können

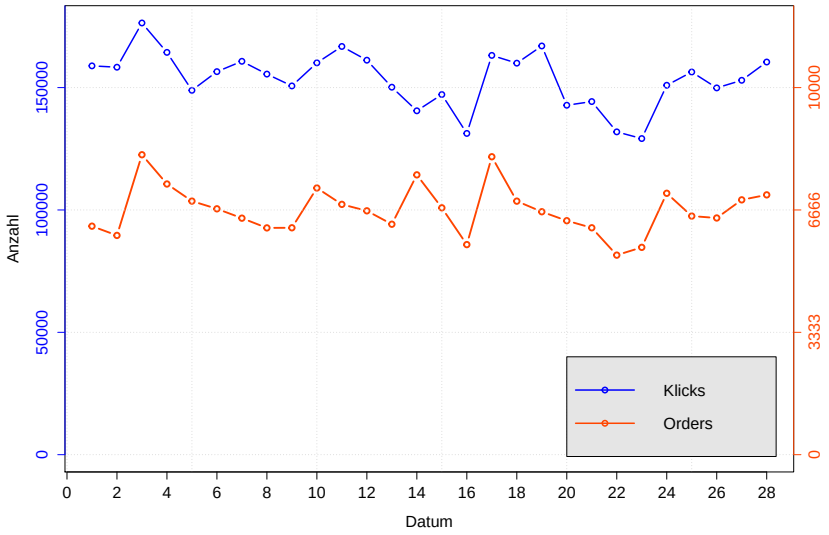

Abbildung 3.11: Klicks und Orders im Monatsverlauf

durch das Laden eines Banner-Pixels oder Postview gezählt werden. Ein Interesse an der genauen Dokumentation besteht vor allem bei Publishern aus dem Bereich Media. Um die Validität der Daten zu gewährleisten, wurde die Analyse auf diese Publisher und den Werbemitteltyp Bannerlink eingeschränkt. Außerdem werden nur Publisher einbezogen, die einen Key Account besitzen. Durch die enge Zusammenarbeit mit dem Netzwerk kann eine hohe Datenqualität sichergestellt werden. Die nachfolgenden Analysen beziehen sich auf diese relevanten Publisher.
Die tägliche Anzahlen an Impressions bzw. Klicks unterscheiden sich je nach Partnerschaft recht stark. Die Publisher-Websites sind unterschiedlich stark frequentiert, was bei den Views zur Heterogenität zwischen verschiedenen Websites führt. Große Publisher erzeugen pro Tag viele Ad-Impressions und Klicks auf Werbemittel, während kleinere Publisher wenige User erreichen. Abbildung 3.12 zeigt exemplarisch die täglichen Klick- und Impressionzahlen für einen Advertiser mit verschiedenen Partnern. Jede Kurve in der Graphik symbolisiert die Partnerschaft mit einem Publisher.
Es ist nachvollziehbar, an welchem Tag ein Werbemittel beim Publisher gesehen wurde, allerdings nicht wie lange und zu welchen Tageszeiten es geschaltet war. Durch die Aggregierung gehen diese Informationen verloren. Diese Restriktion muss für weitere Analysen berücksichtigt werden.
Durch die Heterogenität der Views pro Partnerschaft ist ein direkter Vergleich der Affiliates nicht möglich. Es ist erforderlich die Klicks zu den Impressions pro Partnerschaft in Relation zu setzen. Das Verhältnis zwischen Klicks und Impressions bezeichnet man im Online Marketing als Click-through Rate (CTR). Die Click-trough Rate kann als Maß für den Erfolg einer Werbemittelschaltung herangezogen werden. Sie beschreibt den Anteil der Nutzer, der nach einer Impression auf ein Werbemittel

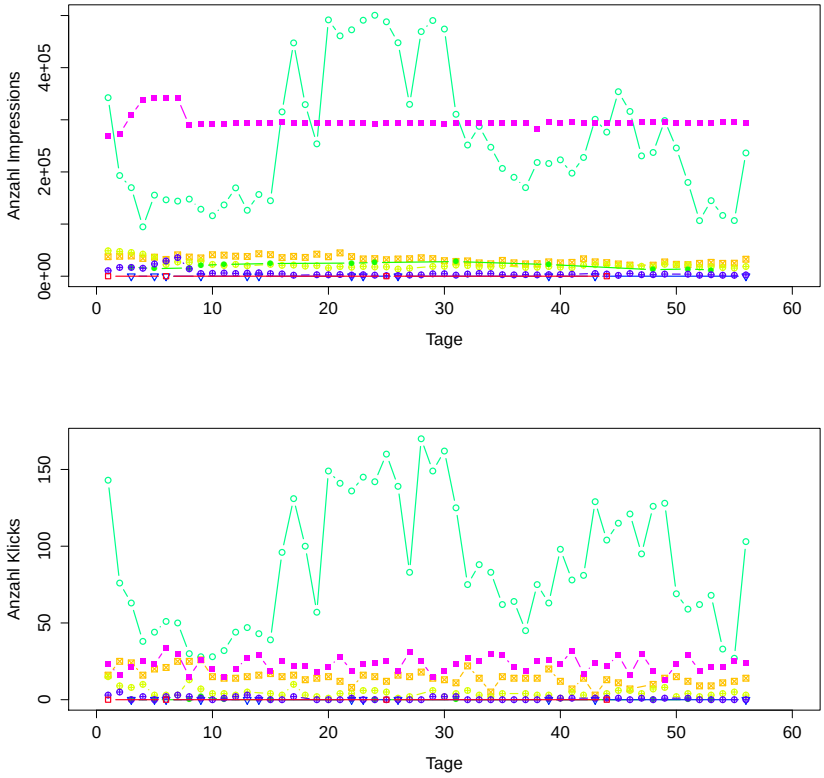

Abbildung 3.12: Impressions (oben) bzw. Klicks (unten) für einen Advertiser pro Publisher

geklickt hat. Die Click-through Rate ist definiert als

$$D_{ijt}^{CTR} = \frac{y_{ijt}}{v_{ijt}} \in [0, 1] \tag{3.1}$$

wobei v_{ijt} die Anzahl der Impressions einer Partnerschaft zwischen Publisher j und Advertiser i am Tag bzw. zum Zeitpunkt t ist. y_{ijt} bezeichnet entsprechend die Anzahl an Klicks von Affiliates i und j. Dabei wird die Annahme getroffen, dass der Zeitpunkt von Impression und Klick zusammenfallen. Da es sich um tageweise aggregierte Daten handelt, ist diese Annahme durchaus realistisch. Aus (3.1) wird ersichtlich, dass eine hohe Click-trough Rate auf zweierlei Arten zustande kommen kann. Entweder gibt es bei gegebener Anzahl von Klicks sehr wenige Impressions oder es gibt bei gegebener Anzahl von Impressions sehr viele Klicks.
Abbildung 3.13 zeigt die Click-through Rate eines Advertisers für seine verschiedenen Partnerschaften. Die meisten Kurven in der Graphik schwanken um einen individuellen Wert. Die blaue Kurve zeigt eine Partnerschaft, bei der sich die Click-through Rate sprunghaft ändert. Solche Sprünge werden in der Regel dadurch verursacht, dass beim Publisher wenige Impressions generiert werden und somit schon ein Klick/wenige Klicks zur sprunghaften Veränderung der Klickrate führen. Abbildung 3.15 (rechts) zeigt den Zusammenhang zwischen der Anzahl der Klicks und der

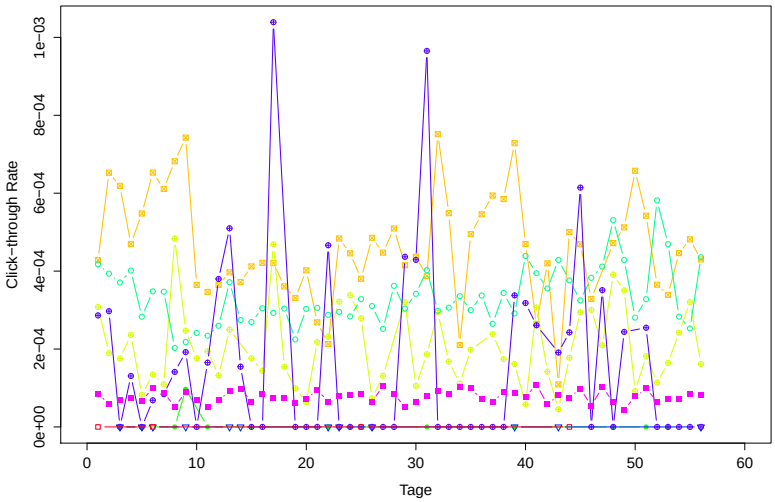

Abbildung 3.13: Click-through Rate für einen Advertiser pro Publishern

Click-through Rate. Die CTR ist bei Partnerschaften mit wenigen Klicks häufig erhöht, was im Umkehrschluss bedeutet, dass hier auch wenige Impressions vorliegen. Betrachtet man tägliche Klicks und Impressions in Absolutwerten (Graphik 3.14), dann fallen sofort die unterschiedlichen Größenordnungen auf. Gemittelt über alle Partnerschaften werden rund 20 Klicks pro Tag bei rund 87.000 Impressions pro Tag erzeugt. Zwar spiegelt das eine ungefähre Größenordnung des Verhältnisses von Klicks und Impressions wieder, allerdings muss beachtet werden, dass die Datengrundlage hier Media-Publisher sind. Bei selbigen werden Orders häufig mittels Postview getrackt und somit nicht alle Klicks erfasst. Beim Postview-Tracking werden den jeweiligen Publishern Orders über generierte Impressions zugeordnet. Eine Dokumentation von Klicks findet in diesen Fällen nicht statt. Es ist davon auszugehen, dass die wahre Click-through Rate deutlich höher liegt. Aufgrund dieser Tatsache ist man in den nachfolgenden Analysen nicht an der Bestimmung oder Prognose von Werten der Conversion Rates interessiert. Vielmehr geht es darum, Wechselwirkungen zwischen den beobachteten Rates und (zeitlichen) Einflussgrößen zu identifizieren. Um Prognosemodelle für erwartete Rates konstruieren zu können, müssten die Impressions umfassender für alle Partnerschaften dokumentiert sein und eine Differenzierung der einzelnen Tracking-Arten vorliegen.
Bei der Click-through Rate gibt es im Zeitverlauf immer wieder Ausreißer mit Raten von bis zu 50% (vgl. Abbildung 3.15 links). Allerdings kommen diese bei Werbemitteln mit wenigen Impressions zustande, bei denen jeder Klick zu hohem D_{ijt}^{CTR} führt. Abbildung 3.15 (rechts) zeigt, dass die Standardabweichung für Affiliates mit wenigen Klicks (und bei hoher Rate folglich wenigen Impressions) deutlich höher ist als für häufig gesehene Werbemittel. Auch das entspricht den Erwartungen, da die Varianz $\mathrm{Var}(\frac{Y}{V})$ für kleine V ansteigt.

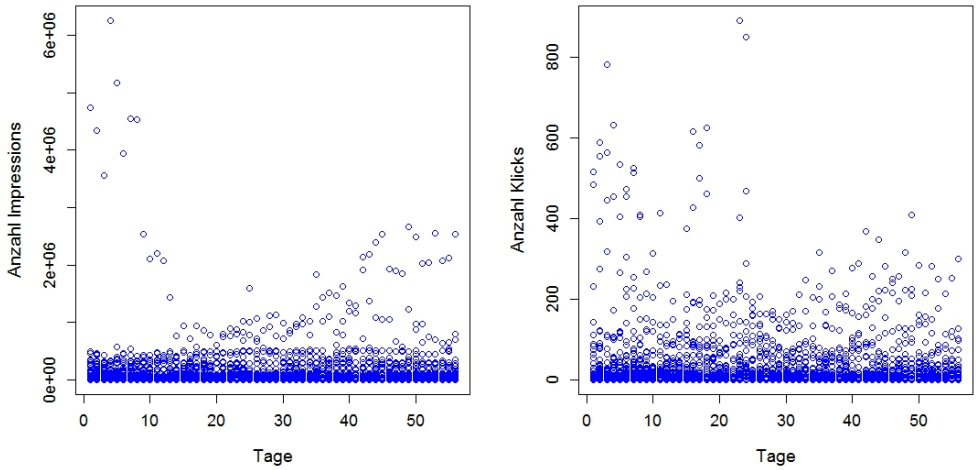

Abbildung 3.14: Tägliche Impressions bzw. Klicks pro Advertiser mit allen Publishern

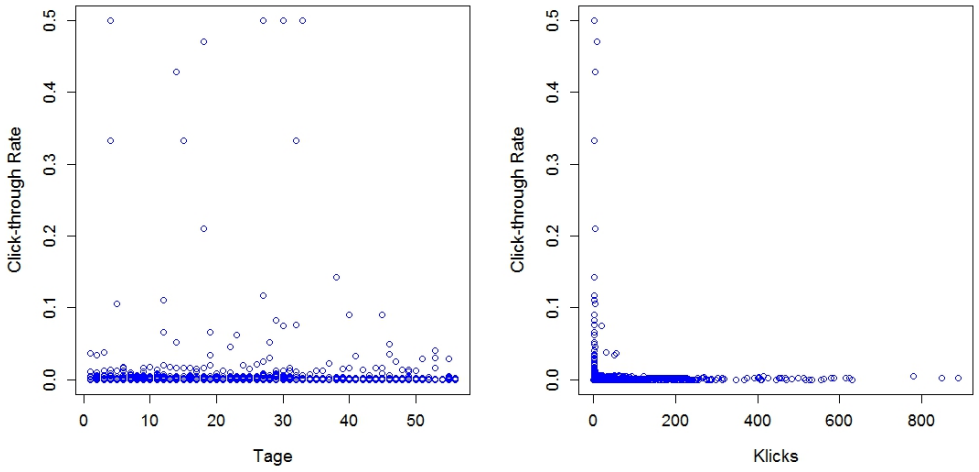

Abbildung 3.15: Click-trough Rate pro Tag pro Advertiser mit allen Publishern (links) und CTR in Abhängigkeit von der Anzahl Klicks pro Tag pro Advertiser (rechts)

3.4 Jahresdaten

Während in den folgenden statistischen Analysen die Jahresdaten auf Media-Publisher mit Key Account eingeschränkt werden, um der Erfassung von Impressions gerecht zu werden, wird nun ein Blick auf die absoluten Orders im kompletten Datensatz geworfen. Die absoluten täglichen Anzahlen an Orders, aggregiert über alle Partnerschaften, waren im Jahr 2012 zwischen Januar und Oktober relativ konstant (ver-

gleiche Abbildung 3.16 oben). Dabei unterlag die Anzahl der Orders zwar tageweisen Schwankungen (Punkte), die Moving Average geglättete Zeitreihe (Linie) weist allerdings in diesem Zeitraum keine erkennbaren Trends auf. Anders ist das in den Monaten November und Dezember. Ungefähr am 1. November beginnt der Anstieg der Orderzahlen, bedingt durch das Weihnachtsgeschäft. Dieser Anstieg erreicht seinen Höhepunkt eine Woche vor Heilig Abend, wobei hier die Zahl an Bestellungen gegenüber dem normalen Niveau verdoppelt bis verdreifacht ist (Abbildung 3.16 unten). Von diesem Höhepunkt fallen die Orders bis zum 24. Dezember rapide bis auf ein ganzjähriges Minimum ab. Das hängt wohl damit zusammen, dass viele Online Retailer vor Weihnachten eine Deadline ausgeben, was die Liefergarantie bis Weihnachten betrifft. Somit wird in den Tagen vor Weihnachten bereits weniger bestellt, weil viele Leute befürchten, dass die Artikel nicht rechtzeitig geliefert werden können.

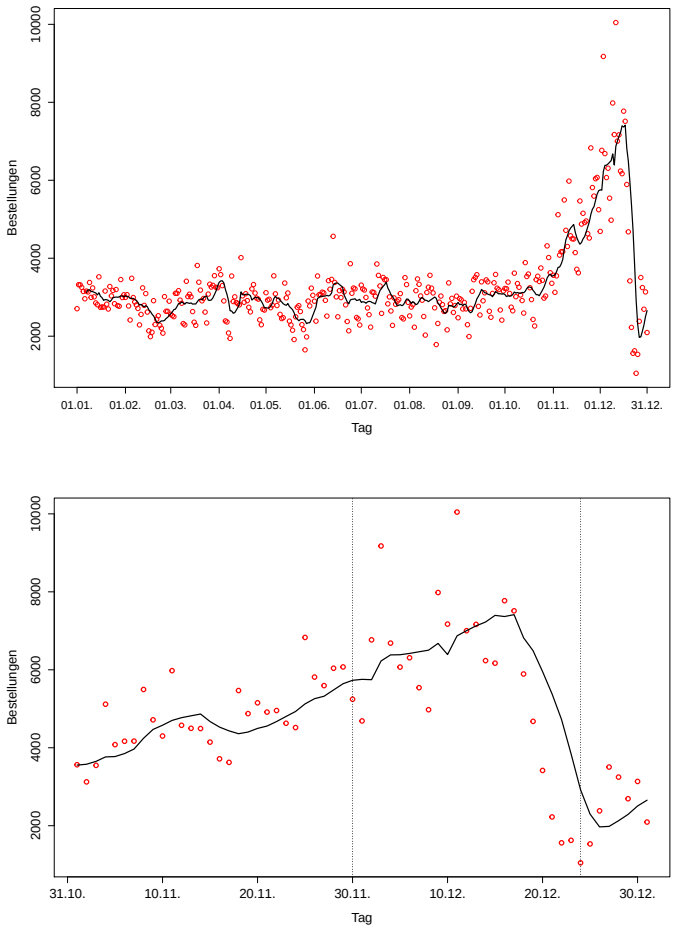

Abbildung 3.16: Orders im Jahresverlauf bzw. für die Monate November/Dezember mit geglätteter Zeitreihe

Kapitel 4

Statistische Grundlagen

4.1 Generalisierte Lineare Modelle

Regressionsmodelle kommen zum Einsatz, wenn der Einfluss einer oder mehrerer Variablen auf eine abhängige Zielgröße modelliert werden soll. Die klassische lineare Mehrfachregression

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \ , \quad i = 1, \ldots, n \tag{4.1}$$

modelliert den Einfluss mehrerer Kovariablen X auf eine Zielgröße Y. Dabei sind $\boldsymbol{\epsilon}$ die Residuen des Modells. Diese streuen mit konstanter Varianz um die Null. Es gilt

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \ .$$

Diese Störgrößen kommen zum Beispiel zustande, weil Einflussgrößen nicht im Modell erfasst wurden oder erfasst werden konnten. Das lineare Regressionsmodell baut auf restriktiven Annahmen wie Normalität, Homoskedastizität und Linearität auf. Durch Generalisierte Lineare Modelle (GLMs, Nelder und Wedderburn, 1972) werden diese Restriktionen aufgeweicht. Im GLM kann die Zielgröße durch verschiedene Verteilungen aus den einparametrischen Exponentialfamilien modelliert werden. Häufig verwendete Verteilungen sind hier Bernoulli-, Gamma-, Poisson- oder Inverse Gaußverteilung. Gegeben der Kovariablen $\boldsymbol{x}_i$ nimmt die Dichte einer einparametrischen Exponentialfamilie für die Zielvariable y_i die Form

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\ \omega_i + c(y_i, \phi, \omega_i)\right) \tag{4.2}$$

an. Der Parameter θ_i heißt in diesem Zusammenhang natürlicher Parameter und ϕ Dispersionsparameter. ω_i wird als Gewichtsfaktor bezeichnet und regelt den Umgang mit gruppierten Daten. Parameter ϕ und die Normierungskonstante $c(y_i, \phi, \omega_i)$ sind unabhängig von θ_i. Existieren erste und zweite Ableitung von $b(\theta_i)$, dann gelten für bedingten Erwartungswert und bedingte Varianz

$$\begin{aligned} \mathbb{E}(y_i|\boldsymbol{x}_i) &= \mu_i = b'(\theta_i) \\ \mathrm{Var}(y_i|\boldsymbol{x}_i) &= \phi\ b''(\theta_i)/\omega_i \ . \end{aligned}$$

Dank Einführung von Link- und Responsefunktion ist keine Transformation der Daten notwendig. Der bedingte Erwartungswert ist mit dem linearen Prädiktor $\boldsymbol{x}_i'\boldsymbol{\beta}$

mittels Responsefunktion $\mu_i = h(\eta_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ verknüpft. Die Linkfunktion g ist die Umkehrung der Responsefunktion, es gilt $g = h^{-1}$ und damit $\eta_i = g(\mu_i)$. Bei gegebener Verteilung der Zielvariable kann aus einer Vielzahl von Linkfunktionen gewählt werden. Jede Exponentialfamilie besitzt darüber hinaus einen kanonischen Link, für den $\theta_i = \eta_i$ gilt (siehe auch Fahrmeir, Kneib und Lang, 2009).
Für jedes GLM muss also eine Verteilungsannahme über die Verteilung der Zielgröße und eine Strukturannahme über die Gestalt der Linkfunktion getroffen werden. Durch die obige Form lassen sich verschiedene Verteilungen der Zielgröße für die Inferenz in eine einheitliche Form bringen. In der GLM-Theorie ist das klassische lineare Regressionsmodell enthalten. Es handelt sich dabei um ein GLM mit normalverteilter Zielgröße und Identitäts-Link.
Im Unterschied zum linearen Regressionsmodell wird im GLM der Erwartungswert $\mu = \mathbb{E}(\mathbf{Y})$ verknüpft durch die Linkfunktion

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

und die Responsevariable unterliegt einer Verteilung aus der Exponentialfamilie, d.h.

$$Y_i \sim f_{\theta_i}(y_i) \ ,$$

mit dem natürlichen Parameter θ_i.

4.1.1 Parameterschätzung

In Generalisierten Linearen Modellen können die Parameter $\boldsymbol{\beta}$ mittels Maximum Likelihood geschätzt werden. Für die i.i.d. Zufallsvariablen Y_i ergibt sich die Likelihood für den Parametervektor $\boldsymbol{\beta}$ zu

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i) \ . \tag{4.3}$$

Für den allgemeinen Fall lautet die Log-Likelihood zum Parametervektor $\boldsymbol{\beta}$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)] = \sum_{i=1}^{n} \frac{y_i\theta_i - b_i(\theta_i)}{\phi} \ \omega_i + c_i(y_i, \phi_i) \ . \tag{4.4}$$

Die Maximierung der Likelihood erfolgt über das Gleichungssystem

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \ \forall j \ , \tag{4.5}$$

mit den Gewichtungen $V(\mu_i)$, vergleiche Wood (2006). In der Regel können diese aus der Scorefunktion resultierenden Gleichungssysteme nicht analytisch gelöst werden. Die Lösung erfolgt numerisch mithilfe von Fisher Scoring oder des Newton-Raphson Ansatzes (siehe z.B. Fahrmeir, Kneib und Lang, 2009).
Vergleichbar zu den Residuenquadratsummen im linearen Modell ist in GLMs die Devianz. Sie dient zum Beispiel zur Analyse der Modellgüte und berechnet sich zu

$$D = 2[\ell(\hat{\boldsymbol{\beta}}_{\max}) - \ell(\hat{\boldsymbol{\beta}})] \ \phi \tag{4.6}$$

wobei $\ell(\hat{\boldsymbol{\beta}}_{\max})$ die Log-Likelihood des saturierten Modells ist (Wood, 2006).

4.1.2 Variablenselektion

In der Regel ist es notwendig, aus einer Menge von in Frage kommender Kovariablen diejenigen auszuwählen, die für das Modell relevant sind. Eine Vorauswahl der Kovariablen sollte anhand fachlicher Gesichtspunkte getroffen werden. Andererseits könnten im Modell auch Einflüsse eine Rolle spielen, die nicht erfasst werden können, weil z. B. keine Daten vorliegen. Diese kommen im späteren Modell als Störgrößen zur Geltung.
Eine Möglichkeit der Modellvalidierung bietet der Vergleich über die Likelihood. Problematisch ist, dass die Likelihood von der Anzahl der Parameter und auch der Anzahl an Beobachtungen abhängt. Das hat zur Folge, dass die Likelihood mit der Aufnahme zusätzlicher Kovariablen meist steigt und zwar sogar dann, wenn die betreffenden Einflussgrößen für das Modell völlig irrelevant sind (Overfitting). Abhilfe schafft in diesem Fall das Akaike Informationskriterium (AIC, Schwarz, 1978). Im AIC wird eine Bestrafung der Anzahl an Parametern vorgenommen und die Likelihood entsprechend angepasst. Das AIC ist definiert als

$$\text{AIC} \ = -2\ell(\hat{\boldsymbol{\theta}}) + 2p \ , \tag{4.7}$$

wobei $\hat{\boldsymbol{\theta}}$ der Parametervektor der ML-Schätzer und p die Anzahl der Parameter ist. Der hintere Teil bestraft ein überparametrisiertes Modell. Eine andere Gefahr besteht darin, dass im Modell relevante Variablen fehlen (Underfitting). Das AIC soll eine optimale Balance zwischen Over- und Underfitting gewährleisten. Da die Log-Likelihood mit negativem Vorzeichen in das Kriterium aufgenommen wird, wählt man das Modell, welches das AIC minimiert.
Ein weiteres Modellwahl-Kriterium ist das Bayesian Information Criterion (BIC, Akaike, 1973). Es ist definiert als

$$\text{BIC} \ = -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)p \ , \tag{4.8}$$

wobei n die Anzahl der Beobachtungen im Modell ist. Im BIC wird also neben der Parameterzahl auch der Datenumfang berücksichtigt und es kommt zu einem anderen Bestrafungsterm. Zur Variablenselektion wird wiederum das Modell mit minimalem BIC gewählt. Formal unterscheiden sich AIC und BIC nur wenig voneinander. Bei der Variablenselektion mit dem BIC werden vergleichsweise sparsame Modelle bevorzugt. Approximativ wird das Modell mit der größten Posteriori-Wahrscheinlichkeit ausgewählt (Fahrmeir, Kneib und Lang, 2009).
Es gibt verschiedene Methoden um eine Variablenselektion nach AIC/BIC in der Praxis durchzuführen. Heuristische Verfahren sind zum Beispiel Forward-, Backward- und Stepwise-Selection. Bei der Forward-Selection startet man mit dem Intercept-Modell und nimmt bei jeder Iteration diejenige Kovariable in das Modell auf, die die größte Verbesserung des AIC/BIC liefert. Kann das AIC/BIC durch Aufnahme einer weiteren Variable in das Modell nicht mehr reduziert werden, so bricht der Algorithmus ab und man wählt das letzte BIC-minimale Modell. Bei der Backward-Selection werden ausgehend vom vollen Modell mit allen zur Verfügung stehenden Kovariablen sukzessive Einflussgrößen entfernt. Dies geschieht wiederum, bis keine Verbesserung des AIC/BIC mehr erzielt werden kann. Die Stepwise-Selektion kombiniert Forward- und Backward-Selection. In jedem Schritt kann eine Variable entfernt oder hinzugefügt werden (vergleiche Fahrmeir, Kneib und Lang, 2009).

4.1.3 Modelldiagnose

Wie im linearen Modell werden bei GLMs Residuen zur Modelldiagnose verwendet. Meist werden hier nicht standardisierte Residuen, sondern Pearson- oder Devianzresiduen betrachtet. Da die Pearson-Residuen in praktischen Anwendungen oftmals asymmetrisch sind, empfiehlt Wood (2006) die Verwendung von Devianzresiduen. Die gesamte Devianz ergibt sich als Summe der Devianzen d_i aller n Beobachtungen

$$D = \sum_{i=1}^{n} d_i \ .$$

Die Devianzresiduen sind dann definiert als

$$\hat{\epsilon}_i^d = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i} \ . \tag{4.9}$$

Zur Überprüfung des Regressionsmodells werden zum Beispiel die Residuen gegen die gefitteten Werte $\hat{\mu}_i$ geplottet. Liegt ein korrektes Modell vor, dann weisen die Devianzresiduen keine Systematik auf, sondern streuen mit konstanter Varianz um die Null. Die Residuen verhalten sich bei sehr gutem Fit annähernd wie standardnormalverteilte Zufallsvariablen (siehe z.B. Wood, 2006).

4.1.4 Poisson-Regression

Verteilungsannahme

Zähldaten werden in der Regel anhand der Poisson-Verteilung modelliert, sprich $Y_i \sim Po(\lambda_i)$. Die Poisson-Verteilung ist diskret mit dem Träger $\mathcal{X} = \mathbb{N}_0$. Ihre Wahrscheinlichkeitsfunktion ergibt sich zu

$$P(Y = y) = \begin{cases} \frac{\lambda^y}{y!}\exp(-\lambda) & y > 0 \\ 0 & \text{sonst.} \end{cases} \tag{4.10}$$

Diese lässt sich als Exponentialfamilie darstellen. Durch Logarithmieren von (4.10) ergibt sich

$$\log\big(f(y|\lambda)\big) = y\log(\lambda) - \lambda - \log(y!) \tag{4.11}$$

und mit $\theta = \log(\lambda)$ als natürlichen Parameter folgt

$$\log\big(f(y|\theta)\big) = y\theta - \exp(\theta) - \log(y!) \tag{4.12}$$

mit den Parametern $b(\theta) = \exp(\theta) = \lambda$, $\phi = 1$ und $c(y, \phi) = -\log(y!)$, vergleiche Kapitel 4.1. Durch Ableiten von $b(\theta)$ erhält man $\mu = \mathbb{E}(y) = \text{Var}(y) = \lambda$, den „gedächtnislosen“ Parameter der Poisson-Verteilung, der auch als Intensität bezeichnet wird.

Strukturannahme

Der Log-Link $\eta_i = \log(\mu_i)$ bildet die natürlichen Linkfunktion der Poisson-Verteilung. Mit der dazugehörigen Responsefunktion $\mu_i = \exp(\eta_i)$ ergibt sich der bedingte Erwartungswert zu

$$\mu_i = \mathbb{E}(y_i|\boldsymbol{x}_i) = \lambda_i = \exp(\eta_i) \ . \tag{4.13}$$

Der Log-Link gewährleistet aufgrund der Eigenschaften der Exponentialfunktion eine einfache Interpretation der Effekte auf den Response. Ein Parameterwert β_j im linearen Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ führt zu einer multiplikativen Senkung/Steigerung des Response um den Faktor $\exp(\beta_j x_{ij})$. Gegeben nichtnegativer Kovariablenwerte kommt es bei positiven Werten β_j zum Anstieg des Erwartungswerts, vice versa. Das Vorzeichen der Parameter entspricht der Richtung des Effekts auf die erwartete Zielgröße. Die Parameterschätzer $\hat{\beta}_j$ lassen sich also ohne größere Transformationen interpretieren. Durch den Log-Link wird außerdem die Nichtnegativität der Erwartungswerte sicher gestellt.

Parameterschätzung

Die Poisson-Verteilung wird durch den Parameter λ charakterisiert, der als Intensität bezeichnet wird und sowohl Erwartungswert als auch Varianz der Poisson-verteilten Zufallsvariable ist. Im log-linearen Poisson-Modell

$$Y_i \sim Po(\lambda_i)$$

gilt für den Parameter

$$\begin{aligned} \lambda_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \ , \\ \log(\lambda_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i \ . \end{aligned}$$

Für die Likelihood-Inferenz wird die Dichte der Poisson-Verteilung herangezogen. Dann ergibt sich für i.i.d.-Zufallsvariablen y_i die Likelihood zu

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \ \exp(-\lambda_i) \ . \tag{4.14}$$

Die Abhängigkeit von $\boldsymbol{\beta}$ ergibt sich mit $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Die Log-Likelihood ist dann

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log(\lambda_i) - \lambda_i) \ , \tag{4.15}$$

wobei die nicht von $\boldsymbol{\beta}$ abhängige Konstante vernachlässigt wird. Mit dem Zusammenhang $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ erhält man

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) = \sum_{i=1}^{n} (y_i\eta_i - \exp(\eta_i)) \tag{4.16}$$

und die Scorefunktion

$$\mathbf{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i\big(y_i - \exp(\eta_i)\big) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \lambda_i) \ , \tag{4.17}$$

sowie die Fisher-Information

$$\mathbf{F}(\boldsymbol{\beta}) = \mathbb{E}\big(\mathbf{s}(\boldsymbol{\beta})\mathbf{s}'(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\lambda_i \ , \tag{4.18}$$

wobei $\mathbb{E}(y_i - \lambda_i)^2 = \lambda_i$, vergleiche Fahrmeir, Kneib und Lang (2009).
Der ML-Schätzer kann nicht analytisch berechnet werden, es kommen numerische Verfahren wie Fisher-Scoring oder Newton-Raphson-Verfahren zur Anwendung (mehr dazu z.B. in Fahrmeir, Kneib und Lang, 2009, Anhang B).
Außerdem besteht zwischen Poisson-Verteilung und Exponentialverteilung ein spezieller Zusammenhang. Ist die Anzahl von Ereignissen innerhalb eines bestimmten Zeitraums Poisson-verteilt mit Parameter λ, dann ist die Wartezeit zwischen zwei Ereignissen exponentialverteilt mit Parameter λ.

4.1.5 Quasi-Poisson/Quasi-Likelihood Schätzung

Bei Zähldaten liegt häufig Überdispersion (Overdispersion) vor. Das ist der Fall, wenn die Variation in den Daten größer ist als im zugrunde liegenden Modell. Im Poisson-Modell wird die Varianz zu

$$\mathrm{Var}(y_i) = \mathbb{E}(y_i) = \lambda_i \tag{4.19}$$

berechnet. Overdispersion kann unter anderem an Ausreißern in den Daten liegen. Die Poisson-Verteilung kann diese Tails nicht modellieren. Abbildung 4.1 zeigt die Wahrscheinlichkeitsfunktion für eine Poisson-Verteilung mit $\lambda = 5$. In diesem Fall geht die Wahrscheinlichkeit für Werte größer als 17 bereits gegen Null. Bei empirischen Daten liegen jedoch häufig Ausprägungen in extremen Bereichen vor. Daher

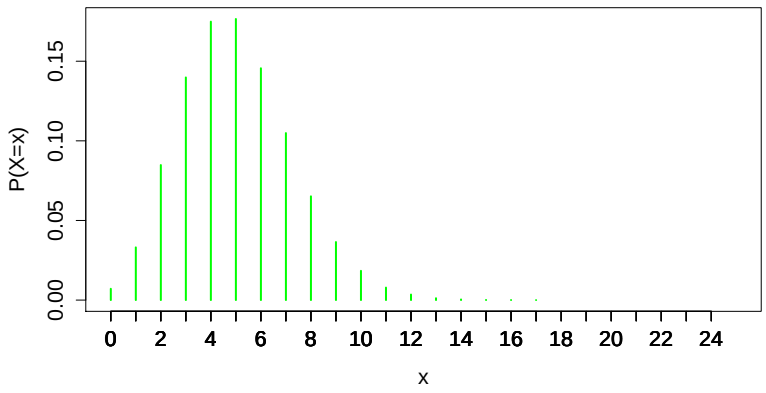

Abbildung 4.1: Poisson-Wahrscheinlichkeitsfunktion für Intensität $\lambda = 5$

ist es notwendig, die Modell-Varianz entsprechend anzupassen. Statt des normalen Poisson-Ansatzes wird ein Quasi-Poisson-Modell verwendet (siehe McCullagh und Nelder, 1989). Hier wird ein Dispersionsparameter $\phi \neq 1$ eingeführt. Die bedingte Varianz ergibt sich dann zu

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi\ \lambda_i\ .$$

Bei vorliegender Überdispersion ist $\phi > 1$ und die bedingte Varianz steigt schneller als der Erwartungswert (Fahrmeir, Kneib und Lang, 2009). Der Dispersionsparameter kann aus den Daten geschätzt werden mit

$$\hat{\phi}_D = \frac{1}{n-p} D\ . \tag{4.20}$$

Die Residualdevianz D wird also durch die Anzahl der Freiheitsgrade der Residuen geteilt. Liegt keine Überdispersion vor, dann nähert sich der Quotient dem Wert 1 an.
Die Modifikation der Varianz hat Implikationen für die Schätzung der Parameter. Die Schätzung der Parameter erfolgt im Quasi-Poisson-Ansatz nicht mehr durch den normalen ML-Schätzer, sondern über die Quasi-Likelihood. Die Quasi-Likelihood kommt im Allgemeinen zum Einsatz, wenn die Verteilung nicht genau spezifiziert werden kann. Es reicht hier bereits aus, das Verhältnis zwischen Mittelwert und Varianz bestimmen zu können. Im Folgenden wird der allgemeine Ansatz für GLMs ausgeführt (vergleiche Wood, 2006). Für eine Beobachtung y_i einer Zufallsvariable ergibt sich die Log-Quasi-Likelihood als

$$q_i(\mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - z}{\phi V(z)} \, dz \ , \tag{4.21}$$

wobei μ_i der Erwartungswert und $V(\mu_i)$ die Varianzfunktion der Zufallsvariable ist. Im Quasi-Likelihood-Ansatz wird davon ausgegangen, dass der Dispersionsparameter ϕ unbekannt ist. Die Log-Quasi-Likelihood der Stichprobe ist dann

$$q(\boldsymbol{\mu}) = \sum_{i=1}^{n} q_i(\mu_i) \ . \tag{4.22}$$

In ihren Eigenschaften ist $q(\boldsymbol{\mu})$ der Log-Likelihood ähnlich. Die Quasi-Likelihood lockert die Verteilungsannahme auf und ermöglicht es, Modelle anhand der Beziehung zwischen Mittelwert und Varianz zu fitten. Besonders bei Zähldaten sind Verteilungsannahmen häufig zu restriktiv, weil in der Empirie Überdispersion vorliegt. Die Parameterschätzer ergeben sich als Lösung des Gleichungssystems

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \ \forall \ j \ ,$$

und lassen sich wiederum numerisch bestimmen. Für den Quasi-Likelihood-Ansatz berechnet sich die Devianz zu

$$D_q = -2q(\hat{\boldsymbol{\mu}})\phi \ .$$

4.1.6 Gamma-Regression

Die Gamma-Verteilung ist eine häufig verwendete Verteilung für asymmetrische, nicht-negative und stetige Merkmale. Ihre Dichte ist gegeben durch

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} \exp(-\beta x), \quad x > 0 \ ,$$

wobei α der Shape-Parameter und β der inverse Skalen-Parameter ist. Erwartungswert und Varianz sind definiert als

$$\mathbb{E}(X) = \frac{\alpha}{\beta} \text{ bzw. } \mathrm{Var}(X) = \frac{\alpha}{\beta^2} \ .$$

Eine Zufallsvariable $X \sim G(\alpha, \beta)$ heißt gammaverteilt mit den Parametern α und β. Die Modellierung der Einflüsse auf gamma-verteilte Zielgrößen erfolgt mittels Gamma-Regression. Der natürliche Link zur Gammafunktion ist der inverse Link

$$\eta_i = \frac{1}{\mu_i} \ .$$

Für eine bessere Interpretierbarkeit der Effekte greift man jedoch in der Regel auf den Log-Link $\eta_i = \log(\mu_i)$ zurück (Fahrmeir, Kneib und Lang, 2009; Fox, 2008).

4.1.7 Inverse Normalverteilung

Eine weitere Exponentialfamilie, die im Kontext von GLMs verwendet werden kann, ist die Inverse Normalverteilung. Sie eignet sich, ähnlich wie die Gammaverteilung, zur Modellierung nicht-negativer, stetiger Merkmale, die rechtschief verteilt sind (Madsen und Thyregod, 2011). Die Inverse Normalverteilung hat die Dichte

$$f(x) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\} \frac{1}{x^{3/2}} \ , \quad \text{für } x > 0 \ ,$$

mit den Parametern $\mu, \lambda > 0$. Eine Zufallsvariable $X \sim IG(\mu, \lambda)$ heißt inversnormalverteilt. Der Erwartungswert und die Varianz ergeben sich zu $\mathbb{E}(X) = \mu$ bzw. $\text{Var}(X) = \mu^3/\lambda = \phi\mu^3$. Die Varianz steigt also mit dem Mittelwert. Für höhere Werte μ nimmt die Schiefe zu, für höhere Werte λ nimmt sie ab. μ ist der Skalenparameter der Verteilung (Fox, 2008).
Auch bei der Inversen Normalverteilung empfiehlt sich die Verwendung des Log-Links. Der kanonische Link der inversen Normalverteilung lautet

$$\eta_i = \frac{1}{\mu_i^2} \ ,$$

was eine Interpretation der Parameterschätzer erschwert.

4.2 Generalisierte Additive Modelle

Hastie und Tibshirani (1990) haben das Generalized Linear Model aus Kapitel 4.1 zum sogenannten Generalized Additive Model (GAM) erweitert. Ein GAM ist ein GLM, in dem die Zielgröße von glatten Funktionen metrischer Kovariablen abhängt. Die Struktur des GLMs erweitert sich dann zu

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \epsilon_i \ , \tag{4.23}$$

wobei $g(.)$ wieder die Linkfunktion ist. Dabei werden im semiparametrischen Ansatz k Kovariablen linear und q metrische Einflussgrößen nicht-parametrisch einbezogen. Es gilt $\mu_i = \mathbb{E}(y_i)$ und Y sei beliebig nach einer Exponentialfamilie verteilt. Der erwartete Response hängt im GAM von einer oder mehreren glatten Funktion $f_1, \ldots, f_q$ ab. Bei $z_1, \ldots, z_q$ handelt es sich um metrische Kovariablen, die den Response vermutlich nicht-linear beeinflussen und flexibel modelliert werden sollen. Diese Funktionen bleiben zunächst unspezifiziert und werden ebenfalls aus den

Daten geschätzt. Hinreichende Glattheit der Funktionen wird unterstellt, d.h. die Funktionen dürfen keine Sprungstellen enthalten. Zu Identifikationszwecken sind die glatten Funktionen stets um die Null zentriert. Für die Glättung der Funktionen steht ein großes Instrumentarium bereit. In dieser Arbeit soll lediglich ein kurzer Abriss über gängige Methoden gegeben werden. Häufig wird Regression unter Einbezug von glatten Funktionen auch als nicht-parametrische Regression bezeichnet.

4.2.1 Glättung

Das Fitten flexibler Funktionen lockert zwar die Restriktionen vorangegangener Regressionsmodelle auf, dennoch wird der Anwender vor neue Herausforderungen gestellt. Man gehe von einem univariaten Glättungsproblem aus. Abbildung 4.2 beschreibt den Scatterplot eines funktionalen Zusammenhangs zweier metrischer Variablen X und Y, der als glatte Funktion dargestellt werden soll. Dabei ist der wahre funktionale Zusammenhang, welcher hier durch die Linie dargestellt wird, unbekannt. Das univariate Glättungsproblem ergibt sich zu

$$y_i = f(x_i) + \epsilon_i \ ,$$

mit $\mathbb{E}(\epsilon_i) = 0$. Es stellt sich nun die Frage, wie die unspezifizierte Funktion f(.) aus den Daten (x_i, y_i) geschätzt werden kann. Zum einen muss eine Routine zur Glättung der Funktionen bestimmt werden, zum anderen muss man sich Gedanken darüber machen, wie glatt oder rau die resultierende Funktion sein darf.

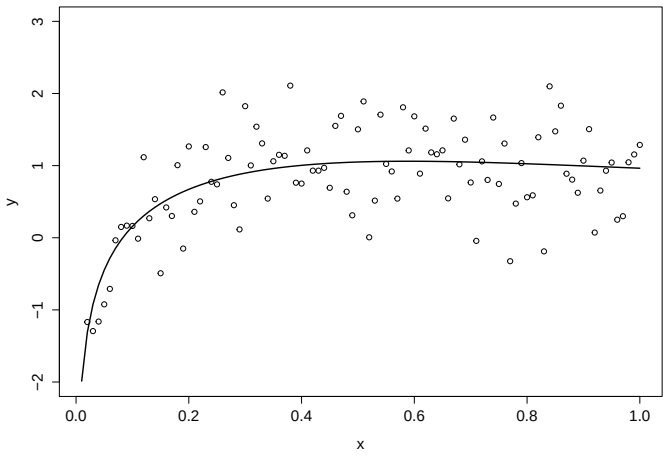

Abbildung 4.2: Scatterplot mit funktionalem Zusammenhang

Zur Glättung werden üblicherweise Regressions-Splines verwendet. Die Funktion wird dann durch stückweise Polynome repräsentiert, welche an den sog. Knoten zu einer stetigen und zweifach differenzierbaren Funktion verbunden werden. Die Glätter lassen sich somit mittels multipler Regression unter Verwendung von Basisvektoren berechnen. Regressions-Splines werden vor allem wegen ihrer computationalen Vorzüge zur Glättung verwendet. Eine glatte Funktion $f(.)$ lässt sich dann

mittels Basisfunktionen darstellen als

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j \ , \tag{4.24}$$

wobei $b_j(x)$ die bekannten Basisfunktionen und β_j unbekannte Parameter für $j = 1, \ldots, q$ sind. Dabei ist q die Basisdimension. Es gibt eine Vielzahl möglicher Basisfunktionen, die im Kontext von Regressionsplines Verwendung finden.
Durch die Wahl der Anzahl und Lage der Knoten kann die Anpassung des Glätters an die Daten gesteuert werden. Diese Wahl sollte in der Regel automatisiert erfolgen. Dies kann zum Beispiel mittels Kreuzvalidierung geschehen. Es besteht auch die Möglichkeit die Anzahl der Knoten unverändert zu lassen, ihren Einfluss auf die geglättete Funktion jedoch zu begrenzen. Dieses Vorgehen wird in den sog. Penalisierungsansätzen angewandt.

Penalisierungsansätze

In Penalisierungsansätzen wird die Rauheit der glatten Funktionen bestraft. Das Optimierungsproblem

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \ \rightarrow \min!$$

erweitert sich dann zu

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda \int_0^1 [f''(x)]^2 dx \ .$$

Der hintere Teil bestraft zu raue Funktionen. Dabei wird λ als Glättungsparameter bezeichnet. Der Tradeoff zwischen Glattheit und Datentreue kann über die Wahl von λ gesteuert werden. Für $\lambda \rightarrow 0$ ergibt sich eine raue Funktion und damit hohe Datentreue. Ein hohes λ liefert eine glatte Funktion mit geringer Anpassung an die Daten (vergleiche Graphik 4.3). Für penalisierte Regressionssplines wird dann

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}' \boldsymbol{S} \boldsymbol{\beta}$$

über $\boldsymbol{\beta}$ minimiert, wobei $\boldsymbol{S}$ eine Matrix mit bekannten Koeffizienten ist. Als Lösung ergibt sich

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X}' \boldsymbol{y} \ ,$$

vergleiche auch Ruppert, Wand und Carroll (2003) oder Wood (2006).

Trunkierte Potenzen

Für die Darstellung von Splines bestehen verschiedene Möglichkeiten. Eine in der Praxis häufig verwendete Variante sind die trunkierten Potenzen. Im Allgemeinen lassen sich Polynom-Splines vom Grad p als

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{p} \beta_{pk} (x - \kappa_k)_+^p$$

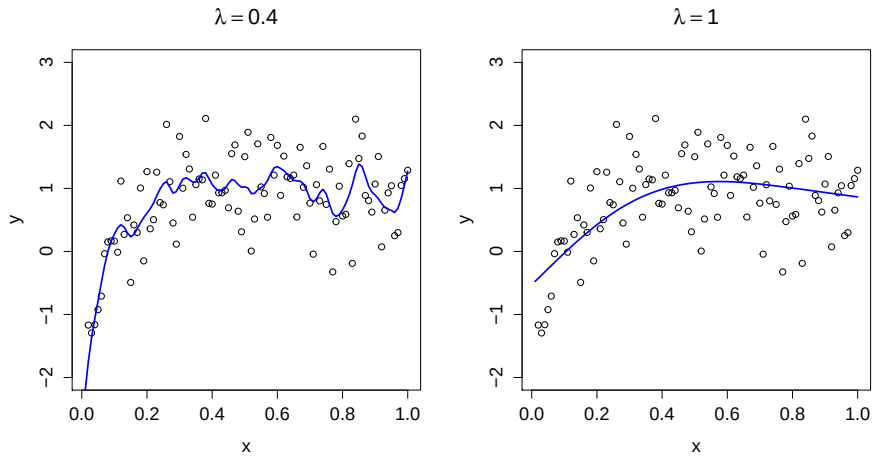

Abbildung 4.3: Scatterplot-Glätter für $\lambda = 0.4$ und $\lambda = 1$

schreiben (Ruppert, Wand und Carroll, 2003), mit den Knoten κ_k, wobei

$$(x - \kappa_k)_+^p = \begin{cases} (x - \kappa_k)_+^p & x \geq \kappa_k \\ 0 & \text{sonst} \ . \end{cases}$$

Es handelt sich um eine Linearkombination von Basisfunktionen. Die Basis vom Grad p lautet für trunkierte Potenzen $1, x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_k)_+^p$.

B-Splines

Eine alternative Darstellung bieten B-Spline-Basen. Sie haben gegenüber trunkierter Potenzen numerische Vorteile und sind in ihrer Anwendung weit verbreitet. Eilers und Marx (1996) verwenden B-Spline-Basen mit Penalisierung für ihre P-Splines.
B-Splines werden aus $q+1$ Polynomstücken vom Grad q zusammengesetzt und sind an den Knoten $q-1$ mal stetig differenzierbar. Sie bilden eine lokale Basis und sind nur im Intervall über $q+2$ Knoten positiv, ansonsten Null. Außer in den Randbereichen überlappen die Basisfunktionen mit $2q$ benachbarten Basisfunktionen. Der Wertebereich von B-Spline-Basisfunktionen ist beschränkt. Die Funktion f(.) kann mithilfe von B-Splines als Linearkombination von $d = m + q + 1$ Basisfunktionen als

$$f(x) = \sum_{k=1}^{d} \beta_k B_k(x)$$

dargestellt werden, wobei m die Anzahl an inneren Knoten ist. Die B-Spline-Basisfunktionen werden rekursiv definiert über die Formel

$$B_k^q(x) = \frac{x - \kappa_k}{\kappa_{k+q} - \kappa_k} B_k^{q-1}(x) + \frac{\kappa_{k+q+1} - x}{\kappa_{k+q+1} - \kappa_{k+1}} B_{k+1}^{q-1}(x) \ ,$$

mit

$$B_k^{-1}(x) = \begin{cases} 1 & \kappa_k \leq x < \kappa_{k+1} \\ 0 & \text{sonst} \ . \end{cases}$$

Für äquidistanten Knoten, wie sie in den P-Splines verwendet werden, haben alle Basisfunktionen die gleiche Gestalt (siehe Abbildung 4.4). Die Ableitungen für die Basisfunktionen ergeben sich zu

$$\frac{\partial B_k^q(x)}{\partial x} = q \cdot \left(\frac{1}{\kappa_{k+q} - \kappa_k} B_k^{q-1}(x) - \frac{1}{\kappa_{k+q+1} - \kappa_{q+1}} B_{k+1}^{q-1}(x) \right) ,$$

als Differenz von benachbarten B-Splines (Fahrmeir, Kneib und Lang, 2009). Eilers und Marx (1996) verwenden in den P-Splines diese Differenzen zur Penalisierung.

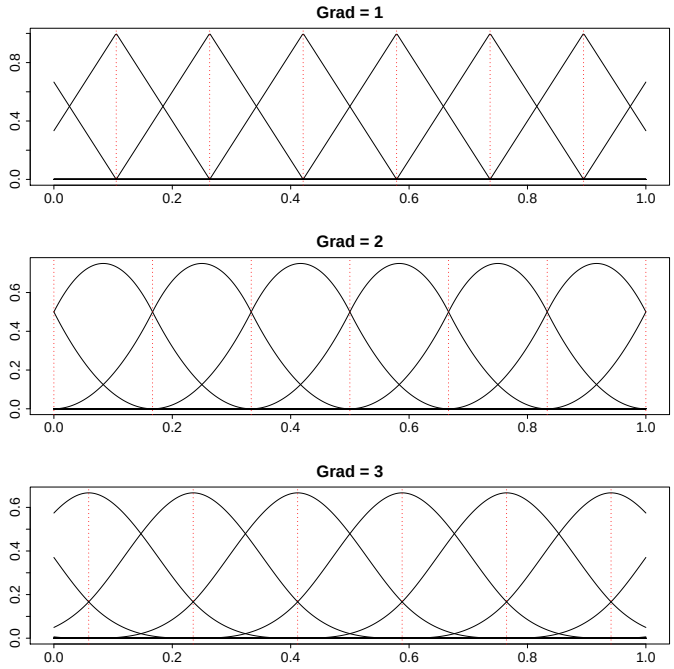

Abbildung 4.4: B-Spline-Basen für äquidistante Knoten

Kreuzvalidierung

Eine Herausforderung beim Glätten von Funktionen ist es, die perfekte Balance zwischen Glattheit und Datentreue zu finden. Der Grad der Glattheit kann mittels Kreuzvalidierung bestimmt werden. Das R-Package mgcv bedient sich für diese Zwecke dem Generalisierten Kreuzvalidierungs-Kriterium (Generalized Cross Validation, GCV). Für GAMs geht man unter Verwendung der Modelldevianz $D(\boldsymbol{\beta})$ von der Minimierung der Gleichung

$$D(\boldsymbol{\beta}) + \sum_{j=1}^{m} \lambda_j \boldsymbol{\beta}' \mathbf{S}_j \boldsymbol{\beta}$$

durch Ableiten nach $\boldsymbol{\beta}$ aus. Der Generalized Cross Validation Score ist dann definiert als

$$V_g = \frac{nD(\hat{\boldsymbol{\beta}})}{[n - \mathrm{tr}(\mathbf{H})]^2} , \tag{4.25}$$

mit der Prädiktionsmatrix **H** (Hat-Matrix), vergleiche auch Hastie und Tibshirani (1990) sowie Wood (2006).

4.2.2 Interaktionen

In linearen Regressionsmodellen können Interaktionseffekte einbezogen werden, die Wechselwirkungen zwischen Kovariablen modellieren. Dies ist auch für GLMs und GAMs möglich (Fahrmeir, Kneib und Lang, 2009). Denkbar ist zum Beispiel ein Effekt zwischen einer metrischen Kovariable und einer kategorialen Variable. Der Ansatz des GAM aus (4.23) erweitert sich zu

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{z_1|x_1}(z_{i1})\ x_{i1} + \epsilon_i\ , \tag{4.26}$$

wobei x_1 eine Faktorvariable sei. Zwischen dieser und der glatten Funktion der metrischen Variable z_1 besteht nun eine Interaktion. Es wird also eine glatte Funktion für jede Ausprägung der Faktorvariable x_1 erzeugt. Es ist auch ein Modell ohne Aufnahme der Haupteffekte denkbar. Dann ergibt sich der Response als

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_{z_1|x_1}(z_{i1})\ x_{i1} + f_2(z_{i2}) + \ldots + f_q(z_{iq}) + \epsilon_i\ , \tag{4.27}$$

wobei die Funktion $f_1(z_{i1})$ entfällt. Der Term $f_{z_1|x_1}(z_1) + \beta_1$ beschreibt dann den variierenden Effekt des Faktors auf den Response. Während die glatten Funktionen aus Ansatz (4.26) als Modifikation des glatten Haupteffekts zu sehen sind, wird in Ansatz (4.27) für jeden Faktor ein eigener Smooth erzeugt.

4.2.3 Generalized Additive Models in R

Die Statistik Software R (www.r-project.org) stellt verschiedene Packages zur Implementierung von GAMs zur Verfügung. In der vorliegenden Arbeit wurde in erster Linie das Package `mgcv` von Wood (2006) verwendet. Die Routine zum Fitten von GAMs ist der Befehl `gam()` bzw. `bam()` für große Datenumfänge (Big Data). Das Package kann als Erweiterung zu den linearen Modellen `lm()` und generalisierten linearen Modellen `glm()` verstanden werden. Viele Befehle und Tools gehen konform mit den vorhandenen Implementierungen. Zum Beispiel stehen dieselben Exponentialfamilien und Linkfunktionen wie in der Funktion `glm()` zur Verfügung.
Für die nicht-parametrische Regression sind im Package verschiedene Basisfunktionen implementiert, z.B. Thin Plate, kubische bzw. penalisierte Splines, sowie Tensorprodukt-Basen. Die Schätzung des Modells basiert im Allgemeinen auf einem penalisierten Regressions-Spline-Ansatz mit automatisierter Wahl des Glättungsparameters. Die Wahl der Glattheit erfolgt zum Beispiel mittels GCV, Unbiased Risk Estimator (UBRE) oder Restricted Maximum Likelihood (REML). Die Schätzung des Modells erfolgt über Maximierung der penalisierten Likelihood oder Quasi-Likelihood.
Ein weiteres Package für GAMs ist `gamlss` von Rigby und Stasinopoulos (2005). Darin wird die Verteilungsannahme der Zielgrößen nach Exponentialfamilien aufgeweicht und ein größeres Instrumentarium an Verteilungen zur Verfügung gestellt. Hier sind unter anderem auch Zero-inflated Models (vergleiche Kapitel 4.4) implementiert.
Die Funktion `gamlss()` fittet Modelle mittels (penalisierter) ML-Schätzung, verwendet allerdings andere Algorithmen wie `gam()` aus dem Package `mgcv`.

4.3 Mixed Models

In klassischen linearen Regressionsmodellen und GLMs (Fixed Effect Models) wird davon ausgegangen, dass die Zielgrößen y_i unabhängig sind. Bei Vorliegen von Longitudinal- oder Clusterdaten ist diese Annahme in der Regel verletzt. Werden an verschiedenen Individuen wiederholte Messungen/Zählungen durchgeführt, so ist davon auszugehen, dass die resultierenden Zielgrößen eines Individuums oder Clusters korreliert sind.
Zur Modellierung solcher Clusterdaten wurden die sog. Mixed Models eingeführt. In diesen werden neben den Fixed Effects auch zufällige, individuenspezifische Effekte (Random Effects) modelliert. Praktische Ansätze werden in Pinheiro und Bates (2000) vorgestellt. Außerdem sind für GAMs Random Effects im Package `mgcv` von Wood (2006) implementiert.
Bei Clusterdaten liegen die Zielgrößen y_{ij} für Cluster i (mit $i = 1, \ldots, m$) und die j-te Messwiederholung ($j = 1, \ldots, n_i$) bei n_i Messwiederholungen pro Cluster vor. Dabei können auch einzelne Individuen ein Cluster bilden. Es wäre denkbar, die individuenspezifischen Effekte als Fixed Effects aufzunehmen. Allerdings ist man in der Regel nicht an den Schätzern für bestimmte Individuen, sondern an allgemeingültigen Ergebnissen interessiert. Durch Aufnahme anderer Individuen in die Stichprobe würden andere Schätzer für die Fixed Effects errechnet werden. Aus diesem Grund werden in Mixed Models die individuenspezifischen Effekte als Random Effects geschätzt. Der übliche lineare Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ erweitert sich für korrelierte Clusterdaten zu

$$\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_i$$

wobei $\boldsymbol{u}_{ij}$ meist ein Teilvektor der Kovariablen ist und γ_i die individuenspezifischen Effekte beschreibt. In der Regel geht man von normalverteilten Random Effects $\gamma_i \sim N(0, \boldsymbol{D})$ mit der Kovarianzmatrix $\boldsymbol{D}$ aus. Bei einfaktoriellen Random Effects gilt $\boldsymbol{u}_{ij} = \mathbf{1}$ und γ_{0i} bezeichnet die individuenspezifische Abweichung vom Populationsmittelwert. Weiterhin gilt dann $\gamma_{0i} \sim N(0, \tau_0^2)$, also die Normalverteilung dieser zufälligen Effekte. Für das Poisson-Modell ergibt sich dann

$$\lambda_i = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \gamma_{0i}) = \exp(\gamma_{0i}) \cdot \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \ ,$$

siehe auch Fahrmeir, Kneib und Lang (2009).

4.4 Zero-inflated Poisson Regression

Eine weitere Herausforderung bei der Modellierung von Zählprozessen ist der Umgang mit einer Vielzahl von Nullen (Excess Zeros). Diese Situation tritt ein, wenn sich im Datensatz Individuen/Einheiten befinden, die sehr wenige Counts erzeugen oder gar nicht unter Risiko für einen Count stehen. Wäre bekannt, welche Individuen nicht unter Risiko stehen, könnten diese Fälle einfach ausgeschlossen werden. In der Regel sind diese Informationen jedoch latent. Zum Umgang mit Excess Zeros schlägt Lambert (1992) das Zero-inflated Poisson Model (ZIP) vor. Darin wird unterstellt, dass mit Wahrscheinlichkeit p für die Zielgröße nur eine 0 beobachtet werden kann und die Zielgröße mit Wahrscheinlichkeit $1-p$ einer Poisson-Verteilung $Po(\lambda_i)$ folgt.

Damit folgt für unabhängige Y_i die Mischverteilung

$$\begin{aligned}\mathbb{P}(Y_i = 0) &= p_i + (1 - p_i)\exp(-\lambda_i) \\ \mathbb{P}(Y_i = k) &= (1 - p_i)\exp(-\lambda_i)\ \frac{\lambda_i^k}{k!}\ , \quad k \in \mathbb{N}^+\ .\end{aligned}$$

Die Zugehörigkeit zur latenten Klasse der Individuen mit Zero Counts hängt wie λ_i von einem Prädiktor ab. Für die Wahrscheinlichkeit p_i ergibt sich mit dem Logit-Link

$$\log \frac{p_i}{1 - p_i} = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_q z_{iq}\ .$$

Der Response μ_i (bzw. λ_i) wird mittels Log-Link mit den Kovariablen verknüpft. Die Wahrscheinlichkeit p_i kann von den gleichen Regressoren wie λ_i abhängen. Für das Zero-inflated Poisson Model gilt

$$\begin{aligned}\mathbb{E}(Y_i) &= (1 - p_i)\mu_i \\ \mathrm{Var}(Y_i) &= (1 - p_i)\mu_i(1 + p_i\mu_i)\ .\end{aligned}$$

Wie beim Quasi-Poisson-Ansatz steigt für $p_i > 0$ die Varianz schneller als der Erwartungswert. Da nicht bekannt ist, welche Beobachtungen zur latenten Klasse gehören, wird die Schätzung des Modells komplizierter. Die Log-Likelihood lautet dann

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{y_i=0} \log\left(\exp(\boldsymbol{z}_i'\boldsymbol{\gamma}) + \exp\big(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\big)\right) \\ &+ \sum_{y_i>0} \left(y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \\ &- \sum_{i=1}^{n} \log\left(1 + \exp(\boldsymbol{z}_i'\boldsymbol{\gamma})\right) - \sum_{y_i>0} \log(y_i!)\ .\end{aligned}$$

4.5 Verweildaueranalyse

4.5.1 Grundlagen

Die Verweildaueranalyse (Lebensdaueranalyse) findet schwerpunktmäßig Anwendung in der Biostatistik. Hier wird die Lebensdauer von Individuen unter Beobachtung analysiert. Als Lebensdauer wird die Zeit vom Beginn der Beobachtung bis zum Eintreten eines interessierenden Ereignisses bezeichnet. Ein Individuum, bei dem das interessierende Ereignis eintreten kann, steht unter Risiko. In der Statistik-Software R sind im Package `survival` von Therneau (2013) die gängigsten Methoden der Verweildaueranalyse implementiert. Neben dem Einsatz in der Biostatistik lässt sich die Lebensdaueranalyse auch in vielen anderen Kontexten anwenden. In dieser Arbeit soll die Dauer zwischen Klick und Order mittels Verweildauern analysiert werden. Das interessierende Ereignis ist hier das Tätigen einer Order durch den Internet-Nutzer.
Eine Besonderheit bei der Analyse von Lebensdauern ist, dass in der Regel nicht alle Individuen über die komplette Dauer beobachtet werden können. In diesem Fall

kommt es zur Zensierung der Daten. Das ist zum Beispiel der Fall, wenn ein Individuum vorzeitig aus der betreffenden Studie ausscheidet. In diesem Fall liegt eine Rechtszensierung der Daten vor. Man muss zwischen wahren Lebensdauern T und Zensierungszeiten C unterscheiden. Bei der Rechtszensierung wählt man die kürzere der beiden, also

$$t_i^* = \min\{t_i, c_i\} \ .$$

Ob es sich um die wahre Lebensdauer oder um eine zensierte Beobachtung handelt, wird durch den Zensierungsindikator angezeigt. Dieser ist definiert als

$$\delta_i = I\{T \leq c_i\} \quad \Leftrightarrow \quad \delta_i = \begin{cases} 1 & , T^* = t_i \\ 0 & , T^* = c_i \end{cases}$$

Bei der Zensierung von Daten unterscheidet man grundsätzlich zwischen zwei Fällen. Bei TypI-Zensierungen wird die Beobachtung der Individuen nach einer festgelegten Dauer abgebrochen. Bei TypII-Zensierungen bricht die Beobachtung nach Erreichen einer festgelegten Anzahl von Ereignissen ab. Neben diesen beiden häufigen Zensierungsarten sind noch viele weitere denkbar. Ein in der Realität häufig auftretender Fall ist das sogenannte Random Censoring. Hier scheidet das Individuum aufgrund eines zufälligen Ereignisses aus der Studie aus.
Lebensdauern T stellen nicht-negative Zufallsvariablen dar. Man stellt daher gewisse Forderungen an ihre Verteilungen. Nicht alle Verteilungen sind geeignet, um Verweildauern zu modellieren. Häufig werden zum Beispiel Exponential-, Weibull- und log-logistische Verteilung verwendet.
Eine zentrale Größe in der Lebensdaueranalyse ist die sogenannte Hazardrate $h(t)$. Sie beschreibt das Riskio bzw. die Chance, dass für ein Individuum im nächsten Moment ein Ereignis eintritt, gegeben dass das Individuum den Zeitpunkt t überlebt hat. Sie ist definiert als

$$h(t) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \ .$$

Die kumulierte Hazardrate zum Zeitpunkt t ergibt sich zu

$$H(t) = \int_0^t h(u)du \ .$$

Im Gegensatz zur Verteilungsfunktion $F(t) = \mathbb{P}(T \leq t)$ ist man bei der Lebensdaueranalyse verstärkt daran interessiert, wieviele Individuen einen Zeitpunkt t überlebt haben. Diesen Zusammenhang drückt die Survivalfunktion $S(t)$ aus, mit

$$S(t) = \mathbb{P}(T \geq t) = 1 - F(t) \ .$$

Zwischen diesen Größen lassen sich einige Zusammenhänge herstellen. So lässt sich die Hazardfunktion als Quotient aus Dichtefunktion und Survivalfunktion

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

schreiben. Die Survivalfunktion lässt sich mithilfe der kumulierten Hazard ausdrücken als

$$S(t) = \exp\{-H(t)\} = \exp\Big\{-\int_0^t h(u)du\Big\} \ ,$$

siehe auch Kleinbaum und Klein (2005).

4.5.2 Nelson-Aalen- und Kaplan-Meier-Schätzer

Die kumulierte Hazard $H(t)$ und die Survivalfunktion $S(t)$ lassen sich nicht-parametrisch aus den Daten schätzen. Instrumente dafür sind der Nelson-Aalen-Schätzer für die kumulierte Hazard und der Kaplan-Meier-Schätzer für die Survivalfunktion (siehe Therneau und Grambsch, 2000).
Der Nelson-Aalen-Schätzer basiert auf der Theorie der Zählprozesse. Man stellt die Beziehung zwischen Individuen unter Beobachtung und der Anzahl an Ereignissen her. Für rechtszensierte Daten sei $Y_i(t) = I\{Y_i \geq t\}$ die Indikatorfunktion, ob Individuum i zum Zeitpunkt t unter Beobachtung steht, wobei für ein beobachtetes Individuum $Y_i = 1$ gilt. $N_i(t)$ sei die Anzahl der Ereignisse von i zum Zeitpunkt t. Aggregiert ergibt sich dann für alle n Individuen

$$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t) \text{ bzw. } \bar{N}(t) = \sum_{i=1}^{n} N_i(t) \ ,$$

sprich, die gesamte Anzahl von Individuen unter Beobachtung zum Zeitpunkt t bzw. die gesamte Anzahl an Ereignissen bis einschließlich t. Der Nelson-Aalen-Schätzer ist definiert als

$$\hat{H}(t) = \sum_{i:\ t_1 \leq t} \frac{\Delta \bar{N}(t_i)}{\bar{Y}(t_i)} \ , \tag{4.28}$$

wobei $\Delta \bar{N}_i(t)$ die Anzahl der Ereignisse zum exakten Zeitpunkt t ist. Der Nelson-Aalen-Schätzer schätzt die durchschnittliche Anzahl an Ereignissen im Zeitraum $(0, t]$ für die Individuen unter Risiko. Mit (4.28) ergibt sich für den Schätzer eine Treppenfunktion mit Sprungstellen zu den Ereigniszeitpunkten. Die Varianz des Schätzers ist dann

$$\hat{\sigma}^2_H(t) = \sum_{i:\ t_1 \leq t} \frac{\Delta \bar{N}(t_i)}{\bar{Y}^2(t_i)} \ .$$

Mit der Varianz des Schätzers lassen sich punktweise Konfidenzintervalle bestimmen. Diese berechnen sich zu

$$\hat{H}(t) \pm z_{1-\alpha/2}\ \hat{\sigma}_H(t) \ ,$$

mit den Normalverteilungsquantilen z. Bei der Lebensdaueranaylse ist man häufig daran interessiert, wieviele Individuen einen bestimmten Zeitpunkt überlebt haben, sprich der Survivalfunktion $S(t)$. Ein Schätzer für $S(t)$ ist der Kaplan-Meier-Schätzer. Die kumulierte Hazard und die Survivorfunktion hängen über $S(t) = \exp\{-H(t)\}$ zusammen. Der Kaplan-Meier-Schätzer ist definiert als

$$\hat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{\Delta \bar{N}(t_i)}{\bar{Y}(t_i)}\right) \ . \tag{4.29}$$

Auch hier ergibt sich eine Treppenfunktion mit Sprungstellen an den Ereigniszeitpunkten. Im Gegensatz zum Nelson-Aalen-Schätzer handelt es sich hier jedoch um eine fallende Funktion. Im Laufe der Zeit „sterben“ immer mehr Individuen, d.h. der Anteil der Überlebenden wird mit der Zeit kleiner. Mit der Kaplan-Meier-Kurve wird sozusagen die „Überlebenszeit“ modelliert. Die geschätzte Varianz des Kaplan-Meier-Schätzers beträgt

$$\widehat{\mathrm{Var}}\big(\hat{S}(t)\big) = \big(\hat{S}(t)\big)^2 \ \cdot \sum_{i:\ t_i \leq t} \frac{\Delta \bar{N}(t_i)}{\bar{Y}(t_i)[\bar{Y}(t_i) - \Delta \bar{N}(t_i)]}$$

Konfidenzbänder für den Schätzer berechnen sich dann zu

$$\hat{S}(t) \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}\big(\hat{S}(t)\big)} \ .$$

Durch gruppierte Kaplan-Meier-Kurven lassen sich Unterschiede bei der Überlebenszeit von Individuen verschiedener Gruppen veranschaulichen. Ein schnelleres Abfallen der Kurve zeigt eine kürzere Lebensdauer in der betreffenden Gruppe an.

4.5.3 Der Log-Rank-Test

Die Unterschiede in den Hazard-Rates von verschiedenen Gruppen lassen sich auf statistische Signifikanz überprüfen. Dies erfolgt üblicherweise mit dem Log-Rank-Test. Dieser testet die Nullhypothese, dass in allen Gruppen i dieselbe Hazardrate h_i vorliegt. Für zwei Gruppen ergibt sich die Nullhypothese

$$H_0 : \ h_1 = h_2 \ ,$$

vergleiche Kleinbaum und Klein (2005). Die Log-Rank-Statistik lautet

$$LR = \frac{(O_i - E_i)^2}{\mathrm{Var}(O_i - E_i)} \overset{H_0}{\sim} \chi^2(1)$$

mit $O_i - E_i = \sum_{t=1}(m_{it} - e_{it})$ für $i = 1, 2$, wobei m_{it} die beobachtete Anzahl an Ereignissen in Gruppe i zum Zeitpunkt t ist. Außerdem ist die erwartete Anzahl an Ereignissen für Gruppe 1 definiert als

$$e_{1t} = \left(\frac{n_{1t}}{n_{1t} + n_{2t}} \right) \cdot (m_{1t} + m_{2t}) \ .$$

Der Nenner der Log-Rank-Statistik ergibt sich zu

$$\mathrm{Var}(O_i - E_i) = \sum_t \frac{n_{1t} n_{2t} (m_{1t} + m_{2t})(n_{1t} + n_{2t} - m_{1t} - m_{2t})}{(n_{1t} - n_{2t})^2 (n_{1t} + n_{2t} - 1)}$$

Der Test lässt sich auch für G (> 2) Gruppen durchführen. Die Berechnung der Teststatistik wird dann weitaus aufwendiger (Kleinbaum und Klein, 2005).

4.6 Der χ^2-Anpassungstest

Mit dem χ^2-Anpassungstest lässt sich statistisch überprüfen, ob ein Merkmal einer bestimmten Verteilung folgt. Seien $h_1, \ldots, h_k$ die absoluten Häufigkeiten für die Beobachtung eines kategorialen Merkmals $X \in \{1, \ldots, k\}$ bei n unabhängigen Wiederholungen eines Zufallsexperiments. Nimmt man an, dass alle Ausprägungen mit derselben Wahrscheinlichkeit auftreten, dann lautet die Nullhypothese

$$H_0 : \ \mathbb{P}(X = i) = \pi_i = \frac{1}{k} \quad \text{für } i = 1, \ldots, k \ .$$

Die χ^2-Teststatistik ist definiert zu

$$\chi^2 = \sum_{i=1}^{k} \frac{(h_i - n\pi_i)^2}{n\pi_i} \overset{H_0}{\underset{a}{\sim}} \chi^2(k-1) \ ,$$

mit der beobachteten absoluten Häufigkeit h_i. Die Statistik ist für große n unter H_0 approximativ χ^2-verteilt mit $k - 1$ Freiheitsgraden, siehe Fahrmeir et al. (2010).

Kapitel 5

Statistische Modellierung

5.1 Modellierung der Conversion Rate (Impression-to-Order)

5.1.1 Datengrundlage und Modellannahmen

Der Erfolg einer Werbemittelschaltung bemisst sich an der Anzahl an Klicks auf das ausgelieferte Werbemittel sowie den daraus resultierenden Orders beim Advertiser. Da die Anzahlen von Klicks und Orders stark davon abhängig sind, wie oft das Werbemittel ausgeliefert wurde, sprich von der Anzahl der erzeugten Impressions, werden im Folgenden nicht Absolutwerte, sondern Conversion Rates mit statistischen Methoden modelliert.
Aufgrund der Tatsache, dass die Daten in verschiedenen Teildatensätzen vorlagen, mussten einige Anpassungen und Einschränkungen vorgenommen werden. Für das nachfolgende Modell wurden Daten aus dem Monat Februar 2013 betrachtet. Wie im aggregierten Datensatz wurde Publisher-seitig auf das Business Model „Media“ eingeschränkt, um die Impressions-Zahlen in das Modell einfließen lassen zu können. Es handelt sich dabei um Publisher mit Key Account Manager. Da im Business Model Media ein beträchtlicher Anteil an Orders mittels Postview getrackt wird, erschien es hier sinnvoll, die Rate Impression-to-Order (ITO) als Maß für den Werbeerfolg zu modellieren. Beim Postview-Tracking werden den Bestellungen die zugehörigen Publisher anhand der Impressions zugeordnet. Ein Klick wird in diesen Fällen nicht dokumentiert. Aus diesem Grund erweisen sich für Media-Publisher die Click-through Rate und die Conversion Rate (Click-to-Order) als weniger zuverlässig.
Für das Modell sind nur Advertiser relevant, die am betreffenden Tag Impressions generierten. Nur wenn ein Werbemittel tatsächlich Impressions beim Publisher erzeugte, kann dadurch eine Order verursacht werden. Um die Informationen aus den Teildatensätzen kombinieren zu können, wird die Annahme getroffen, dass die Zeitpunkte von Impression, Klick und Order zusammenfallen, d.h. dass Order, Klick und zugehörige Impression innerhalb der gleichen Tagesstunde erfolgten. Diese Annahme muss getroffen werden, da sich der Weg eines Users von der Impression beim Publisher zur Order nicht nachvollziehen lässt. Die Verweildaueranalyse in Kapitel 5.5 zeigt, dass zwischen Klick und Order regelmäßig weniger als eine Stunde vergeht. Damit ist die Annahme, dass Impression und Order innerhalb einer Stunde stattfin-

0 €	0-100 €	100-500 €	500-1000 €
-	-	4	2
1000-2000 €	2000-5000 €	5000-10000 €	>10000 €
6	14	4	2

Tabelle 5.1: Verteilung der Advertiser auf Umsatzklassen im GAM (ITO)

den, nicht unrealistisch.
Filtert man die Daten nach Media-Publishern, dann verbleiben 32 Advertiser im Datensatz. Bei diesen Advertisern werden nur Fälle berücksichtigt, in denen mindestens eine Impression pro Stunde registriert ist. Es ist bekannt, dass es im Verhältnis zu den Impressions nur sehr wenige Orders gibt. Darum wurden Impressions und Orders pro Advertiser über alle Media-Publisher aggregiert. Insgesamt stehen im Teildatensatz 47 Publisher zur Verfügung. Statt einzelner Partnerschaften werden also sämtliche Partnerschaften eines Advertisers mit Media-Publishern modelliert. Die Datenbasis für das Modell bilden rund 18500 stündliche Beobachtungen. Durch diese Verkleinerung des Datenumfangs ergeben sich Verwerfungen bei der Besetzung der Advertiser-Kategorien. Wie aus den Tabellen 5.1 und 5.2 ersichtlich, gibt es im Teildatensatz keine Advertiser mit weniger als 100 EUR Umsatz. Außerdem wird die Kategorie Home & Accessoires von einem einzigen Advertiser gebildet. Für die spätere Interpretation der Schätzparameter sollte man diese Eigenschaften der Datengrundlage im Hinterkopf behalten.

	Elektro	Fashion	Home&Acc.	Kinder	Sonstige	Vollvers.	
Nein	-	2	-	3	-	2	7
Ja	5	9	1	2	4	4	25
	5	11	1	5	4	6	32

Tabelle 5.2: Verteilung auf die Kategorien Advertiser Account Manager und Branche im GAM (ITO)

Für die Zielgröße Orders wird zunächst eine geeignete Verteilung spezifiziert, welche sich mittels Generalisierter Additiver Modelle modellieren lässt. Da es sich um Zähldaten handelt, bietet sich die Verwendung eines Poisson-Modells mit

$$z_i \sim Po(\lambda_i), \quad i = 1, \ldots, n \tag{5.1}$$

an, wobei λ_i die Rate (Intensität) des zugrundeliegenden Poisson-Prozesses bezeichnet. Unter Verwendung der natürlichen Linkfunktion, d.h. des Log-Links, ist der Erwartungswert der Zielgröße definiert als

$$\mathbb{E}(z_i) = \lambda_i = \exp(\eta_i) \ , \tag{5.2}$$

mit dem Prädiktor η_i, der im Folgenden ebenfalls noch genauer spezifiziert wird.
Da sowohl der Effekt der Tageszeit als auch des Monatsverlaufs untersucht werden soll, werden als Zielgröße die Orders pro Stunde pro Advertiser mit allen Media-Publishern betrachtet. Die Zielgröße z_{it} wird definiert als die Anzahl der Orders bei Advertiser i zu Zeitpunkt t, welche durch Impressions von Werbemitteln auf den

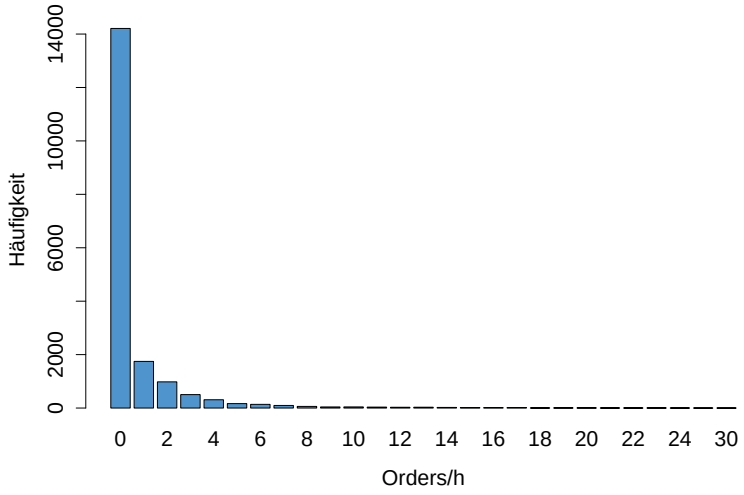

Abbildung 5.1: Verteilung z_{it} (Orders pro Stunde pro Advertiser)

Websites der Media-Publisher generiert wurden. Der Zeitpunkt t zerfällt in Datum und Uhrzeit.
Bei bloßer Betrachtung von Orders bleibt jedoch unbeachtet, dass bei den Impressions der Publisher ganz unterschiedliche Größenordnungen vorliegen. Die Anzahl der Klicks und Orders hängt wohl vor allem davon ab, wieviele Impressions generiert wurden. Die stundengenaue Berücksichtigung der Impressions stellt sich aufgrund der Aggregierung der Daten als problematisch dar. Die stundengenauen Order-Daten müssen mit tageweise aggregierten Impressions-Zahlen gematcht werden. Daher wird hier eine vereinfachende Annahme getroffen. Es wird eine Gleichverteilung der Impressions im Tagesverlauf unterstellt. Die stündlichen Impressions am Tag d ergeben sich dann zu

$$V(\text{Tag} = d) = \frac{V_d}{24} \ . \tag{5.3}$$

Es wären auch andere Ansätze für die Gewichtung der Impressions denkbar. Allerdings müssten dafür zusätzliche Informationen über den Verlauf der Impressions innerhalb eines Tages gegeben sein. Eine Gewichtung der Impressions gemäß des Order-Tagesverlaufs hätte zur Folge, dass keine Effekte auf die Rates erwartet würden. Da die Impressions-Zahlen als Offset in erster Linie das Ziel verfolgen, die Frequentierung der Publisher zu beschreiben, erscheint die einfache Gewichtung mittels Gleichverteilung sinngemäß. Die ITO-Rate ergibt sich nach (5.4) als Anteil der Orders z_{it} an den Impressions v_{it} zu

$$D_{it}^{ITO} = \frac{z_{it}}{v_{it}} \quad \Leftrightarrow \quad z_{it} = D_{it}^{ITO} \cdot v_{it} \ . \tag{5.4}$$

Es wird also angenommen, dass die Orders proportional zu den Impressions sind. Im zugrunde liegenden Poisson-Prozess entstehen mit konstanter Rate Counts (hier: Orders), welche proportional zur Anzahl an Impressions sind. Um der Abhängigkeit von Orders und Ad-Impressions Rechnung zu tragen, müssen die Orders auf die Größenordnung der Impressions heruntergebrochen werden. Zu diesem Zweck

kommt ein sogenannter Offset zum Einsatz.
Die abhängige Variable „Orders pro Stunde je Advertiser“ ist deutlich rechtsschief verteilt. Besonders „0-Beobachtungen“ (Zero Counts) sind stark ausgeprägt (siehe Abbildung 5.1). Das heißt, es kommt relativ selten zum Order-Erfolg und mehrfache Orders pro Stunde werden nur mit geringer Wahrscheinlichkeit realisiert. In der Datengrundlage liegen bei über 75% der stündlichen Beobachtungen keine Orders vor (Excess Zeros). Aufgrund dieser Tatsache ist die Modellierung mit einem reinen Poisson-Modell, wie vorgeschlagen, problematisch. Um der Vielzahl an 0-Beobachtungen Rechnung zu tragen, wird hier das Poisson-Modell zum Zero-inflated Poisson Model erweitert (vergleiche Kapitel 4.4). Mittels ZIP-Modell soll die Tatsache berücksichtigt werden, dass in den Daten neben einer Vielzahl von 0-Beobachtungen auch Beobachtungen von bis zu 30 Orders pro Stunde vorliegen. Das ZIP-Modell soll diese Heterogenität in den Daten auffangen und die Modellgüte gegenüber einem normalen Poisson-Modell verbessern. Man geht davon aus, dass manche Advertiser, aufgrund unbekannter Ursachen, keine Counts erzeugen können. Es ergibt sich die Mischverteilung

$$z_i \sim \begin{cases} 0 & \text{mit Wahrscheinlichkeit } p_i \\ Po(\lambda_i) & \text{mit Wahrscheinlichkeit } 1 - p_i \ , \end{cases}$$

wobei p_i die Wahrscheinlichkeit ist, dass im vorliegenden Fall nur eine 0 vorliegen kann. Mit der Gegenwahrscheinlichkeit $1 - p_i$ ist die Variable gemäß des oben beschriebenen Poisson-Modells verteilt.
Die zeitlichen Effekte, Stunde und Tag, werden als glatte Funktionen in das Modell aufgenommen. Dazu wird das Modell zu einem Generalisierten Additiven Modell (GAM) erweitert (siehe Hastie und Tibshirani, 1990). Im GAM werden nun in den linearen Prädiktor auch nicht-lineare, nonparametrische Funktionen aufgenommen. Diese werden ebenfalls aus den vorliegenden Daten geschätzt. Hier werden die zeitlichen Effekte als glatte Funktionen einbezogen. Dabei handelt es sich u.a. um die Tageszeit in Stunden. Da die Daten stundengenau zusammengefasst wurden, wird hier die Konvention getroffen, dass $h_t = 1$ die Beobachtungen zwischen 0.00 Uhr und 0.59 Uhr bezeichnet, sprich, die erste Stunde des Tages. Die Funktion $f(d_t)$ beschreibt den zeitlichen Effekt der Tage im Monatsverlauf. Dabei ist $d_t = 1$ der erste Tag im Monat und $d_t \in \{1, \ldots, 28\}$, da der Monat Februar als Betrachtungszeitraum vorliegt. Da es weniger um den Effekt einzelner Tage als um den von Abschnitten wie Monatsmitte oder -ende geht, lassen sich die Resultate problemlos auf andere Monate übertragen. Um einen möglichst glatten Effekt im Monatsverlauf zu erzielen, wird für den Glätter eine niedrige Basisdimension festgelegt (hier $k = 3$). Der Wochentags-Effekt wird kategorial in das Modell einbezogen und soll nicht in der glatten Funktion durchschlagen.
In den Modellen wird bei den zeitlichen Kovariablen generell mit zyklischen Splines geglättet. Sie besitzen die wünschenswerte Eigenschaft, dass Start- und Endpunkt den gleichen Wert annehmen. Beispielweise die Tageseffekte müssen an jedem neuen Tag auf demselben Niveau beginnen, damit sich die Ergebnisse verallgemeinern lassen. Das Package `gamlss` verwendet penalisierte zyklische B-Splines `cy()`. Im Package `mgcv` sind zyklische, penalisierte kubische Regressionssplines (`bs='cc'`) implementiert.
Aus der deskriptiven Analyse ist bekannt, dass bei Partnerschaften mit wenigen Im-

pressions aufgrund großer Standardfehler teils höhere Rates erzeugt werden. Deshalb werden die logarithmierten Impressions ebenfalls als glatte Funktion $f(\log(v_{it}))$ aufgenommen. Diese sollen den Effekt zunehmender Impressions beschreiben. Die Rate wird dann modelliert als

$$D_{it}^{ITO} = \frac{z_{it}}{v_{it}} = \eta_{it}^{k} + f(h_{it}) + f(d_{it}) + f\big(\log(v_{it})\big) + \epsilon_{it} \ . \tag{5.5}$$

Mit dem natürlichen Link zur Poisson-Familie, dem Log-Link, folgt die Responsefunktion

$$\mu_{it} = \exp\big\{\eta_{it}^{k} + f(h_{it}) + f(d_{it}) + f\big(\log(v_{it})\big) + \log(v_{it})\big\} \ , \tag{5.6}$$

wobei $\log(v_{it})$ der sog. Offset des Modells ist und η_{it}^{k} den Teil des Prädiktors ausgenommen der nicht-parametrischen Funktionen beschreibt. Im vorliegenden Fall handelt es sich bei den linearen Regressoren ausschließlich um kategoriale Größen. Die spätere Interpretation der Schätzer ist dabei immer in Relation zur Referenzkategorie zu setzen. Bei kategorialen Variablen wird die sogenannte Dummykodierung angewandt. Um die Identifizierbarkeit des Modells zu gewährleisten, gibt es jeweils eine Dummyvariable weniger als die Anzahl der Ausprägungen der kategorialen Variable.

Für das volle Modell wurden insgesamt fünf kategoriale Kovariablen vorgesehen. Die Variable „Wochentag“ besitzt die Referenzkategorie „Montag“. Außerdem wird nach der Branche des Advertisers differenziert. Die Ausprägungen sind „Fashion“, „Home & Accessoires“, „Kinder“, „Sonstige“ und „Vollversender“, mit der Referenzkategorie „Elektro“. Die binäre Variable „Advertiser Account Manager“ gibt an, ob der Advertiser einen Key Account besitzt (Referenz: kein Account Manager). Die unabhängige Variable „Advertisergröße“ ordnet die Advertiser in sechs Kategorien ein, wobei hier die kleinste Kategorie „Umsatz 100€ - 500€“ die Referenz darstellt. Des weiteren wurden spezielle Tage, an denen ein abweichendes Verhalten der User vermutet wird, in der binären Variable „Spezialtag“ erfasst. Es handelt sich im vorliegenden Datensatz um den Rosenmontag (11.02.), Faschingsdienstag (12.02.), Aschermittwoch (13.02.) und den Valentinstag (14.02.). Zwar sind dies keine (bundeseinheitlichen) gesetzlichen Feiertage, dennoch kommt ihnen eine besondere Bedeutung zu.

Die zugehörigen Parameter β_j beschreiben die Parameter aus (5.7), die geschätzt werden müssen. Beim Parameter β_0 handelt es sich um den sogenannten Intercept. Er gibt eine Schätzung für die Baseline an, d.h. den erwarteten Response, wenn alle Kovariablen auf 0 oder ihre Referenzkategorie gesetzt werden, und bei den glatten Funktionen durchschnittliche Effekte vorliegen. Durch die Einschränkung der Publisher auf das Business-Modell „Media“ mit Key Account, konnten Publisher-seitig keine aussagekräftigen Variablen in das Modell einbezogen werden.

Das nachfolgende Modell wurde in `R` mithilfe des Packages `gamlss` geschätzt. Im Gegensatz zum später noch häufiger verwendeten Package `mgcv` ist in diesem eine Vielzahl von Verteilungsfamilien für GLMs und GAMs implementiert, darunter auch das Zero-inflated Poisson Model. Für die geschätzten Modelle lässt sich das AIC evaluieren, mittels dessen die Variablen selektiert wurden. Es wurde eine Backward-Selektion durchgeführt, d.h. man beginnt mit dem vollen Modell inklusive aller zur Verfügung stehenden Kovariablen und entfernt in jedem Schritt eine Variable. Der Algorithmus wird dann abgebrochen, wenn durch weitere Reduktion des Modells das AIC nicht mehr verringert werden kann. Im vorliegenden Fall wurde die Variable Spezialtag aus dem Modell entfernt. Der lineare Prädiktor ergibt sich dann

zu

$$\begin{aligned}\eta^k =& \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{Branche}} + \beta_3 x_{\text{AdvAccountManager}} \\ &+ \beta_4 x_{\text{AdvGröße}} \ .\end{aligned} \tag{5.7}$$

5.1.2 Schätzung der Parameter

In diesem Abschnitt werden die Ergebnisse der Parameterschätzungen vorgestellt. Eine Übersicht über die geschätzten Koeffizienten liefert Tabelle A.1 im Anhang. Die Baseline des Modells ist die Impression-to-Order Rate am Montag bei einem Advertiser aus der Branche „Elektro“, ohne Key Account, aus dem Long Tail Bereich (100 - 500 EUR Umsatz), wenn durchschnittliche zeitliche Effekte vorliegen. Die geschätzten Effekte $\hat{\beta}_j$ lassen sich als Steigerung bzw. Senkung der ITO-Rate gegenüber der Baseline um den Faktor $\exp(\hat{\beta}_j)$ interpretieren. Durch die multiplikativen Eigenschaften des Response beim Log-Link lassen sich die Effekte einzelner Kovariablen losgelöst als Veränderung gegenüber der Referenzkategorie betrachten, unter der Prämisse, dass alle anderen Einflussgrößen konstant gehalten werden (Ceteris Paribus-Betrachtung). Bei positivem Vorzeichen handelt es sich um eine Steigerung, bei negativem Vorzeichen um eine Senkung gegenüber der Baseline.
Die glatten Effekte der zeitlichen Kovariablen sind in Abbildung 5.2 bzw. 5.3 dargestellt. In den Graphiken sind die Konfidenzbänder der Schätzung abgebildet, in denen sich der tatsächliche Effekt mit Wahrscheinlichkeit von 95% befindet.
Graphik 5.2 (oben) zeigt lediglich schwache Effekte im Monatsverlauf. Die Tageszeit (unten) hingegen weist sehr starke Effekte auf den Response auf. Man sieht einen deutlich negativen Effekt in den Nacht- und Morgenstunden. Ab circa 11.00 Uhr liegen im Tagesverlauf fast konstante Effekte vor. Es ergibt sich bis zum Abend lediglich ein geringer Anstieg, der seinen Höhepunkte zwischen 21.00 und 22.00 Uhr findet. Alternativ lassen sich die Effekte auch kombiniert als Kontourplot darstellen (vergleiche Abbildung 5.3).
Für das Zero-inflated Poisson-Modell muss ein weiterer Parameter geschätzt werden. Es handelt sich dabei um die Wahrscheinlichkeit p_i, dass im betreffenden Fall kein Count erzeugt werden kann. Mit dem Logit-Link ergibt sich die Wahrscheinlichkeit für einen Zero Count zu 12,9% ($\log(\hat{p}/(1-\hat{p})) = -1.903$).
Der Wochentag hat einen signifikanten Einfluss auf die erwartete Conversion Rate. Die meisten Orders in Relation zu den Impressions werden sonntags und montags erwartet. Die anderen Wochentage haben ungefähr ein identisches Niveau, wobei für Freitag und Samstag die niedrigsten Rates geschätzt werden (vergleiche auch Abbildung 5.4). Beispielsweise wird für Samstag eine um den Faktor 0.81 ($= \exp(-0.21313)$) geringere ITO-Rate gegenüber der Referenzkategorie Montag erwartet, ceteris paribus. Die Koeffizienten der Schätzer sind betragsmäßig klein, sodass die Wochentage eher geringe Änderungen der Rate verursachen.
Deutlich stärkere Effekte ergeben sich für die Advertiser-Branche. Auf dem gleichen Niveau wie die Referenzkategorie Elektro liegen die erwarteten Rates der Advertiser aus den Bereichen Fashion und Home & Accessoires. Eine deutlich gesteigerte ITO-Rate liegt bei Retailern für Kinderprodukte und den sonstigen Advertisern vor. Das niedrigste Verhältnis zwischen Orders und Impressions wird bei den Vollversendern beobachtet. Besitzt ein Advertiser einen Key Account im Affiliate Netzwerk führt eine Impression im Mittel häufiger zur Order. Bei den Umsatzklassen der Adverti-

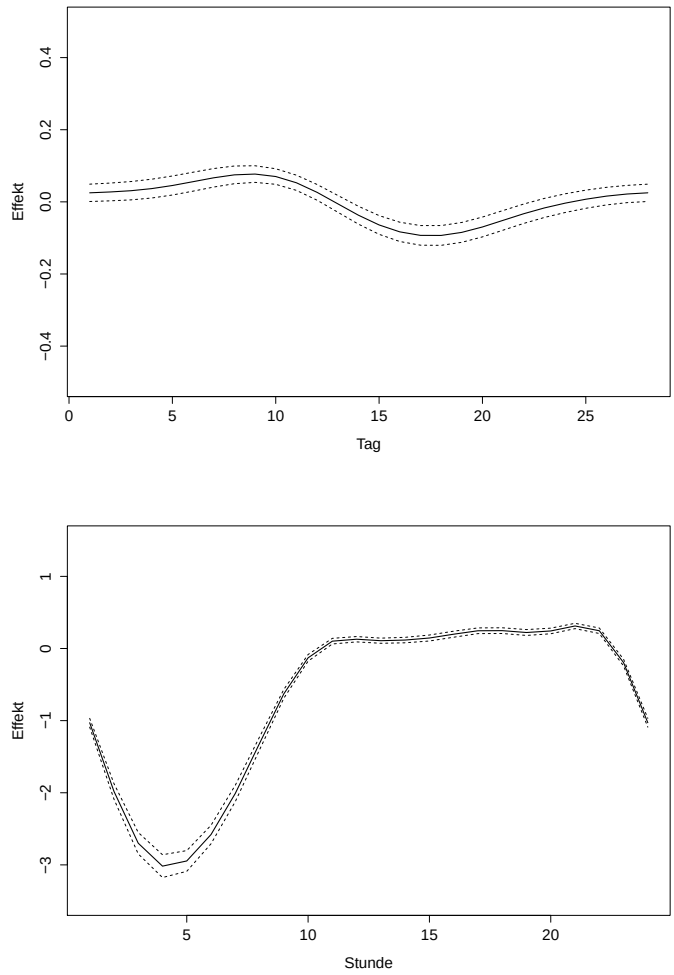

Abbildung 5.2: Zeitliche Effekte Monat und Stunde

ser zeigt sich, dass bei Advertisern mit höherem Umsatz im Netzwerk eine höhere Conversion Rate erwartet wird. Von der kleinsten Umsatzklasse (500-1000 EUR) steigen die Effekte bis zur Klasse 2000-5000 EUR an und bewegen sich auch für die Advertiser mit den höchsten Umsätzen auf diesem Niveau. Für die Umsatzklasse 5000-10000 EUR werden ein bisschen niedrigere Rates als für die Kategorie 2000-5000 EUR erwartet. Die Koeffizienten-Schätzer für Advertiser Key Account und Advertiser-Größe stützen die Vermutung eines vorliegenden Brand Effects. Es wird davon ausgegangen, dass große Advertiser mit starken Markennamen bessere Rates erzeugen.
Abbildung 5.5 zeigt die Effekte der logarithmierten Impressions. Die erwartete Rate ist bei vergleichsweise wenigen Impressions (ca. 100 pro h) maximal und fällt von da ab mit steigender Impressionszahl. Gibt es nahezu keine Impressions, so ist auch die erwartete mittlere Rate niedrig. Der Abfall der Rate bei steigender Impressionszahl ist erwartungsgemäß, da die Standardabweichung bei Vorliegen weniger Impression größer ist. Falls bei den Publishern wenige Impressions generiert werden, können

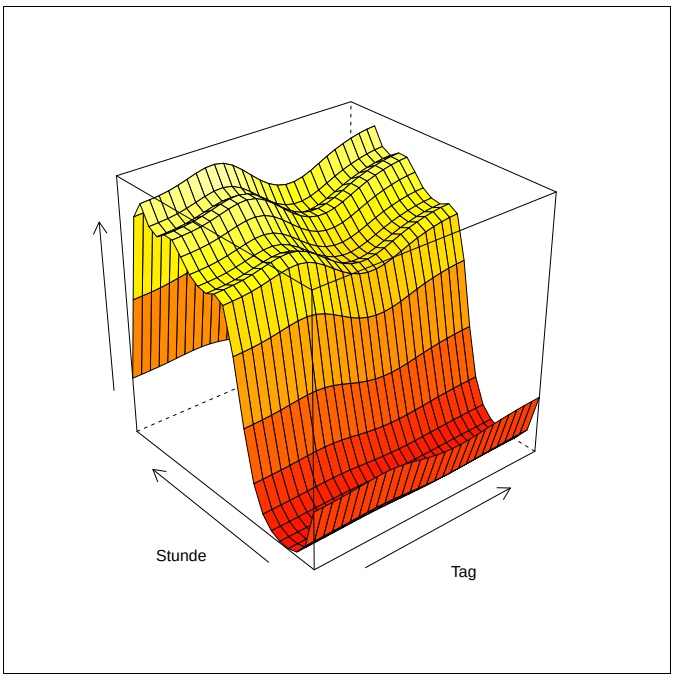

Abbildung 5.3: Kombinierte zeitliche Effekte Monat und Stunde

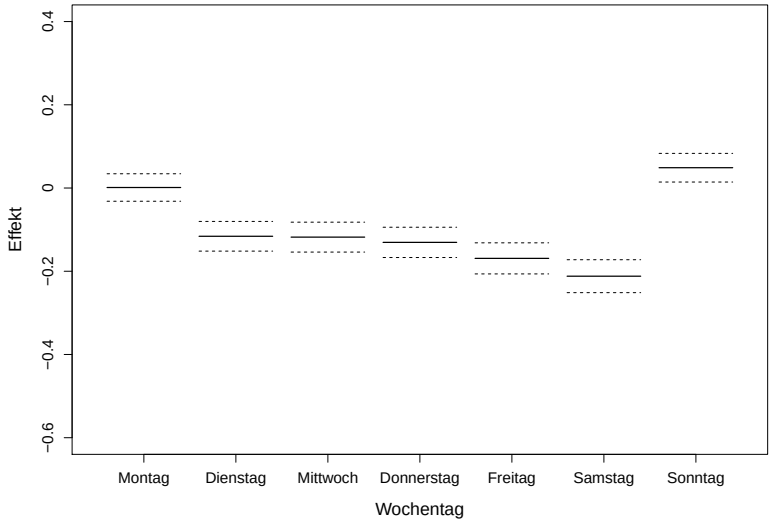

Abbildung 5.4: GAM ITO: Effekte Wochentag

schon wenige Orders zu hohen Rates führen. Überraschend ist daher, dass hier die erwartete Rate bei ca. 50-150 Impressions pro Stunde höher liegt als für Publisher mit 0-50 Impressions pro Stunde.

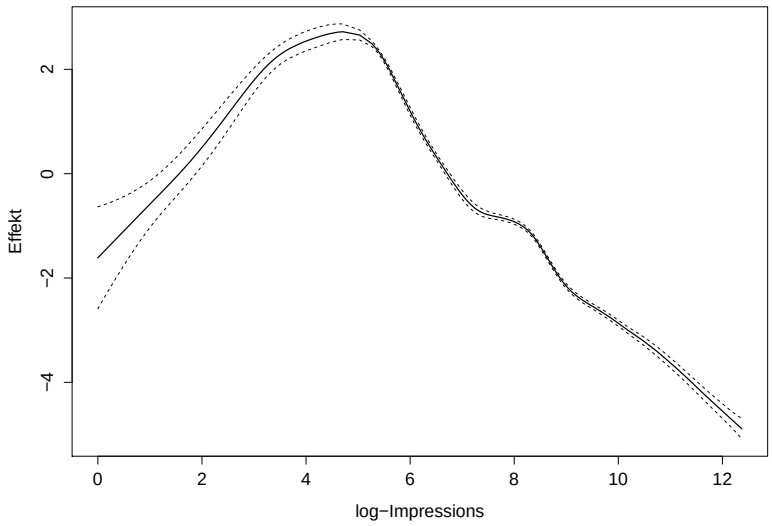

Abbildung 5.5: GAM ITO: Effekte log-Impressions

5.1.3 Modelldiagnose

Ein probates Mittel zur Überprüfung der Modellannahmen und -güte ist die sogenannte Residualanalyse. Residuen bezeichnen die Differenz zwischen den beobachteten und den vom Modell geschätzten Werten. Wie im linearen Regressionsmodell wird auch bei GAMs die Normalität der Residuen angenommen. Außerdem wird im Modell Varianzhomogenität unterstellt, d.h. ungeachtet der Größe der gefitteten Werte müssen die Residuen mit konstanter Varianz um die Null streuen.
Für die Residualanalyse sind im Package `mgcv` unter anderem Devianzresiduen implementiert, während das Package `gamlss` mit standardisierten Residuen arbeitet.
Die Residualplots weisen keine gravierenden Probleme mit der Modellgüte auf (siehe Abbildung 5.6). Die Annahme normalverteilter Residuen wird mittels QQ-Plot und Histogramm der Residuen überprüft. Hier zeigen sich lediglich geringe Abweichungen von der Normalverteilung aufgrund leichter Rechtsschiefe (Residualplots links). Der Residualplot der gefitteten Werte gegen die Residuen (rechts oben) zeigt, dass die Streuung für kleine Werte höher ist als für große Vorhersagewerte. Die Annahme der Varianzhomogenität ist daher nur unter Einschränkungen erfüllt. Allerdings streuen die Residuen gut um die Null. Bei hoher Modellgüte befinden sich die Werte im Plot rechts unten in der Nähe der Winkelhalbierende, d.h. Vorhersagewert und beobachteter Wert stimmen überein. Hier gruppieren sich die Tupel mit einiger Streuung um die Winkelhalbierende. Die Residualplots für das gefittete Modell weisen einige Schwächen bei Erfüllung der Modellannahmen auf, die jedoch insgesamt in einem vertretbaren Rahmen liegen.

5.1.4 Modell mit Interaktionseffekten

Bisher wurden die zeitlichen Effekte für alle vorliegenden Advertiser gemeinsam geschätzt. Ist man besonders an den Unterschieden zwischen den zeitlichen Effekten

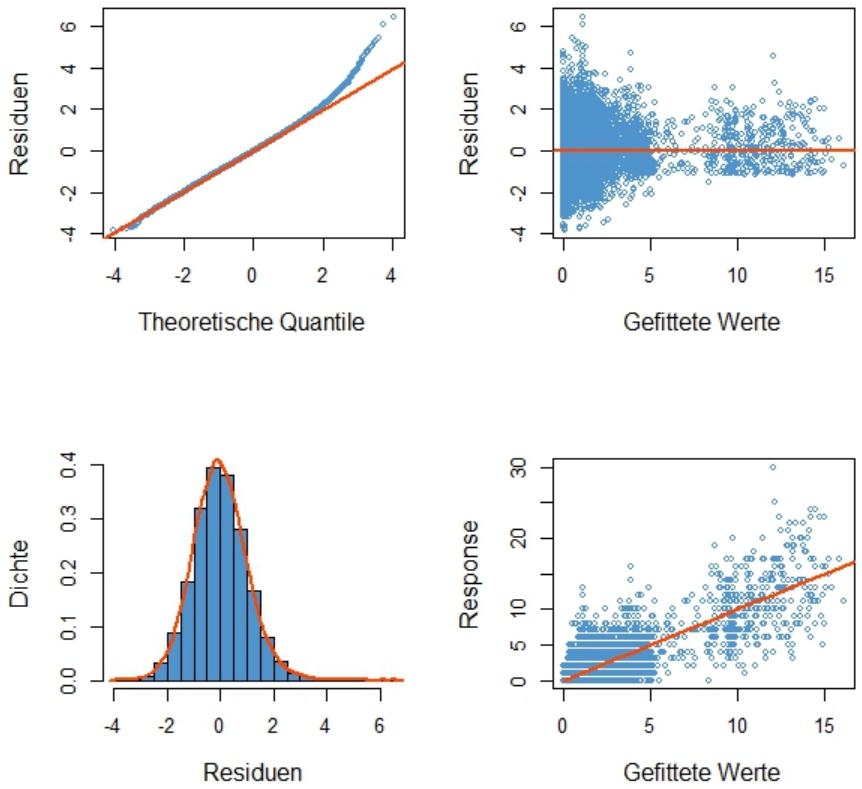

Abbildung 5.6: GAM ITO: Residualplots

in den einzelnen Advertiser-Branchen interessiert, macht es Sinn, das Modell um Interaktionen zu erweitern. So können die zeitlichen Effekte auf die Branche des Advertisers bedingt werden. Gleichung (5.6) wird erweitert zu

$$\mathbb{E}(z_{it}) = v_{it} \cdot \exp\{\eta_{it}^k + f_{h_{it}|x_B}(h_{it}) + f_{d_{it}|x_B}(d_{it}) + f(\log(v_{it}))\}, \tag{5.8}$$

wobei x_B hier die Faktorvariable Advertiser-Branche ist. Durch das Einbeziehen der Interaktionseffekte zwischen der Faktorvariable Branche und den glatten Effekten ändern sich die Parameter für die kategorialen Größen gegenüber dem ursprünglichen Modell (5.6) nur geringfügig. Die Signifikanzen und Richtungen der Effekte bleiben nahezu unverändert.
Alle zeitlichen Effekte, bedingt auf die Advertiser-Branche, ergeben signifikante Schätzer. Beim Tageseffekt zeigen sich für die einzelnen Branchen abweichende Verläufe (siehe Abbildung 5.7). An den Konfidenzbändern lässt sich außerdem erkennen, ob eine Gruppe viele oder wenige Beobachtungen besitzt. Je weniger Beobachtungen, desto höher ist die Standardabweichung des Schätzers und die Breite der Konfidenzbänder nimmt zu. Alle Branchen haben gemeinsam, dass ein negativer Effekt auf die Order-Rate in den Nacht- und frühen Morgenstunden vorliegt. Diese unterscheiden sich jedoch in ihrer Stärke. Bei den Retailern im Bereich Home & Accessoires und Kinderprodukte ist der negative Effekt besonders stark, die ITO-Rate liegt

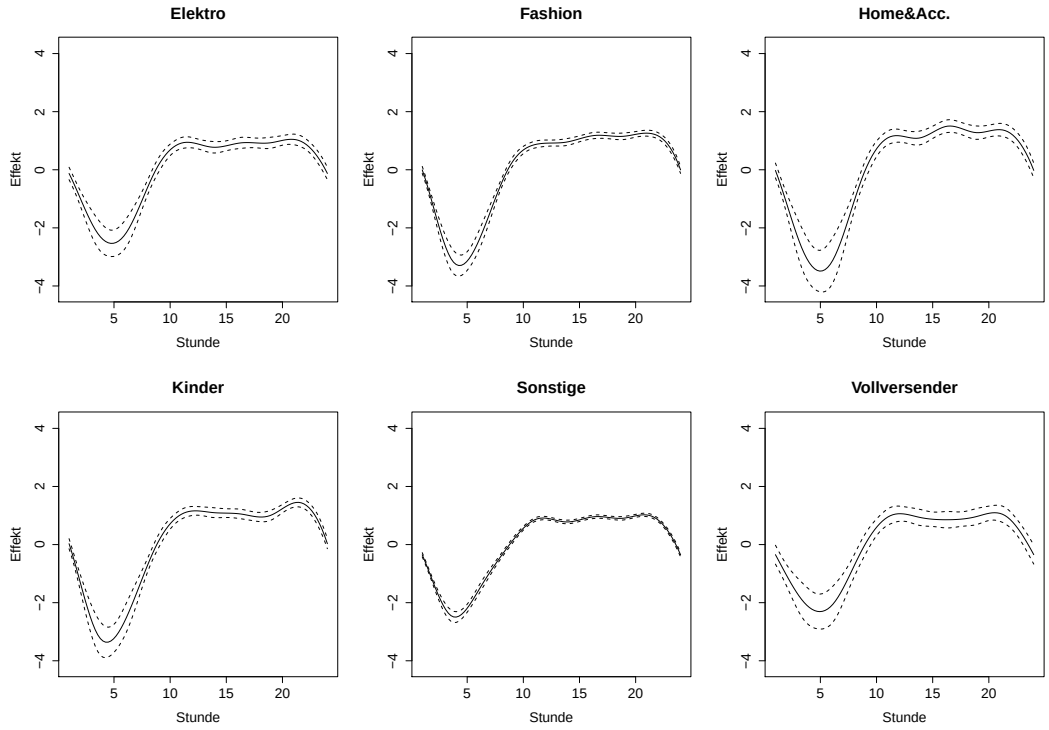

Abbildung 5.7: GAM ITO: Interaktionseffekte Tagesverlauf

nachts deutlich niedriger als tagsüber. Bei den meisten Advertisern wird ein erster Höhepunkt der Order-Rate gegen 11.00 Uhr erreicht. Beim Großteil der Advertiser-Branchen sind die Rates zwischen Mittag und Nachmittag relativ konstant und steigen bis zum Höhepunkt um ca. 21.00 Uhr nur langsam an. Nach diesem Hoch am Abend sinkt der erwartete Order-Erfolg wieder bis auf das Nacht-Niveau. Auffällig ist der deutliche Peak in den mittleren relativen Orders um ca. 21.00 Uhr in der Advertiser-Branche „Kinder“. Die höchste Schätzunsicherheit liegt in der Branche Vollversender vor. Die Konfidenzbänder sind hier relativ breit.
Auch beim Monatseffekt lassen sich Unterschiede zwischen den Branchen erkennen, siehe Abbildung 5.8. Die stärksten Effekte der Monatszeit auf die ITO-Rate liegen in der Branche Elektro vor. Hier wird ein Abfall der erwarteten Rates zur Monatsmitte hin beobachtet. In den anderen Branchen ergeben sich lediglich schwache Effekte im Monatsverlauf. Bei den Advertisern der Branchen Fashion, Vollversender und den Sonstigen werden am Monatsanfang die meisten Orders erwartet, bei den Retailern für Kinderprodukte hingegen in der zweiten Monatshälfte. In der Branche Home & Accessoires gibt es keine signifikanten Effekte. Hier wird allerdings auch nur ein Advertiser betrachtet.

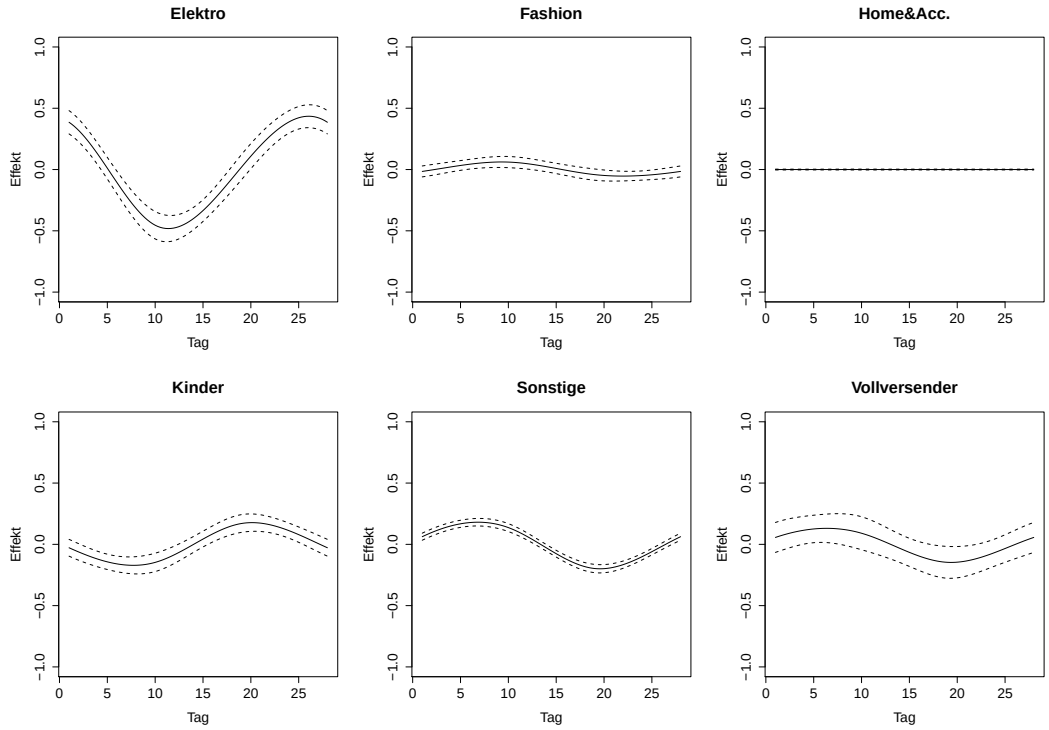

Abbildung 5.8: GAM ITO: Interaktionseffekte Monatsverlauf

5.2 Modellierung der Conversion Rate (CR)

5.2.1 Datengrundlage und Modellannahmen

Der Übergang vom Anklicken eines Werbemittels zum Kauf beim Advertiser lässt sich mit der Conversion Rate (CR) beschreiben. Sie ist definiert als das Verhältnis der Orders zu den getätigten Klicks. Auch hier liegen die Daten von Klicks und Orders in Einzeldatensätzen vor, welche miteinander gematcht werden. Es wird wiederum davon ausgegangen, dass die Zeitpunkte der beiden Ereignisse zusammenfallen. Im vorliegenden Modell werden die Orders pro Advertiser über alle Publisher aggregiert. Zum einen verkleinert das den Datenumfang erheblich, da mehr als 1500 Publisher vorliegen. Zum anderen werden durch die Aggregation zahlreiche 0-Beobachtungen eliminiert. Die Conversion Rate (CR) ergibt sich dann zu

$$D_{it}^{CR} = \frac{z_{it}}{y_{it}} \,, \tag{5.9}$$

wobei z_{it} wiederum die Anzahl der Orders bei Advertiser i zum Zeitpunkt t und y_{it} die Anzahl der Klicks auf die Werbemittel von Advertiser i auf allen Publisher-Websites zum Zeitpunkt t beschreibt.
Die notwendige Voraussetzung für das Zustandekommen einer Order ist, dass mindestens ein Klick generiert wurde. Alle anderen Fälle werden aus der Analyse ausgeschlossen. Für den Februar 2013 liegen dann nach Bereinigung rund 50000 Datenzeilen mit stündlichen Beobachtungen vor. Dabei wurden alle 75 Advertiser mit

Daten von rund 2000 Publishern im Modell berücksichtigt. Die Besetzung der Kategorien entspricht in dieser Stichprobe nahezu dem kompletten Orders-Datensatz (vergleiche Tabellen 3.1 und 3.2). Lediglich ein Advertiser wurde ausgeschlossen, da dieser kaum Klicks bei den Media-Publishern hervorbrachte.
Bei Zähldaten liegt in der Regel Überdispersion vor. Das bedeutet, dass die Varianz in den Daten größer ist als im unterstellten Modell. Bei der Schätzung der Parameter muss dem Rechnung getragen werden. Statt eines normalen Poisson-Ansatzes wählt man dann den sog. Quasi-Poisson-Ansatz. Wird eine normale Poisson-Verteilung angenommen, dann gilt

$$\lambda_i = \mathbb{E}(y_i) = \mathrm{Var}(y_i) \ .$$

Um die erhöhte Varianz in den Daten zu berücksichtigen, wird ein Dispersionsparameter ϕ eingeführt. Die bedingte Varianz lautet dann

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi \ \lambda_i \ .$$

Bei vorliegender Überdispersion ist $\phi > 1$ und die bedingte Varianz steigt schneller als der Erwartungswert. Man erhält einen Schätzer für den Dispersionsparameter, indem man die Residualdevianz durch die Freiheitsgrade der Residuen teilt. Liegt keine Überdispersion vor, dann nähert sich der Quotient dem Wert 1 an. Die Schätzung der Parameter erfolgt im Quasi-Poisson-Ansatz nicht mehr durch den normalen ML-Schätzer, sondern über die Quasi-Likelihood, vergleiche auch Fahrmeir, Kneib und Lang (2009) oder Fox (2008).
Die Zielgröße „Orders pro h pro Advertiser“ ist rechtsschief verteilt (siehe Abbildung 5.9). Eine hohe Anzahl von Orders pro Stunde wird mit geringer Wahrscheinlichkeit beobachtet. Es gibt einige Ausreißer mit über 100 Orders pro Stunde pro Advertiser.

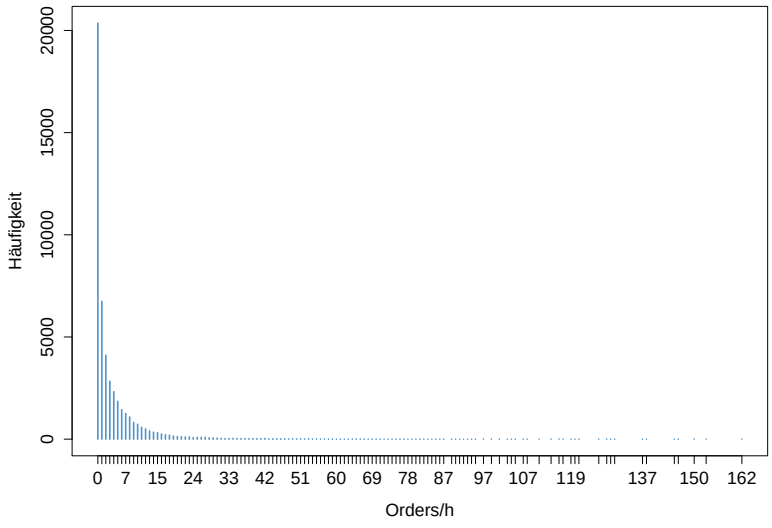

Abbildung 5.9: GAMM Conversion Rate: Verteilung Orders

Das Modell für die Conversion Rate (CR) lässt sich formulieren als

$$D_{it}^{CR} = \frac{z_{it}}{y_{it}} = \eta_{it}^{k} + f_{h_{it}|x_B}(h_{it}) + f_{d_{it}|x_B}(d_{it}) + f(\log(y_{it})) + \epsilon_{it} \ . \tag{5.10}$$

Für die erwartete Anzahl an Orders erhält man

$$\mu = \mathbb{E}(z_{it}) = \exp\{\eta_{it}^k + f_{h_{it}|x_B}(h_{it}) + f_{d_{it}|x_B}(d_{it}) + f(\log(y_{it})) + \log(y_{it})\}.$$

Den Offset bildet hier die Anzahl der Klicks pro Stunde, welche im Datensatz genau erfasst sind. Im Gegensatz zum Modell in Kapitel 5.1 muss hier der Offset nicht angepasst werden. Als glatte Effekte werden wiederum die Tageszeit h_{it} und die Tage im Monatsverlauf d_{it} in das Modell aufgenommen. Auch hier werden mittels Interaktionen die zeitlichen Effekte nach Branche differenziert. $f_{h_t|x_B}(h_t)$ sei der Interaktionseffekt zwischen der glatten Funktion über die Tageszeit und der Faktorvariable „Advertiser-Branche“. Kategoriale Variablen werden wie im Modell für die ITO-Rate mit den gleichen Ausprägungen und Referenzkategorien in das Modell aufgenommen. Für die Kovariable Advertiser-Größe ist in diesem Modell die Kategorie 0€ - 100€ Referenzkategorie.
Für die Modellklasse `gam()` ist keine Funktion zur Modellselektion nach AIC- bzw. BIC-Kriterium implementiert, die vergleichbar mit der Funktion `AICstep()` für GLMs oder lineare Modelle wäre. Man kann jedoch das AIC bzw. BIC von GAMs evaluieren. Hier wird der Poisson-Ansatz zur Modellvalidierung verwendet, da er für diese Zwecke ähnliche Ergebnisse wie der Quasi-Poisson-Ansatz liefert. Die Variablen wurden mittels BIC Backward-Selektion gewählt. Der lineare Teil des vollen Modells vor der Selektion lautete

$$\begin{aligned}\tilde{\eta}^k =& \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{Branche}} + \beta_3 x_{\text{AdvAccountManager}} + \\ & \beta_4 x_{\text{AdvGröße}} + \beta_5 x_{\text{Spezialtag}} + \gamma_{0,\text{Advertiser}}.\end{aligned}$$

Bei der Variablenselektion nach BIC werden die Kovariablen Spezialtag und Advertiser Account Manager entfernt. Somit ergibt sich der Prädiktor als

$$\eta^k = \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{Branche}} + \beta_3 x_{\text{AdvGröße}} + \gamma_{0,\text{Advertiser}} \ .$$

Das Generalisierte Additive Modell wird um zufällige Effekte zum Mixed Model erweitert. Bei den vorliegenden Daten handelt es sich um Clusterdaten. Die verschiedenen Partnerschaften oder Advertiser können als Individuen betrachtet werden, die korrelierte Counts erzeugen. Die Autokorrelation der Counts zeigt sich in der deskriptiven Analyse, vergleiche Abbildungen 3.12 und 3.13. Die Impressions, Klicks und Orders schwanken je nach Advertiser oder Partnerschaft auf verschiedenen Niveaus. Die zufälligen, individuenspezifischen Effekte werden hier mit γ_0 bezeichnet. Es wird berücksichtigt, dass die einzelnen Advertiser Eigenheiten aufweisen, die die Schätzer der Fixed Effects nicht beeinflussen sollen.

5.2.2 Zeitliche Effekte

Das Modell wurde in `R` mithilfe des Packages `mgcv` von Wood (2006) gefittet. Die Variablen Branche und Advertisergröße weisen keine signifikanten Effekte auf. Sonntags ist die erwartete mittlere Conversion Rate signifikant höher als in der Referenzkategorie Montag. Freitags und samstags wird die niedrigste Erfolgsrate geschätzt. Ansonsten wird das vorliegende Modell lediglich durch glatte Funktionen und Random

Effects beschrieben. Tabelle A.2 im Anhang gibt eine Übersicht über die Schätzparameter des Modells.
Entfernt man die Random Effects aus dem Modell, dann ergeben sich teils hochsignifikante Kovariablen-Effekte für die kategorialen Einflussgrößen. Jedoch verschlechtert sich die Modellgüte gegenüber dem Mixed Model. Das deutet darauf hin, dass hier anscheinend beträchtliche Effekte aufgrund der individuellen Eigenschaften der Advertiser generiert werden. Aufgrund der besseren Modellgüte wurden die zufälligen Effekte im Modell belassen und im Folgenden die glatten Funktionen des GAMs interpretiert.
Die Stärke der Effekte im Monatsverlauf ist wie bei der ITO-Rate eher gering (vergleiche Abbildung 5.10). Die erwarteten Orders steigen bei Advertisern der Branchen Elektro und Fashion zur Monatsmitte hin an und erreichen einen Tiefpunkt zum Monatsende. In diesen Branchen sind die Effekte betragsmäßig auch am größten. Ein nahezu gegensätzlicher Verlauf ergibt sich für Advertiser der Branchen Kinderprodukte und Vollversender. Weite Konfidenzbänder und recht schwache Effekte lassen sich bei der Advertiser-Branche Home & Accessoires beobachten. Bei den sonstigen Advertisern sind die Rates am Monatsbeginn erhöht.

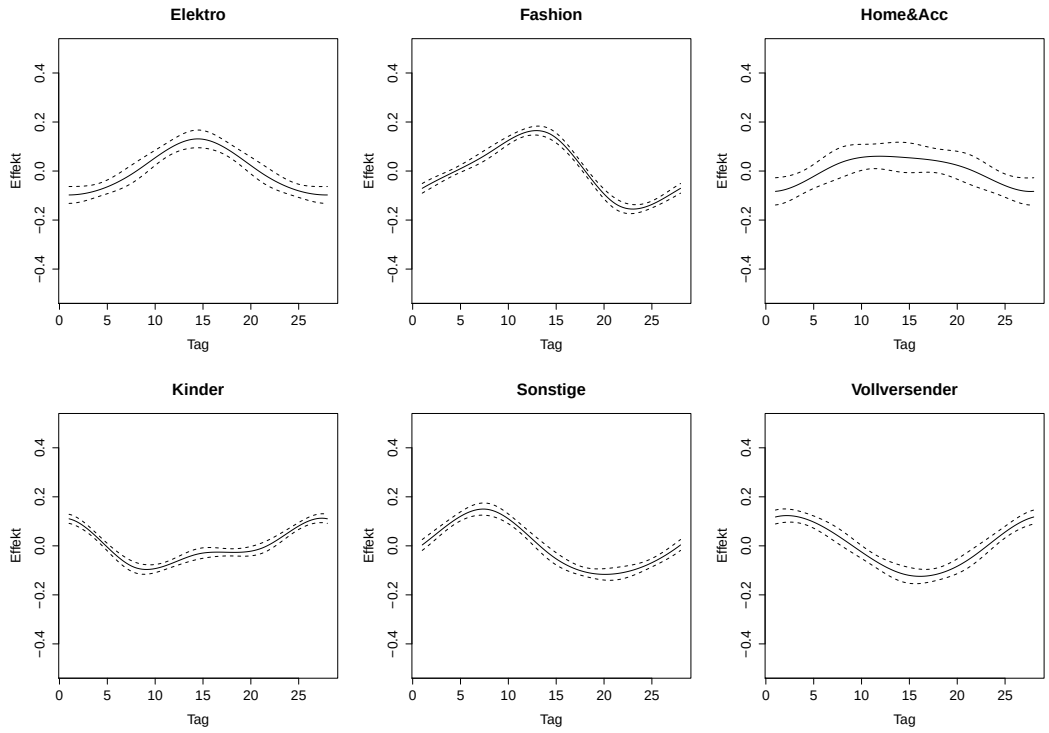

Abbildung 5.10: GAMM Conversion Rate: Effekte im Monatsverlauf

Wie schon in den anderen Modellen zeigen sich starke Effekte auf die Conversion Rates im Tagesverlauf. Auf den ersten Blick zeigen die geglätteten Funktionen in Abbildung 5.11 recht ähnliche Verläufe für alle Branchen. In den Nachtstunden ist der erwartete Kauferfolg bei Advertisern aller Branchen unterdurchschnittlich mit einem globalen Tiefpunkt zwischen 4.00 Uhr und 5.00 Uhr morgens in allen Branchen. Bei genauerer Betrachtung fallen allerdings einige Unterschiede zwischen den

Branchen auf. In den Branchen Home & Accessoires und vor allem den Retailern für Kinderprodukte ist die erwartete Kaufrate gegen 22.00 Uhr deutlich erhöht. Bei den Vollversendern ist die Conversion Rate im Tagesverlauf vergleichsweise konstant. Hier fallen vor allem die negativen Effekte in den Nachtstunden deutlich geringer aus als in anderen Branchen. Die größten Schwankungen im Tagesverlauf ergeben sich bei den Conversion Rates in der Branche Home & Accessoires.

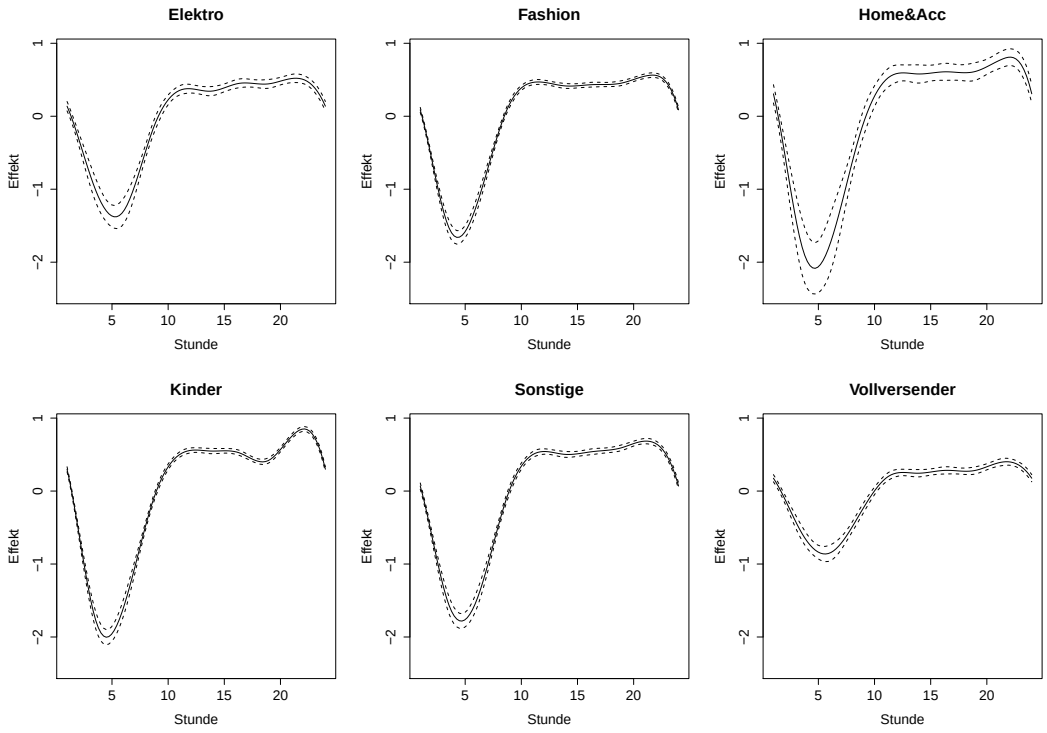

Abbildung 5.11: GAMM Conversion Rate: Effekte im Tagesverlauf

Für steigende Klick-Zahlen nimmt die erwartete Conversion Rate ab. Für den Effekt der Klick-Zahlen liegen betragsmäßig starke Effekte vor. Die Conversion Rate bei Advertisern mit einer hohen Anzahl an Klicks liegt deutlich niedriger als für Advertiser, deren Werbemittel nur selten Klicks generieren, vergleiche Abbildung 5.12 (links).
Die Aufnahme der Advertiser als zufällige Effekte funktioniert recht gut. Der Normal-Quantil-Plot zeigt abgesehen von den Effekten eines Advertisers keine gravierenden Abweichungen von der angenommenen Normalverteilung (siehe Abbildung 5.12 rechts).
Um neben den zeitlichen Einflussgrößen weitere Kovariablen-Effekte analysieren zu können, wird in einem weiteren Modell in Kapitel 5.3 ein alternativer Ansatz gewählt. Hier werden gleichartige Partnerschaften von Advertisern aus verschiedenen Branchen herausgegriffen und anhand einer größeren Menge von Kovariablen untersucht.

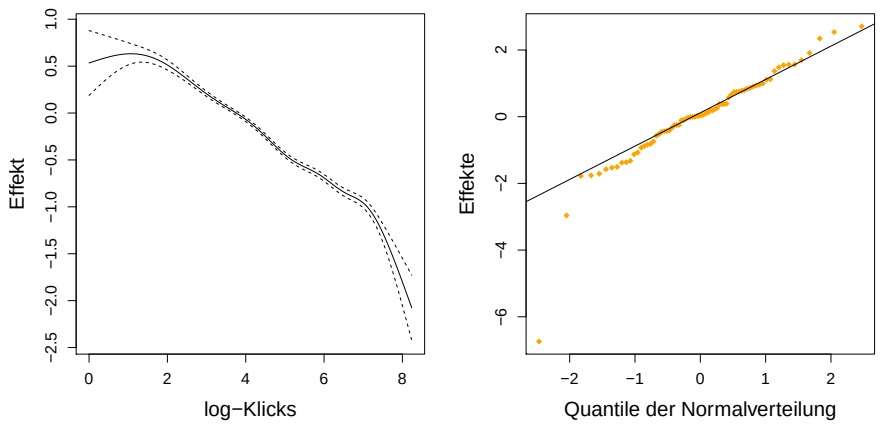

Abbildung 5.12: GAMM Conversion Rate: Effekte logarithmierte Klicks (links) und Normal-Quantil-Plot der zufälligen Effekte (rechts)

5.2.3 Modelldiagnose

Die Schätzung mit der Funktion `gam` weist ein adjustiertes Bestimmtheitsmaß von 82.8% auf. Dabei werden 68.5% der auftretenden Devianz durch das Modell erklärt. Diese Maße weisen auf eine ordentliche Modellgüte hin. Sorgen hingegen macht die Normalverteilungsannahme der Devianzresiduen. Die Diagnoseplots in Abbildung B.1 weisen deutliche Anpassungsschwächen bei der Normalverteilungsannahme auf. Die Residualplots zeigen zwar eine einigermaßen symmetrische Verteilung der Residuen, die jedoch gegenüber der Normalverteilung zu stark gekrümmt ist.
Für die Anpassungsprobleme kann es verschiedene Ursachen geben. Beispielsweise können fehlende Einflussgrößen ausschlaggebend sein. Außerdem werden für Regressionsmodelle homogene Zielgrößen unterstellt. Diese Annahme ist im vorliegenden Fall problematisch. Zwar soll der Heterogenität durch die Random Effects entgegengewirkt werden, allerdings lässt sich diese nicht vollständig beseitigen. Die einzelnen Advertiser fließen als Random Effects ein. Die Advertiser haben jedoch eine Vielzahl verschiedener Publisher, die wiederum individuenspezifische Eigenschaften besitzen. Durch die Aggregation der Daten kann nicht jede Partnerschaft zusätzlich als Random Effect modelliert werden.
Der Dispersionsparameter für das Quasi-Poisson-Modell wird auf 1.46 geschätzt. In den analysierten Daten liegt also erwartungsgemäß eine leichte Überdispersion vor. Die bedingte Varianz steigt also schneller als der Erwartungswert.
Die Tatsache, dass sowohl die meisten Kovariablen als auch der Intercept des Modells keine signifikanten Werte annimmt, lässt auf eine unzureichende Anpassung des Modells schließen. Dennoch wurde das Modell nicht verworfen. Zum einen lassen sich die glatten Effekte der zeitlichen Variablen trotzdem interpretieren. Zum anderen wurde aus dem vorliegenden Modell ein alternativer Ansatz entwickelt (Kapitel 5.3). Somit können zumindest die nicht-parametrischen Funktionen der beiden Ansätze verglichen werden.

5.3 Modellierung der Conversion Rate für ausgewählte Partnerschaften

5.3.1 Datengrundlage und Modellannahmen

Bisher wurden die Datensätze in aggregierter Weise analysiert. In Kapitel 5.2 wurden die Orders pro Advertiser über alle Publisher zusammengefasst. Daraus folgt zum einen, dass seitens der Publisher keine Einflussgrößen aufgenommen werden konnten. Zum anderen weist das Modell in seiner Anpassung deutliche Schwächen auf. Generell ist man auch daran interessiert, einzelne Partnerschaften in disaggregierter Form, d.h. die Anzahl der Orders zwischen Advertiser i und Publisher j, zu modellieren. Auch diese Modellierung ist mit einigen Schwierigkeiten verbunden. Die Datensätze sind sehr umfangreich (vergleiche Kapitel 2). Legt man die bisher betrachteten 75 Advertiser mit rund 2000 Publishern aus dem Klicks-Datensatz zugrunde, so explodiert die Anzahl an stündlichen Beobachtungen. Im Zeitraum Februar 2013 liegen für die Advertiser bei jeweils 20 Publishern bereits 1.000.800 stündliche Beobachtungen vor. Dabei ist die Anzahl von 20 Publishern pro Advertiser noch niedrig angesetzt. Außerdem wären der Großteil dieser Beobachtungen sogenannte Zero Counts, d.h. Tagesstunden in denen kein Klick und/oder keine Bestellung stattgefunden haben. Das liegt teils daran, dass bei vielen Partnerschaften die Publisher-Website nicht an jedem Tag ein Werbemittel des Advertisers schaltet. Es ist nicht bekannt, wann tatsächlich Impressions stattgefunden haben. Außerdem sind die Größenordnungen der Partnerschaften sehr unterschiedlich. Eine vielbesuchte Publisher-Website erzeugt in der Regel mehr Klicks/Orders als weniger bedeutende Werbeträger. Aus diesem Grund führt eine disaggregierte Modellierung aller Partnerschaften zu einer nicht mehr modellierbaren Heterogenität der Daten.
Um dennoch die Conversion Rate von Partnerschaften modellieren zu können, wurden für das nachfolgende Modell bestimmte Affiliates selektiert. Dabei handelt es sich um die zehn Partnerschaften mit den meisten Sales je Advertiser-Branche. Insgesamt bilden also 60 Partnerschaften mit einer großen Anzahl an Bestellungen die Datengrundlage. Somit liegen vergleichsweise wenige 0-Beobachtungen und einander ähnliche Partnerschaften vor. Die unerwünschte Heterogenität in den Daten wird somit begrenzt. Das nachfolgende Modell modelliert also zeitliche Verläufe und weitere Kovariablen speziell für große, erfolgreiche Partnerschaften. Hier ist die stündliche Conversion Rate

$$D_{ijt}^{CR} = \frac{z_{ijt}}{y_{ijt}} \ ,$$

wobei z_{ijt} die Orders der Partnerschaft zwischen Advertiser i und Publisher j, und y_{ijt} entsprechend die Klicks dieser Affiliates zum Zeitpunkt t sind.
Nach Bereinigung des Datensatzes liegen rund 30000 stundenweise Beobachtungen vor. Die Orders pro Stunde sind wieder rechtsschief verteilt (vergleiche Abbildung 5.13). Sogar bei diesen großen Advertisern sind die „0-Orders“ wieder stark repräsentiert, allerdings mit circa 40% Anteil deutlich weniger als in den vorangegangen Modellen.
Zur Modellierung wird der Quasi-Poisson-Ansatz mit Log-Link herangezogen, um vorhandene Overdispersion zu berücksichtigen. Außerdem wird das Modell zum gemischten Modell erweitert, indem die Partnerschaften als Random Effects berück-

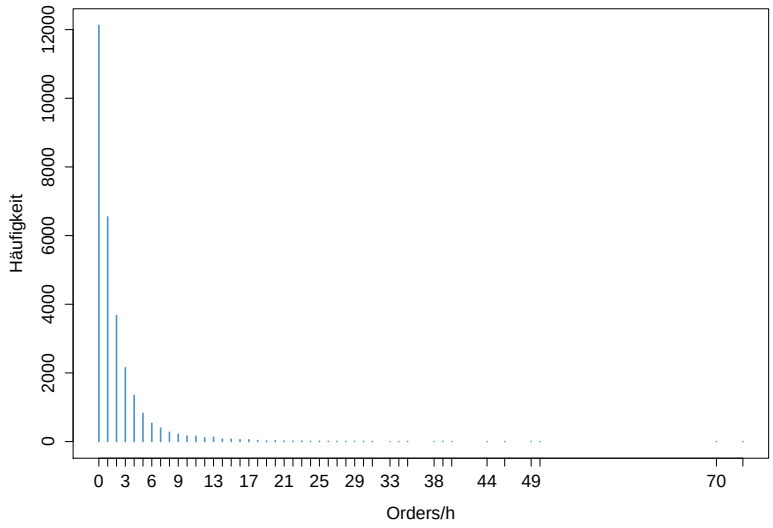

Abbildung 5.13: GAMM CR Partnerschaften: Verteilung Orders

sichtigt werden.
Die Modellgleichung für die Conversion Rate (CR) entspricht (5.10), außer dass es sich jetzt um Beobachtungen von Paaren statt über alle Publisher aggregierte Beobachtungen handelt. Als glatte Effekte werden wiederum die Tageszeit h_t und die Tage im Monatsverlauf d_t in das Modell aufgenommen. Diese fließen als Interaktionseffekte mit der Advertiser-Branche ein. Im Vergleich zu den vorangegangenen Modellen steht eine größere Zahl von Einflussgrößen zur Verfügung. Für die Kovariable Advertiser-Größe ist in diesem Modell die Kategorie 500€ - 1000€ Referenzkategorie, da keine kleineren Advertiser im Modell vertreten sind. Für das Modell kommen Publisher-seitig zwei weitere Einflussgrößen in Betracht. Zum einen wird das Business Model der Publisher aufgenommen. Referenzkategorie ist hier „Cash Back Site“. Außerdem kommt die binäre Kovariable Publisher Account Manager hinzu, die zwischen Publishern mit und ohne Account Manager unterscheidet.
Es findet eine Modellvalidierung anhand des BIC-Kriteriums unter Verwendung der Likelihood des Poisson-Modells statt. Anhand einer Rückwärts-Selektion wird überprüft, ob ein sparsameres Modell eine bessere Anpassung als das volle Modell mit allen Kovariablen liefert. Der lineare Prädiktor des vollen Modells vor der Selektion lautet

$$\begin{aligned}\tilde{\eta}^k =& \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{Branche}} + \beta_3 x_{\text{AdvAccountManager}} + \\ & \beta_4 x_{\text{AdvGröße}} + \beta_5 x_{\text{Spezialtag}} + \beta_6 x_{\text{PubBusinessModel}} + \\ & \beta_7 x_{\text{PubAccountManager}} + \gamma_{0,\text{Partnerschaft}} \cdot\end{aligned}$$

Im ersten Schritt wird die Kovariable Spezialtag, dann Advertiser Account Manager und im dritten Schritt die Variable Publisher Account Manager entfernt. Dieses sparsamere Modell ist BIC-optimal gegenüber dem vollen Modell. Der lineare Teil

des Prädiktors lautet dann

$$\eta^k = \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{Branche}} + \beta_3 x_{\text{AdvGröße}} + \beta_4 x_{\text{PubBusinessModel}} + \beta_5 x_{\text{PubAccountManager}} + \gamma_{0,\text{Partnerschaft}} \cdot$$

Als individuenspezifische Effekte werden die 60 Partnerschaften aufgenommen. Jede Partnerschaft bildet dabei ein einzelnes Cluster, was aufgrund der Heterogenität der Partnerschaften positive Wirkungen auf die Modellgüte hat. Abbildung 5.14 zeigt individuenspezifische Effekte exemplarisch für einige Partnerschaften. Die stündlichen Conversions der einzelnen Partnerschaften unterscheiden sich in der Tat recht stark. Die Behandlung als einzelne Cluster soll die individuenspezifischen Effekte auffangen.

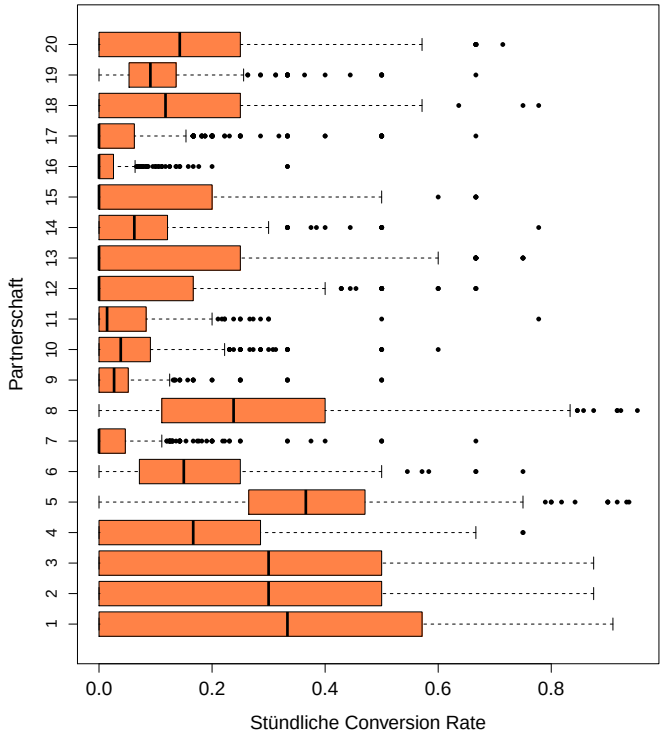

Abbildung 5.14: Conversion Rates für verschiedene Partnerschaften

5.3.2 Parameterschätzer

Die Effekte im Monatsverlauf weisen große Ähnlichkeiten mit denen im Modell in Kapitel 5.2 auf. Auffällig ist dennoch, dass der positive Effekt in der Branche Fashion hier viel stärker ausgeprägt ist. Abgesehen von den Advertisern der Branche Elektro ergeben sich ansonsten kaum nennenswerte, von der Zeit im Monat verursachten Effekte auf die Conversion Rate (siehe Abbildung 5.15).
Einen größeren Einfluss auf die erwarteten relativen Orders hat die Tageszeit. Abbildung 5.16 zeigt den üblichen Verlauf mit den höchsten Rates in den Abendstunden.

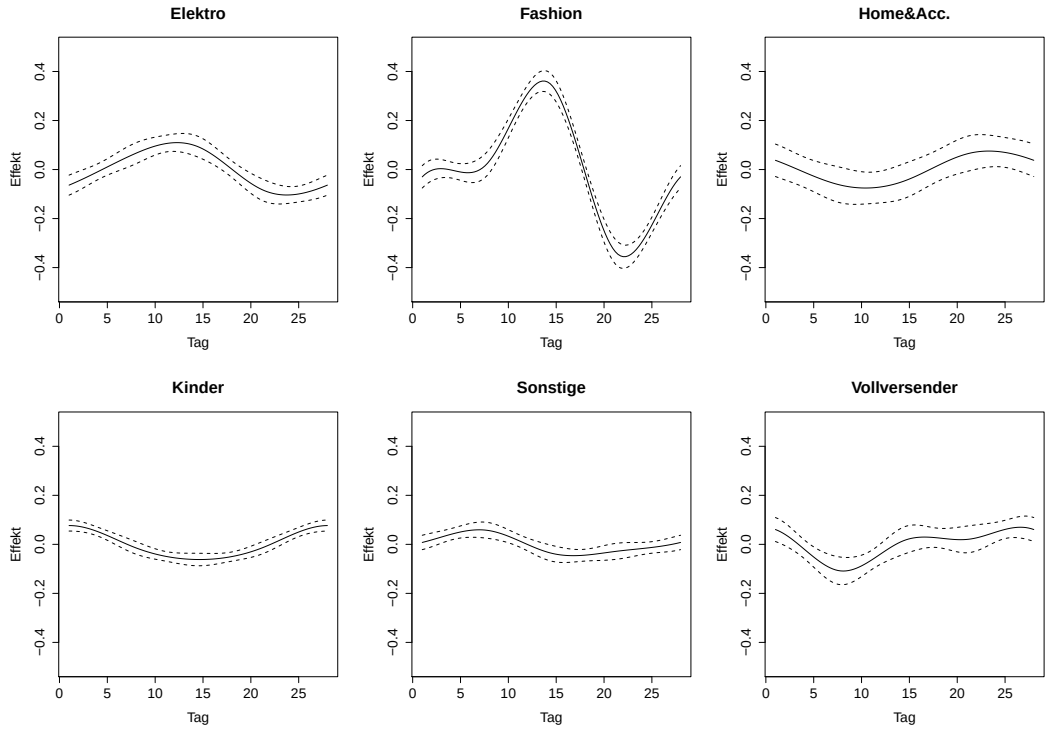

Abbildung 5.15: GAMM CR Partnerschaften: Monatseffekte branchenweise

Die einzelnen Branchen ähneln sich in den Effekten recht stark. Bei den Händlern für Kinderprodukte, aber auch in den Branchen Fashion und Vollversender, ist ein 22.00 Uhr-Peak sichtbar.
Wie im vorangegangen Modell sinken die erwarteten Conversions mit zunehmenden Klick-Zahlen, wobei hier die Effekte betragsmäßig ein wenig schwächer eingeschätzt werden. Der Normal-Quantil-Plot zeigt, dass die Partnerschaften als Cluster für das Mixed Model durchaus Sinn machen. Es gibt keine gröberen Abweichungen von der Normalverteilung, vergleiche auch Abbildung 5.17 (rechts).
Im Gegensatz zum Modell mit aggregierten Daten liegen hier noch einige signifikante Effekte von kategorialen Kovariablen vor. Eine Übersicht der Parameterschätzer ist in Anhang A.3 gegeben. So hat der Wochentag einen signifikanten, wenn auch betragsmäßig schwachen, Einfluss auf die erwarteten Conversions. An Sonntagen wird demnach die höchste Conversion der User erwartet. Am kleinsten sind die Schätzer für die Wochentage Freitag und Samstag, siehe auch Abbildung 5.18.
Starke Unterschiede zeigen sich bei der Differenzierung nach Branchen (Graphik 5.19). Vergleichsweise hohe Conversions liegen bei den Kinderprodukte-Advertisern und den sonstigen Werbeträgern vor. Für die Referenzkategorie „Elektro“ wird die niedrigste mittlere Conversion Rate geschätzt. Sie liegt ungefähr auf demselben Niveau wie die erwartete CR bei den Vollversendern und in der Branche Home & Accessoires. Als Publisher-seitige Kovariable ist die kategoriale Größe „Publisher Business Model“ im Modell verblieben. Hier zeigt sich bei der erwarteten Rate in erster Linie ein großer Unterschied zwischen Usern, die auf Publisher-Websites aus dem Geschäftsmodell „Cash Back“ abgesprungen sind und allen anderen. Bei Klicks, die

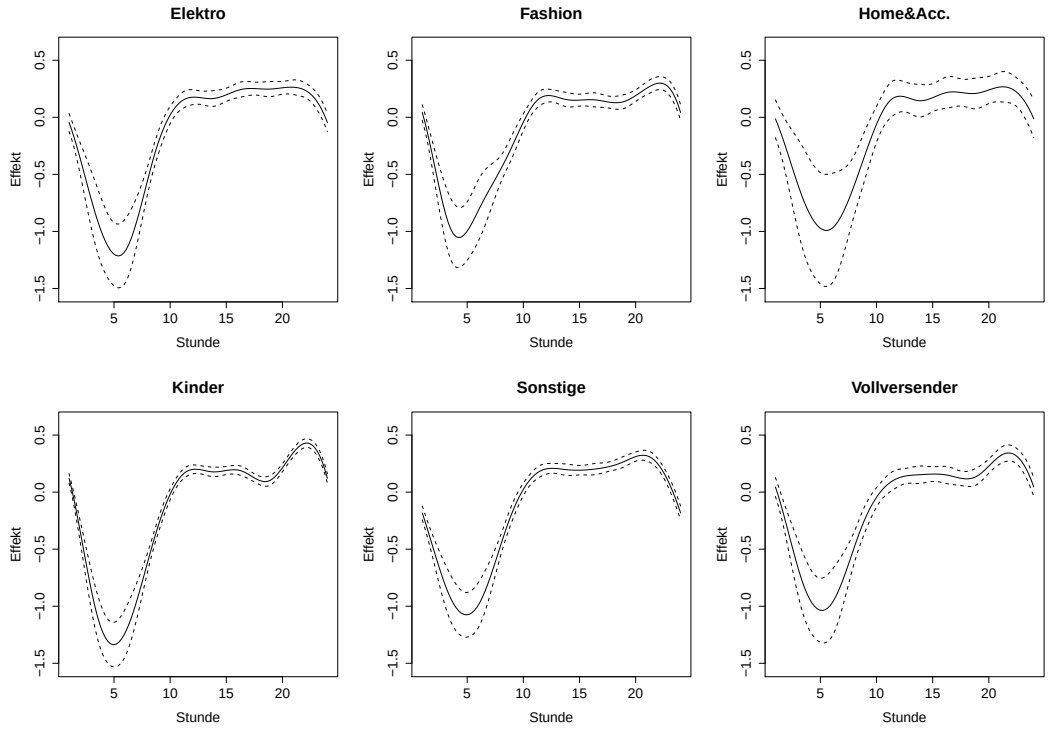

Abbildung 5.16: GAMM CR Partnerschaften: Tageseffekte branchenweise

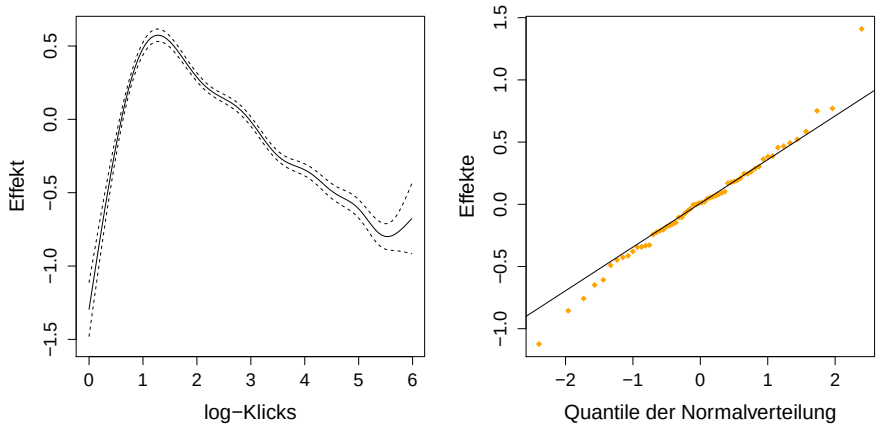

Abbildung 5.17: GAMM CR Partnerschaften: Effekte logarithmierte Klicks (links) und Normal-Quantil-Plot der zufälligen Effekte (rechts)

von Cash Back-Publishern generiert werden, ist die mittlere Conversion Rate deutlich höher. Die Geschäftsmodelle Coupon, Media und Topic Website liegen nahezu auf demselben Niveau. Die verhältnismäßig geringste Zahl an Orders wird bei Usern erwartet, die von einer Publisher-Website zum Thema Preisvergleich abgesprungen sind. Die Umsatzgröße der Advertiser liefert hier keinen signifikanten Einfluss auf

die erwarteten Conversions.

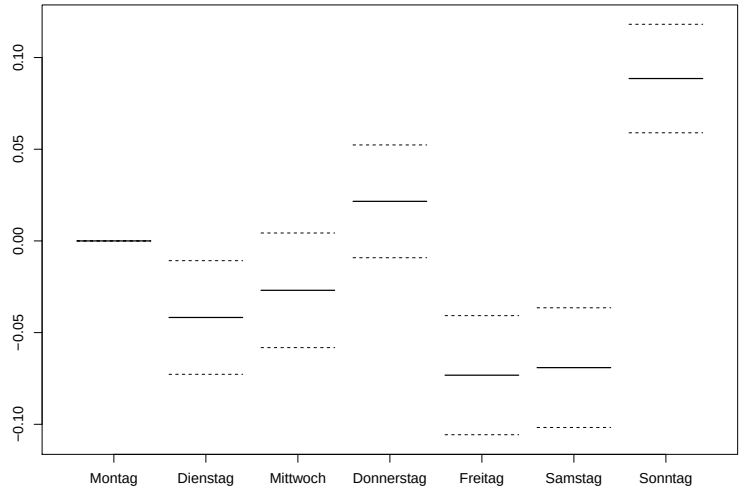

Abbildung 5.18: GAMM CR Partnerschaften: Effekte Wochentag

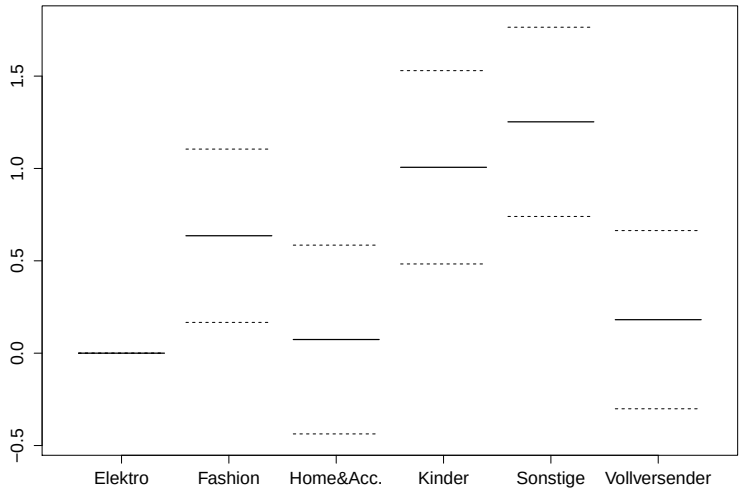

Abbildung 5.19: GAMM CR Partnerschaften: Effekte Branchen

5.3.3 Modelldiagnose

Insgesamt zeigt das Modell eine zufriedenstellende Anpassung an die Daten. Das adjustierte Bestimmtheitsmaß liegt hier bei 79,4% und die erklärte Devianz bei 55,7%. Außerdem wurde durch die Selektion großer, erfolgreicher Partnerschaften die Über-

dispersion aus dem Modell beseitigt. Der geschätzte Dispersionsparameter beträgt 0.95.
Die Residualanalyse bestätigt die Verbesserung der Modellgüte. Auf den ersten Blick liegen hier kaum Verletzungen der Annahmen vor. Der Plot der Residuen (Anhang B.2 rechts oben) zeigt einige Ausreißer in den Residuen an, die wohl aus Ausreißern bei den beobachteten Orders resultieren. Die Anpassung an die unterstellte Normalverteilung der Residuen funktioniert insgesamt jedoch recht gut.
Durch die gezielte Auswahl von großen Partnerschaften, welche in ihren Frequentierungen vergleichbar sind, konnte die Heterogenität in den Daten reduziert werden. Natürlich ergeben sich durch die Selektion der Daten Implikationen für die Interpretation der Schätzer und Verallgemeinerung der Ergebnisse. Auf große Partnerschaften mit täglichen Orders sollten sich die Erkenntnisse jedoch problemlos übertragen lassen.

5.4 Modellierung des Traffics im Jahresverlauf

5.4.1 Datengrundlage und Modellannahmen

Im nachfolgenden Modell wird das Ziel verfolgt, langfristige Effekte zu identifizieren. Dazu wurden Daten über das komplette Jahr 2012 hinweg betrachtet. Fraglich ist, ob es saisonale Effekte auf die Conversions gibt, wie zum Beispiel Steigerungen in der Vorweihnachtszeit oder ein Gefälle zwischen Sommer und Winter. Die Daten für die Untersuchung liegen in tageweise aggregierter Form vor. Intraday-Effekte im Jahresverlauf lassen sich daher nicht analysieren. Im Gegensatz zu den anderen Modellen über die Conversion Rates sind die Daten von Impressions, Klicks und Orders einheitlich in einem Datensatz erfasst. Die Rate Impression-to-Order kann direkt modelliert werden. Der Erfolg (Anzahl Orders) wird abhängig von den Impressions gemessen (Impression-to-Click-Rate). Es handelt sich um die Rate

$$D_{ijt}^{ITO} = \frac{z_{ijt}}{v_{ijt}} \,, \tag{5.11}$$

sprich das Verhältnis von Orders zu Impressions am Tag t pro Partnerschaft zwischen Advertiser i und Publisher j.
Es wurden nur Partnerschaften einbezogen, die über das ganzen Jahr hinweg dokumentiert wurden. Diese Einschränkung trägt zu einer ausbalancierten Stichprobe bei. Ein Hinzukommen und Ausscheiden von Partnerprogrammen im Analysezeitraum könnte im betrachteten Modell aufgrund der ohnehin geringen Anzahl an analysierten Partnerschaften zu starken Verzerrungen führen. Im Modell verbleiben dann 33 Partnerschaften, die sich aus 14 Advertisern und 18 Publishern zusammensetzen. Aufgrund der Dokumentation von Impressions wurden für das Modell die Beobachtungen bei Publishern aus dem Business Model Media mit Account Manager selektiert. Aus dem Schaltjahr 2012 liegen pro Partnerschaft Beobachtungen an 366 Tagen vor. Nach Selektion bilden insgesamt rund 12.000 Beobachtungen mit Tagesdaten die Datengrundlage.
Betrachtet man die neue Datengrundlage genauer, fällt auf, dass sich der Jahresverlauf der absoluten Orders im Subsample gegenüber dem kompletten Datensatz unterscheidet (vergleiche Graphiken 3.16 und 5.20). Zum einen weist die geglättete

Zeitreihe aufgrund der geringeren Anzahl an betrachteten Partnerschaften eine größere Streuung auf. Außerdem ist der Anstieg der Bestellungen in der Vorweihnachtszeit beim Subsample moderater als in den kompletten Jahresdaten. Diese Tatsachen sollte man bei der Interpretation der Effekte im Hinterkopf behalten. Trotz einiger Einschränkungen leistet das Subsample dennoch eine vernünftige Approximation der Gesamtdaten und die Allgemeingültigkeit der Ergebnisse sollte nicht gefährdet sein.

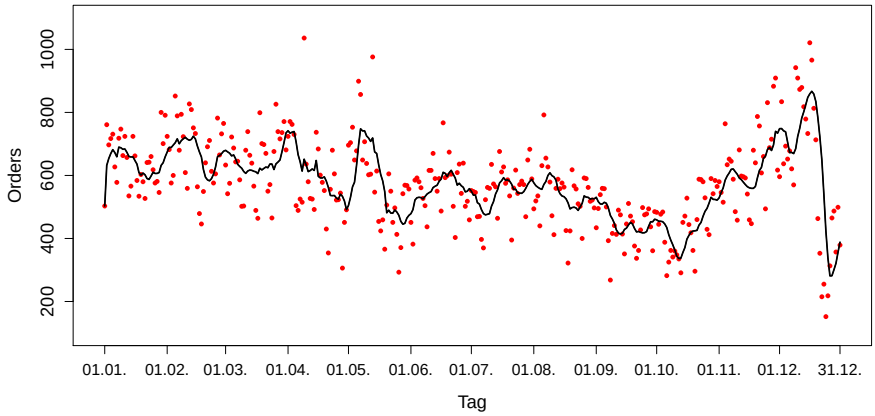

Abbildung 5.20: GAMM Jahresdaten: Absolutwerte Orders

Hier wird der zeitliche Verlauf lediglich über die Tage im Jahr modelliert. Die Aufnahme der Tage im Monat als glatte Funktion liefern keine Verbesserung des Modells. Der Monatseffekt bleibt deshalb unberücksichtigt. Die Anzahl der Orders wird modelliert als

$$z_{ijt} = \exp\{\eta^k_{ijt} + f_{t|x_B}(t) + f\big(\log(v_{ijt})\big) + \log(v_{ijt})\} + \epsilon_{ijt} \ , \tag{5.12}$$

für alle Partnerschaften und Tage in 2012 mit $t = 1, \ldots, 366$. Die zeitliche Komponente (Tag im Jahr) fließt als glatte Funktion mit Interaktion Advertiser-Branche (x_B) in das Modell ein. Zur Identifikation von Größeneffekten wird die logarithmierte Anzahl an Impressions ebenfalls als glatte Funktion modelliert. Was Verteilungs- und Strukturannahme betrifft, orientiert man sich wieder an den vorangegangenen Modellen. Auch hier wird ein Quasi-Poisson-Ansatz mit Log-Link verwendet. Die logarithmierten Impressions dienen als Offset des Modells.
Es ist eine plausible Annahme, dass das Verhalten der Internetnutzer an Feiertagen abweichen könnte. Dazu wurde die neue Variable „Feiertag“ aufgenommen, die bundeseinheitliche gesetzliche Feiertage erfasst. Nach Variablenselektion durch BIC ergibt sich der Prädiktor

$$\begin{aligned}\eta^k =& \beta_0 + \beta_1 x_{\text{Wochentag}} + \beta_2 x_{\text{AdvGröße}} + \beta_3 x_{\text{Branche}} + \\ & \beta_4 x_{\text{Feiertag}} + \gamma_{0,\text{Partnerschaft}} \cdot\end{aligned}$$

Zur Verbesserung der Modellgüte wurden wiederum Random Effects aufgenommen. Im vorliegenden Fall wird davon ausgegangen, dass die einzelnen Partnerschaften das

Modell als zufällige Komponente beeinflussen. Das macht Sinn, da es sich um heterogene Partnerschaften mit spezifischen Eigenschaften handelt. Es kann angenommen werden, dass die Anzahlen der Orders innerhalb einer Partnerschaft abhängig sind.

5.4.2 Schätzung der Koeffizienten

Die zeitlichen Effekte wurden branchenweise differenziert betrachtet. Hier fällt auf, dass für manche Branchen im Jahresverlauf kaum Effekte auf die Conversion Rates erwartet werden. Das ist bei Advertisern in der Branche Home & Accessoires und den Sonstigen der Fall (vergleiche Abbildung 5.22). Die Jahreszeit hat einen stärkeren Einfluss auf die erwarteten Conversions bei den Elektro- und Fashion-Advertisern. Hier ist der Bestellanstieg im Weihnachtsgeschäft deutlich erkennbar. Die erwartete Rate liegt gegen August/September in beiden Branchen jeweils am niedrigsten. Genau zu dieser Jahreszeit liegen hingegen bei den Advertisern für Kinderprodukte die stärksten positiven Effekte vor. Das Weihnachtsgeschäft ist in der Branche „Kinder“ nicht erkennbar, hier liegt die erwartete Conversion am niedrigsten. Auch wenn das auf den ersten Blick kontraintuitiv ist, lassen sich mögliche Begründungen finden. Vor allem muss man sich ins Gedächtnis rufen, dass die Conversion Rate (5.11) auch aufgrund von vermehrter Ausbringung von Werbemitteln sinken kann. Wenn beispielsweise einzelne Branchen zu bestimmten Jahreszeiten verstärkt Werbemittel schalten, sprich die Anzahl an Impressions erhöhen, müssen die Orders um den gleichen Prozentsatz steigen, damit die Rates konstant bleiben. Gerade bei Kinderprodukten ist es denkbar, dass diese in der Vorweihnachtszeit im Übermaß beworben werden und somit die Effekte verloren gehen. In der Branche Home & Accessoires liegen zum Jahresanfang positive Effekte auf die Conversion Rates vor. Außerdem verzeichnet die erwartete Erfolgsrate in der Branche Fashion im Monat März neben der Vorweihnachtszeit noch einen zweiten Höhepunkt. Breite Konfidenzintervalle werden bei den Vollversendern ersichtlich. Anfang April sowie im Monat August liegen hier die erwarteten Conversion Rates am höchsten.

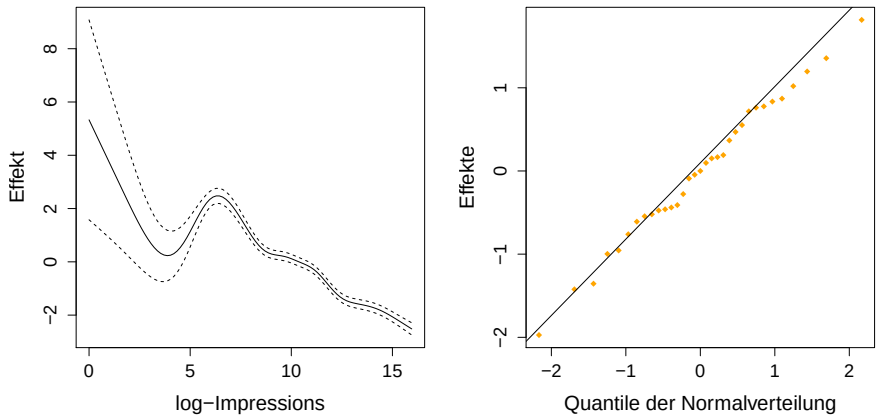

Abbildung 5.21: GAMM Jahresdaten: Effekt Anzahl Impressions und Random Effekte

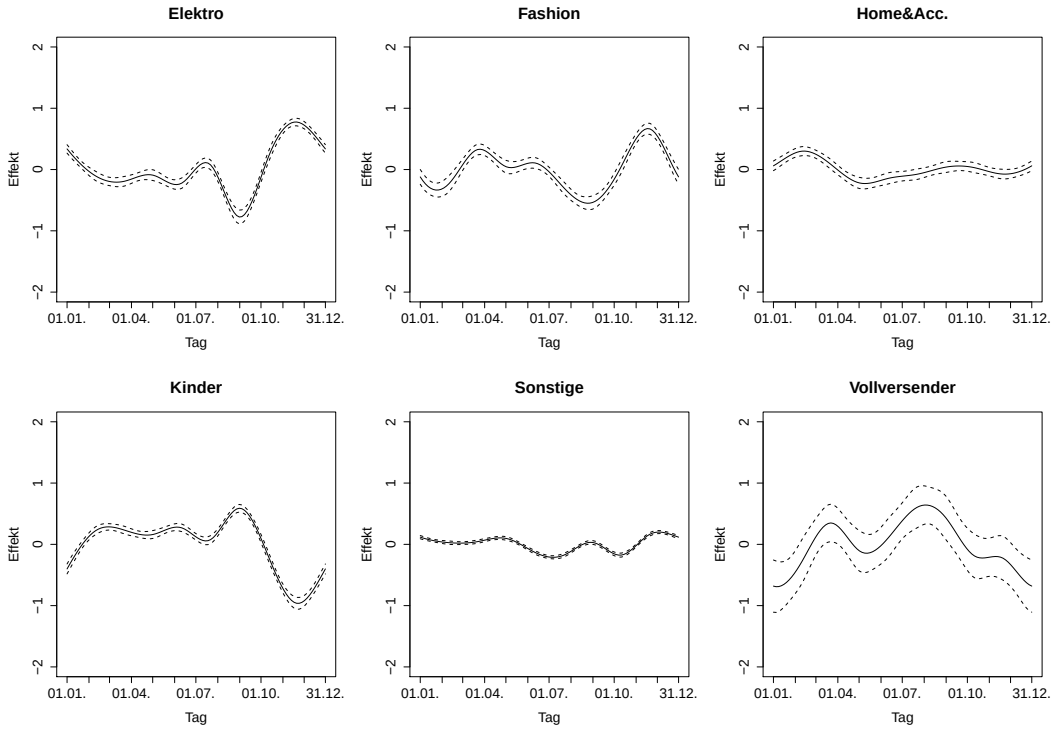

Abbildung 5.22: GAMM Jahresdaten: Zeitliche Effekte branchenweise

Wie vermutet, nimmt die erwartete Conversion Rate mit steigender Anzahl an Impressions wiederum ab. Abbildung 5.21 (links) zeigt die Effekte der logarithmierten Impression-Anzahl. Die vorliegenden Partnerschaften wurden als Random Effects einbezogen. Die zufälligen Effekte liefern eine gute Anpassung an die Normalverteilung (siehe Abbildung 5.21 rechts).
Die erwartete Conversion Rate liegt sonntags und montags am höchsten. Die wenigsten Orders relativ zu den Impressions werden am Freitag und Samstag erwartet, siehe auch Abbildung 5.24. Das Modell sagt für Samstage eine Verringerung der erwarteten Conversion um den Faktor 0.78 gegenüber Montagen voraus. Die höchsten Conversion Rates werden montags und sonntags erwartet. Die geschätzten Koeffizienten sind in Anhang A.4 aufgeführt.
Die Schätzer für die Einflüsse der Kovariable Branche und Advertiser-Größe nehmen relativ große Werte an. Das deutet an, dass die Conversion Rates über diese Faktoren recht heterogen sind. Abbildung 5.24 zeigt die Effekte der Branchenzugehörigkeit im Einzelnen. Dabei sind vor allem die niedrigen erwarteten Erfolgsraten in den Branchen Fashion und Vollversender auffällig. Verhältnismäßig viele Orders werden bei Advertisern mittleren Umsatzes erwartet (1000-2000 EUR bzw. 2000-5000 EUR). Feiertage haben negative Auswirkungen auf die Conversion. Die erwartete Rate liegt um den Faktor 0.92 niedriger.

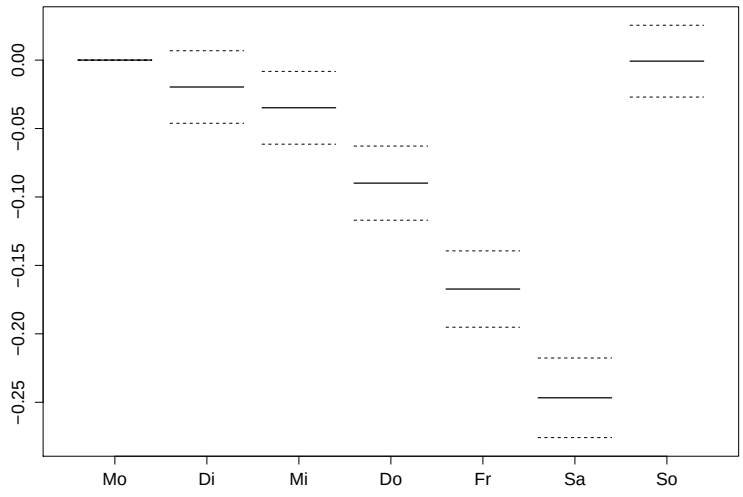

Abbildung 5.23: GAMM Jahresdaten: Effekte Wochentag

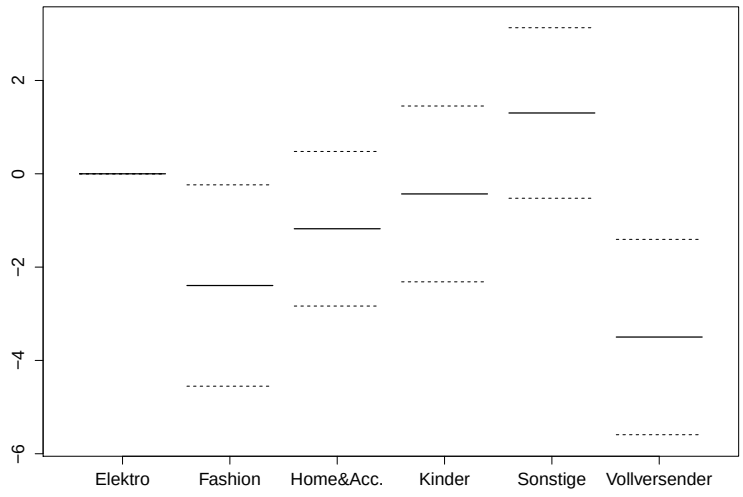

Abbildung 5.24: GAMM Jahresdaten: Effekte Branche

5.4.3 Modelldiagnose

Beim Modell über die Jahresdaten wird eine zufriedenstellende Güte erreicht. Das adjustierte Bestimmtheitsmaß liegt bei 86,3%, d.h. der größte Anteil der Streuung wird durch das Modell erklärt. Auch der Anteil der erklärten Devianz ist mit 94,4% recht hoch.
Wie in den meisten Anwendungen von Zählmodellen liegt auch hier Überdispersion vor. Der Dispersionsparameter wird auf 2,878 geschätzt. Folglich kann die Varianz in

den Daten durch das Modell nur unzureichend erklärt werden. Ein möglicher Grund hierfür können zum Beispiel Ausreißer in den Daten sein. Es gibt bei einigen Partnerschaften immer wieder hohe Bestellzahlen, deren Hintergründe sich aus den Daten nicht nachvollziehen lassen. Gründe für kurzfristige Steigerungen könnten z.B. Rabattaktionen oder neue Kampagnen sein. Diese sprunghaften Änderungen innerhalb von Partnerschaften lassen sich nicht durch Random Effects auffangen.
Die Residualplots (Anhang Abbildung B.3) weisen Verletzungen der Modellannahmen im akzeptablen Bereich auf. Der QQ-Plot (oben links) und das Histogramm der Residuen (unten links) zeigen Probleme mit Normalverteilungsannahme auf. Der QQ-Plot weist auf eine symmetrische, aber leptokurtische Verteilung hin. Schwach negative Residuen haben hier im Vergleich zur Normalverteilung eine zu hohe Wahrscheinlichkeitsmasse. Die Probleme könnten auch von der Vielzahl an Null-Beobachtungen rühren. Mit ca. 25% Nullbeobachtungen ist hier aber noch kein Anlass gegeben, ein Zero-Inflated Model anzusetzen. Die Streuung der Residuen um die Null ist auf den ersten Blick in Ordnung, auch die Annahme Varianzhomogenität sollte halten (rechts oben). Der Vergleich zwischen beobachteten und gefitteten Werten (rechts unten) zeigt einen guten Modellfit. Bis auf einige Ausreißer gruppieren sich alle Punkte der Tupel um die Winkelhalbierende.

5.4.4 Modell ohne Interaktionen

Um die Effekte für das „Gesamtgeschäft“ zu identifizieren, wurde das obige Modell nochmals ohne Interaktionseffekte gefittet. Da sich die teils gegenläufigen saisonalen Effekte der einzelnen Branchen neutralisieren, nimmt die Stärke der Effekte deutlich ab (vergleiche Abbildung 5.25). Im Gesamtmarkt bleibt der relative Kauferfolg im Jahresverlauf nahezu konstant. Ein leicht positiver Ausschlag ist wiederum für das Weihnachtsgeschäft zu beobachten. Unterdurchschnittliche Rates werden in den Monate Juni/Juli bzw. September/Oktober erwartet.

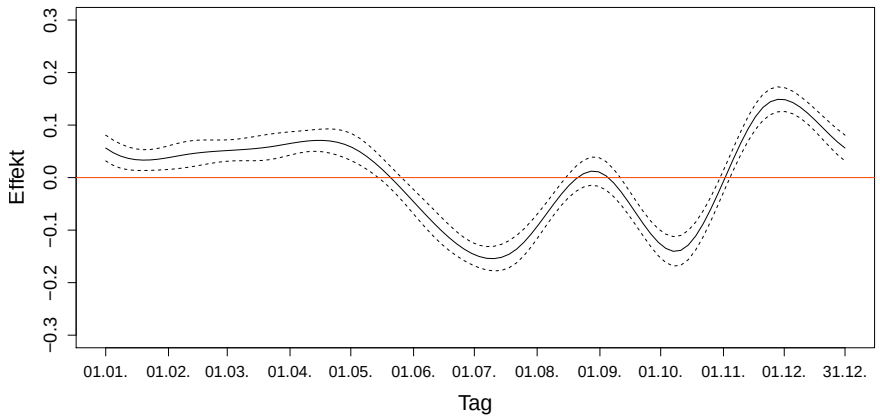

Abbildung 5.25: GAMM Jahresdaten: Zeitliche Effekte alle Advertiser

5.5 Verweildauer zwischen Klick und Order

5.5.1 Datengrundlage

Nicht jeder Sale im Affiliate Marketing erfolgt unmittelbar nach dem Klick auf ein Werbemittel. Oftmals nimmt der Entscheidungsprozess beim Käufer längere Zeit in Anspruch. Durch erneutes Bewerben desselben Produkts könnte der Entscheidungsprozess beeinflusst werden. In den Daten des Netzwerkes wird zu jeder Order die zugehörige Klickzeit erfasst. So kann die Verweildauer zwischen Wahrnehmung eines Produkts und dem tatsächlichen Kauf ausgewertet werden. Die Wirksamkeit einer Werbeschaltung beginnt in diesem Fall, wenn der Internetnutzer auf ein Werbemittel klickt und sich Informationen über ein Produkt einholt. Eine wichtige Fragestellung ist, wieviele Kunden sofort nach dem Klick auf das Werbemittel kaufen und wieviele sich erst später für einen Kauf entscheiden.
Hat der Kunde schon vorher einmal auf das Werbemittel geklickt, dann ist auch dieser Zeitpunkt in den Daten dokumentiert. Mehrfache vorherige Klicks lassen sich allerdings nicht identifizieren. Hier liegen somit linkstrunkierte Daten vor. Es ist nicht identifizierbar, welche der vorliegenden Beobachtungen trunkiert sind. Theoretisch könnten das alle sein, bei denen zwei Klicks erfasst wurden. In Abbildung 5.26 werden diese Fälle veranschaulicht. Bei Fall 1 wird die genaue Lebensdauer beobachtet, da es nur einen Klickzeitpunkt gibt. Bei den Fällen 2a und 2b hingegen wurden zwei Klickzeitpunkte erfasst. Man weiß nicht, ob der 1. Klick eines Users die tatsächliche Beobachtung ist (2a) oder eine trunkierte Beobachtung vorliegt (2b).

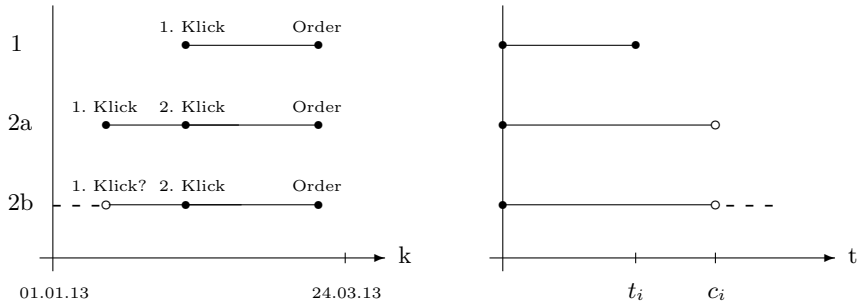

Abbildung 5.26: Verweildauern mit Kalenderzeit (links) und als rechtszensierte Verweildauern

Aufgrund dessen müssen beide Fälle als trunkierte Beobachtungen behandelt werden. Im weiteren werden die Fälle 2a und 2b wie rechtszensierte Beobachtungen behandelt. Es werden also Start- und Ereigniszeitpunkt vertauscht. Dies ist daher möglich, da es in den Daten keine echte Rechtszensierung gibt. Hier werden nur User betrachtet, die eine Order getätigt haben. Es gibt also für jedes Individuum einen Ereigniszeitpunkt und es kommt nicht zu Rechtszensierungen. Die Vertauschung von Start- und Ereigniszeitpunkt bietet sich daher an.
Im Folgenden werden die Cookie Ages bei Eingang einer Order mit Methoden der Verweildaueranalyse (R-Package `survival`) untersucht. Dazu wurde wiederum der

Orders Datensatz mit identischen Bereinigungen verwendet. Außerdem wurden für die Analysen nur die Werbemitteltypen Bannerlink und Textlink betrachtet, da die übrigen Typen nur einen kleinen Anteil ausmachen (3126 Beobachtungen) und die Interpretation der Ergebnisse damit erschwert würde.

5.5.2 Survival- und Hazardfunktion

Zunächst werden nur Sales betrachtet, um Aussagen über die Länge des Entscheidungsprozesses beim Kauf treffen zu können. Es ist zweckmäßig, sich zuerst darüber Gedanken zu machen, wie Hazard- und Survivalfunktion im vorliegenden Beispiel zu interpretieren sind. Der Ereigniszeitpunkt ist hier der Zeitpunkt des Kaufs. Damit ist die Hazardrate die Chance für einen Kauf im nächsten Moment, wenn er bis dahin nicht getätigt wurde. Die Survivalfunktion beschreibt dann die Rate der Internetnutzer, die zum Zeitpunkt t noch nicht gekauft haben und sich noch im Entscheidungsprozess befinden.

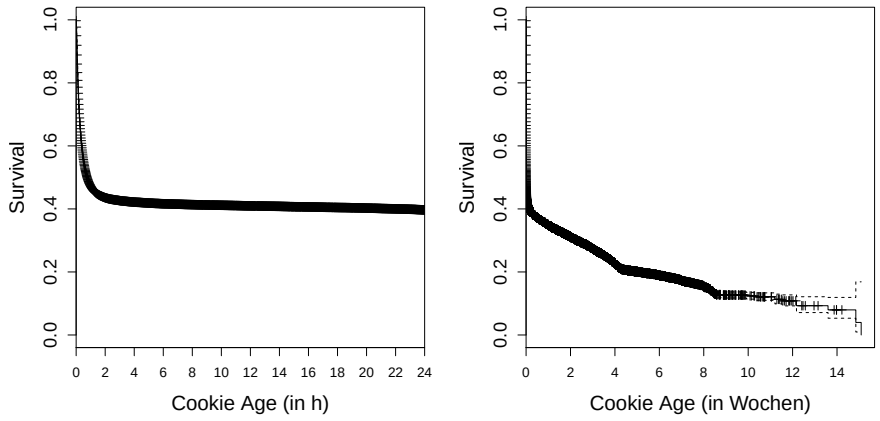

Abbildung 5.27: Eintägige und langfristige Survialfunktion

Die Kaplan-Meier-Kurve als Schätzer für die Survivalfunktion zeigt, dass ein Großteil der Käufe bereits innerhalb der ersten 24 Stunden nach dem Klick erfolgen. Der Anteil der Käufe innerhalb eines Tages liegt bei rund 60%. Betrachtet man die geschätzte Survivalfunktion für die ersten 24 Stunden nach dem Klick (Abbildung 5.27 links), zeigt sich ein schneller Abfall in den ersten 2 Stunden. Danach wird die Funktion schnell flacher. Das bedeutet, dass es sich bei über der Hälfte der Käufe um sofortige Käufe handelt, während bei den anderen Käufen der Entscheidungsprozess länger als einen Tag dauert. Bei der langfristigen Kaplan-Meier-Kurve ist ein Abknicken nach vier Wochen und acht Wochen erkennbar. Diese Strukturbrüche haben wohl technische Hintergründe. Viele Cookies besitzen sogenannte Lifetimes, d.h. sie verfallen nach einer bestimmten Zeit. Das Abknicken findet wohl an Zeitpunkten statt, an denen eine Vielzahl von Cookies ausläuft und sich die Klicks nicht mehr zuordnen lassen. Die wahre Kurve wird also glatter verlaufen. Dabei wurden vier Wochen nach dem Klick schon rund 80% aller Käufe getätigt, für den Fall, dass der Klick zum Kauf führte.

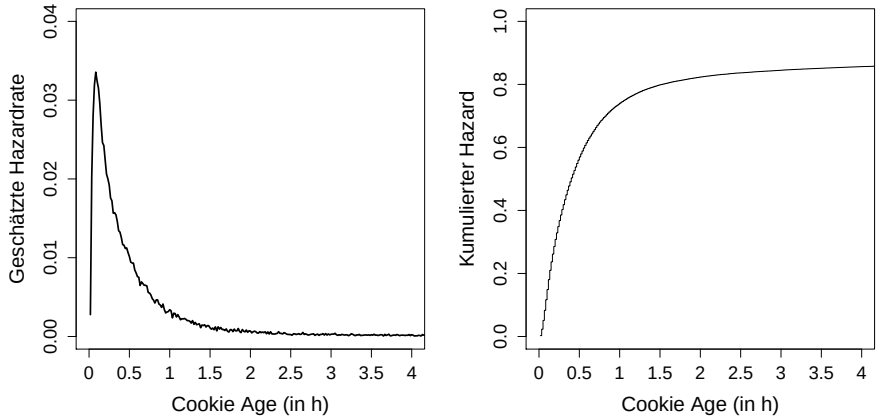

Abbildung 5.28: Geschätzte Harzardrate und kumulierte Hazard für Verweildauern bis zu 4 Stunden

Die hohe Kaufrate unmittelbar nach dem Klick lässt sich besonders gut mit der Hazardrate veranschaulichen. Abbildung 5.28 zeigt die geschätzte Harzardrate und den kumulierten Hazard in den ersten vier Stunden nach dem Kauf. Hier zeigt sich eine erhöhte Rate in den ersten 1.5 bis 2 Stunden nach dem Klick auf das Werbemittel. Die Hazardrate pendelt sich für längere Verweildauern auf einem konstant niedrigem Niveau ein.
Eine branchenweise Differenzierung der Kaplan-Meier-Kurve zeigt kurzfristig Unterschiede bei der Dauer der Kaufentscheidung (Abbildung 5.29). In den ersten zwei Stunden nach dem Klick auf das Werbemittel zeichnet sich ein längerer Entscheidungsprozess bei Käufen in den Advertiser-Branchen Home & Accessoires und Vollversendern ab. Bei den Advertisern anderer Branchen erfolgt eine kurzfristige Kaufentscheidung. Hier liegt der Anteil bereits getätigter Käufe nach 2 Stunden schon bei über 50%. Bei Käufen in der Branche Elektro wird die 50%-Hürde schon binnen einer halben Stunde nach dem Kauf geknackt. Käufer in der Branche Fashion nehmen sich dagegen mehr Zeit beim Online-Shopping. Hier flacht die Survivalkurve nach circa einer Stunde ab.
Abbildung 5.30 zeigt die Kaplan-Meier-Schätzer, differenziert nach Publisher Business Model. Aufgrund des Postview-Trackings bei Media-Publishern wurden diese aus der Analyse generell ausgeschlossen. Innerhalb der ersten zwei Stunden nach dem Klick wird sehr häufig gekauft, wenn der User auf einer Website aus dem Business Model Cash Back oder Preisvergleich abgesprungen ist. Bei mehr als 50% der Käufer, die vorher eine Cash Back Site besucht haben, findet der Kauf innerhalb der ersten 30 Minuten nach dem Klick statt. Das gilt auch annähernd für User von Preisvergleichsseiten, die sich zu einen Kauf entschließen. Bei den Besuchern von Coupon-, Topic-, oder Portal & Communities-Websites, haben zwei Stunden nach dem Klick deutlich weniger Käufer diesen Kauf bereits getätigt. Unabhängig vom Business Model haben aber innerhalb von zwei Stunden nach dem Klick auf ein Werbemittel schon ein Großteil der Käufer den Kauf bereits getätigt.

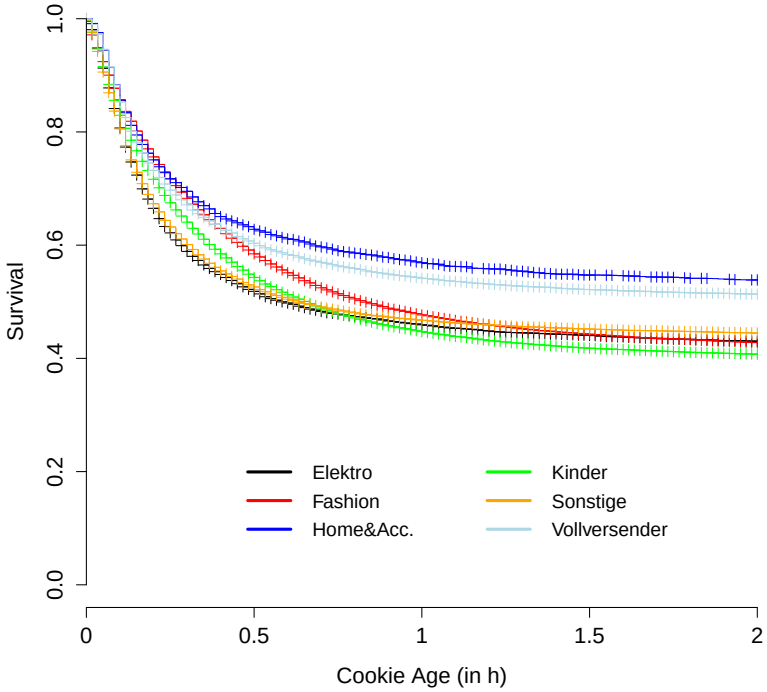

Abbildung 5.29: Kaplan-Meier-Schätzer branchenweise (2 Stunden)

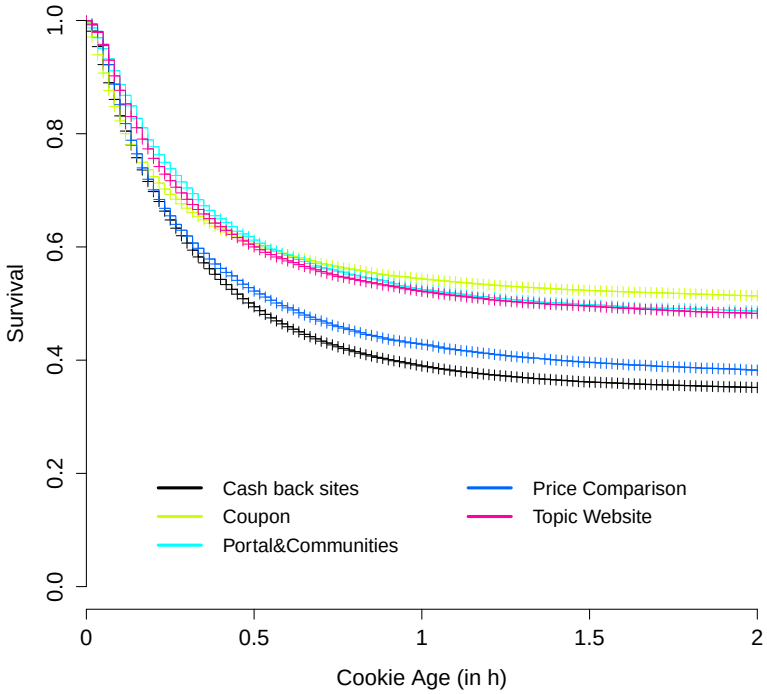

Abbildung 5.30: KM-Schätzer nach Business Model (2 Stunden)

Nimmt man alle Orders, sowohl Sales als auch Leads, in die Analyse auf, zeigt sich, dass Leads deutlich schneller nach dem Klick getätigt werden. Nach 2 Stunden haben lediglich rund 30% der Nutzer ihren Lead noch nicht abgeschlossen, vergleiche Abbildung 5.31. Mittels des Log-Rank-Tests wurde überprüft, ob die Abweichungen in den Survivalkurven statistisch signifikant sind. Bei den gezeigten Differenzierungen nach Business Model, Branche und Order Type wird für den Log-Rank-Test jeweils ein p-Value nahe Null ausgegeben. Die Nullhypothese, dass in allen Gruppen dieselbe Hazardrate vorliegt, kann also jeweils hochsignifikant verworfen werden.
Ähnlich zu den Generalisierten Linearen Modellen kann auch die Verweildauer in Abhängigkeit von Kovariablen modelliert werden. Einen solchen Ansatz für zensierte Daten leistet das Proportional Hazards Modell von Cox (1972). Bei den Bemühungen eine Cox-Regression durchzuführen hat sich gezeigt, dass die Verweildauern zwischen Klick und Orders mit den zur Verfügung stehenden Variablen nicht hinreichend beschrieben werden können. Zum Fitten eines solchen Regressionsmodells wären weitere, aussagekräftige Einflussfaktoren notwendig. Aufgrund dessen werden die Verweildauern hier lediglich mittels Nelson-Aalen- und Kaplan-Maier-Schätzer beschrieben.

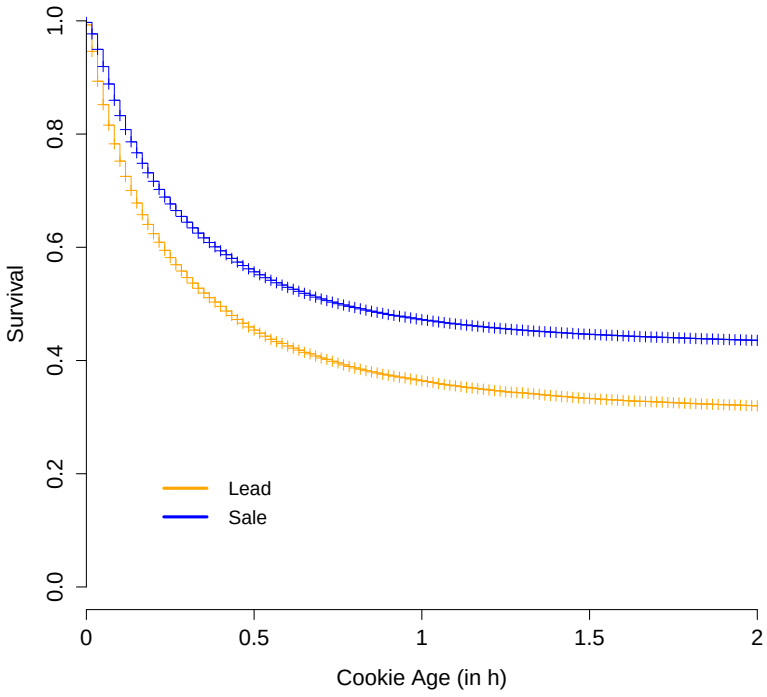

Abbildung 5.31: KM-Schätzer nach Order Type (2 Stunden)

5.6 Warenkorbwerte im Zeitverlauf

5.6.1 Modell und Datengrundlage

Für die Vergütung zwischen Advertiser und Publisher ist der Warenkorbwert beim Sale maßgeblich. Die Publisher werden anteilig zum Warenkorbwert, der beim Advertiser generiert wurde, vergütet. Fraglich ist, ob es im Zeitverlauf Unterschiede in der Höhe des durchschnittlichen Warenkorbwerts gibt. Eine intuitive Hypothese wäre zum Beispiel, dass der durchschnittliche Warenkorbwert zum Monatsende hin geringer ist. Die Vermutung liegt nahe, da viele Arbeitnehmer ihr Gehalt am Monatsersten erhalten und größere Anschaffungen aufschieben. Da auch viele Gehälter zur Monatsmitte gezahlt werden, ist hier dieselbe Vermutung zu prüfen. Des weiteren wird der Frage nachgegangen, ob die mittleren Warenkorbwerte im Tagesverlauf gleichverteilt sind oder zu bestimmten Tageszeiten bevorzugt hochwertige Produkte eingekauft werden.
Die Warenkorbwerte werden auf Grundlage der Orders-Daten analysiert. Zu jedem Sale ist die Höhe des Warenkorbwerts dokumentiert. Nach Filterung auf die Werbemittel Banner- und Textlink, sowie die bisher analysierten Advertiser-Branchen und Publisher-Geschäftsmodelle stehen für die Analysen rund 300.000 Beobachtungen aus dem Zeitraum von 28.01.2013 bis 24.03.2013 zur Verfügung. Die Differenz zu den Orders-Daten aus der deskriptiven Analyse rührt in erster Linie aus der Nichtbeachtung der Leads her. Außerdem wurden sehr hohe Warenkorbwerte (> 1000 EUR) und geringe Bestellwerte (< 5 EUR) ausgeschlossen (ca. 5500 Beobachtungen).
Im vorliegenden Fall wird der Einfluss von Kovariablen auf die Zielgröße „Warenkorbwert in Euro“ geprüft. Die zeitlichen Effekte sollen als glatte Funktionen aufgenommen werden, weshalb wiederum ein Generalisiertes Additives Modell zum Einsatz kommt. Die Verteilung der Zielgröße (Abbildung 3.8) mit nicht-negativem Wertebereich und Rechtsschiefe spricht für die Modellierung mittels Gammaverteilung oder inverser Normalverteilung (Madsen und Thyregod, 2011). Nach BIC liefert hier die inverse Normalverteilung (Inverse Gaussian) eine bessere Anpassung. Dann gilt für den Warenkorbwert

$$W \sim \mathrm{IG}(\mu, \lambda) \ ,$$

mit $\mu, \lambda > 0$. Außerdem wird statt der natürlichen Linkfunktion $\eta = 1/\mu^2$, wieder der Log-Link verwendet. Die multiplikativ-exponentiellen Effekte sind für die Interpretation des Modells besser handzuhaben (vgl. Fahrmeir, Kneib und Lang, 2009). Die Variablen wurden mithilfe des BIC selektiert. Die Tageszeit (in Minuten) und das Datum des Tages werden als glatten Effekte in das Modell aufgenommen. Der erwartete Warenkorbwert ist dann

$$w_{it} = \exp\{\eta_{it} + f_{t|x_B}(min_t) + f_{t|x_B}(d_t)\} + \epsilon_{it} \ , \tag{5.13}$$

für alle Käufe $i = 1, \ldots, n$. Im Modell wird der Zeitpunkt t in die Tageszeit, gemessen in Minuten min_t, und Tag im Monat d_t zerlegt. Außerdem hängen die zeitlichen Effekte über Interaktionen von den Advertiser-Branchen ab. Der lineare Prädiktor ergibt sich zu

$$\begin{aligned}\eta^k =& \beta_0 + \beta_1 x_{\text{Branche}} + \beta_2 x_{\text{AdvAccountManager}} + \beta_3 x_{\text{PubAccountManager}} \\ &+ \beta_4 x_{\text{BusinessModel}} + \beta_5 x_{\text{AdvGröße}} + \beta_6 x_{\text{Wochentag}} + \beta_7 x_{\text{Voucher}} \\ &+ \gamma_{0,\text{Advertiser}} \ .\end{aligned}$$

Es wird angenommen, dass die Advertiser im Bezug auf das Niveau des Warenkorbwertes verschiedene Cluster bilden. Daher werden die 75 Advertiser als zufällige Effekte $\gamma_{0,i} \sim N(0, \tau_0^2)$ einbezogen.

5.6.2 Parameterschätzer

Betrachtet man die glatten Funktionen in den Abbildungen 5.32 und 5.33, dann fällt vor allem die Schwäche der Effekte auf. Bei den meisten Branchen ist wohl abends und in den Nachtstunden der erwartete Warenkorbwert leicht erhöht. Hochsignifikante Effekte im Tagesverlauf ergeben sich nur in den Branchen Elektro-, Kinderprodukte und den Sonstigen. Auch im Monatsverlauf sind die geschätzten Effekte eher schwach. Die größten Schwankungen des erwarteten Warenkorbwerts liegen hier bei den Elektro-Advertisern vor. Aufgrund der schwachen Effekte im Zeitverlauf muss man davon ausgehen, dass die Warenkorbwerte sowohl im Tages- als auch im Monatsverlauf konstant sind und nur zufälligen Schwankungen unterliegen.

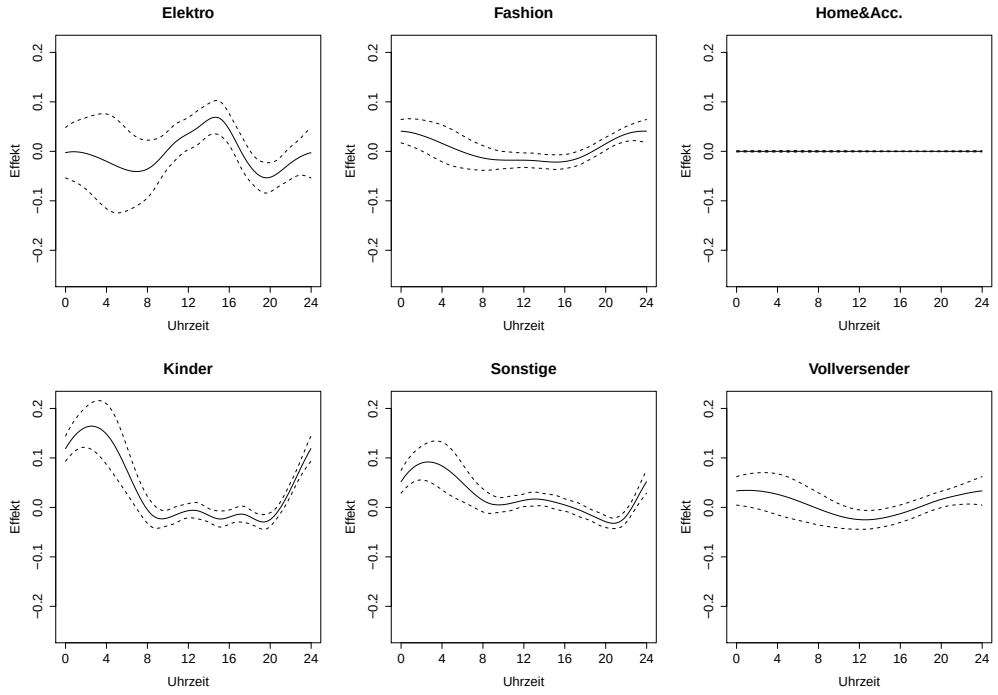

Abbildung 5.32: Effekte auf Warenkorbwert im Tagesverlauf

Über die Wochentage ergeben sich Unterschiede in der Höhe des Warenkorbwerts. Freitags, samstags und sonntags werden leicht erhöhte Bestellwerte erwartet. Die stärksten Effekte auf den Warenkorbwert hat jedoch die Branchenzugehörigkeit der Advertiser. In den Branchen Elektro und Home & Accessoires wird der höchste, in den Branchen Kinderprodukte und den Sonstigen der niedrigste durchschnittliche Warenkorbwert geschätzt.
Gelangt der Kunde über einen Publisher mit Account Manager auf die Website des Advertisers, dann generiert er im Mittel einen höheren Bestellwert. Auch für das Business Model des Publishers ergeben sich signifikante Effekte. Für die Nutzer

von Topic Websites werden demnach die höchsten Warenkorbwerte erwartet. Beim Einlösen eines Vouchers liegt der erwartete Bestellwert um den Faktor 1.11 höher als bei normalen Online-Käufen. Auffällig ist, dass abgesehen von den Brancheneffekten, die Effekte generell eher schwach sind (vergleiche Koeffizienten in Anhang A.5 und A.6). Keine signifikanten Einflüsse haben die Advertiser-Größe und das Vorliegen eines Key Accounts bei den Advertisern.

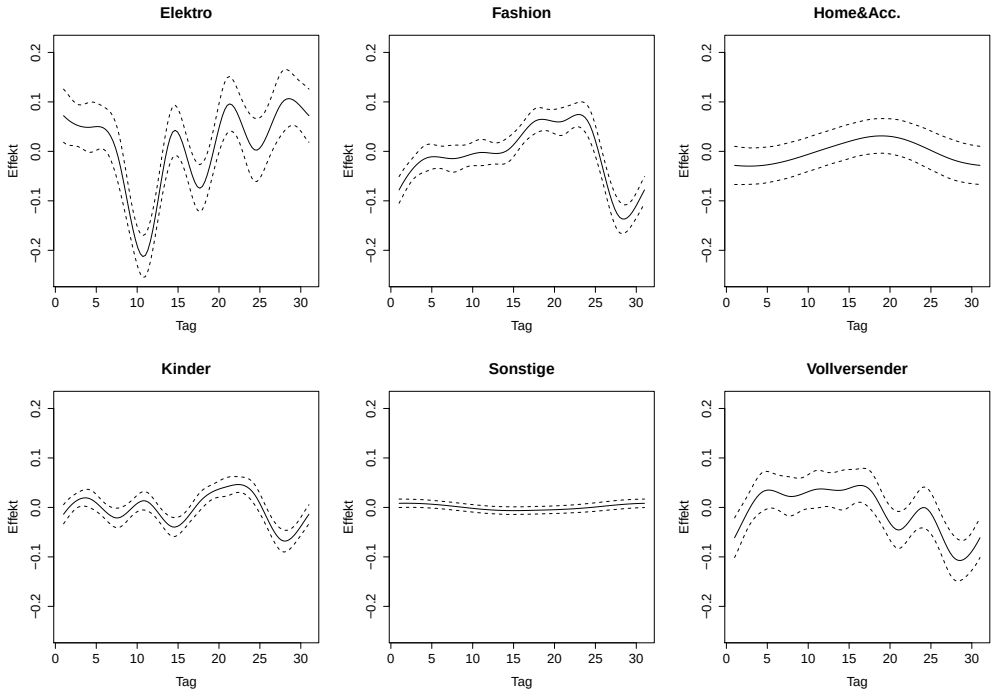

Abbildung 5.33: Effekte auf Warenkorbwert im Monatsverlauf

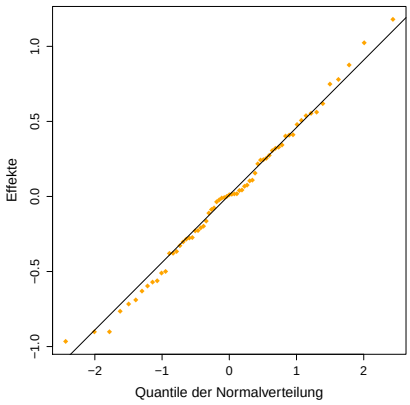

Abbildung 5.34: GAMM Warenkorbwert: Random Effects

5.6.3 Modelldiagnose

Die Residualplots (Abbildung Anhang B.4) weisen erwartungsgemäß Schwierigkeiten mit den Modellannahmen auf. Vor allem die Vorhersage der Werte (Graphik unten rechts) ist unbefriedigend. Die gefitteten Werte sollten eigentlich gegen die Responsewerte gehen, d.h. die Punkte in der Graphik sollten sich um die Winkelhalbierende konzentrieren. Im Modell gibt es eine Vielzahl an Ausreißern, sprich, die Vorhersage der Warenkorbwerte durch das Modell funktioniert nicht besonders gut. Hohe Warenkorbwerte werden vom Modell überhaupt nicht erfasst.
Diese Probleme sind wohl darin begründet, dass nicht alle relevanten Einflussfaktoren vorliegen. Den wichtigsten Einfluss auf den Warenkorbwert hat intuitiv die Art des gekauften Produkts. Es ergibt sich beispielsweise ein erheblicher Unterschied, ob bei einem Elektro-Advertiser ein Fernsehgerät oder geringwertiges PC-Zubehör gekauft wird. Im Datensatz liegt keine Information vor, welches Produkt gekauft wurde. Mit den zur Verfügung stehenden Informationen kann jedoch keine ausreichende Unterscheidung getroffen werden. Die Prognoseleistung des Modells leidet darunter. Die Problematik wird auch durch das adjustierte Bestimmtheitsmaß von 25.6% verdeutlicht. Der Großteil der Streuung kann nicht durch das Modell erklärt werden. Da mit dem Modell in erster Linie das Ziel verfolgt wird, zeitliche Effekte zu identifizieren, und es nicht zur exakten Prognose von Warenkorbwerten eingesetzt werden soll, wurde trotz der Güteprobleme am Modell festgehalten.
Die Verteilung der Residuen geht prinzipiell in Ordnung. Die Verteilung ist symmetrisch, jedoch nach links verschoben. Der QQ-Plot zeigt Abweichungen von der Normalverteilung an. Die Residuen streuen um die Null. Für zunehmende Prognosewerte nimmt die Streuung ab.
Die Advertiser als Random Effects liefern eine gute Gruppierung. Der QQ-Plot für die zufälligen Effekte (Abbildung 5.34) diagnostiziert eine sehr gute Anpassung an die Normalverteilung.

5.7 Klick-/Ordervolumen Wochentage

Die deskriptive Analyse hat beim Klick- und Ordervolumen Unterschiede an den einzelnen Wochentagen aufgezeigt (vergleiche Tabellen 3.3 und 3.7). Es soll getestet werden, ob die Ergebnisse statistisch signifikant sind und nicht rein zufällig zustande kommen. Dazu wurde ein χ^2-Anpassungstest durchgeführt. Er prüft in diesem Fall die Nullhypothese, dass die Anzahlen an Klicks bzw. Orders vom Wochentag unabhängig sind. Bei Vorliegen perfekter Unabhängigkeit würde man an jedem Wochentag die gleiche Anzahl an Orders bzw. Klicks beobachten.
Die p-Values des χ^2-Tests sowohl für die Orders, als auch die Klicks gehen gegen Null. Die Unabhängigkeitsannahme kann daher hochsignifikant verworfen werden. Das stützt die Hypothese, dass die Anzahlen an Klicks und Orders im Affiliate Marketing vom jeweiligen Wochentag abhängig sind. Vor allem an Sonntagen werden zahlreiche Klicks und Orders getätigt. Die Modelle haben gezeigt, dass sonntags nicht nur in Absolutzahlen die höchste Nutzeraktivität vorliegt, sondern auch die Rates signifikant erhöht sind. Hinter dem Wochendurchschnitt bleibt dagegen der Traffic an Freitagen und Samstagen zurück. Anscheinend sind die Internetnutzer zum Ende der Arbeitswoche gegenüber Werbemitteln weniger sensibel.

Kapitel 6

Zusammenfassung und Ausblick

6.1 Zusammenfassung der Ergebnisse

Die Analyse der Daten hat wiederkehrende zeitliche Strukturen im Affiliate Marketing bei Online Retailern aufgezeigt. Sowohl Absolutwerte als auch Conversion Rates unterliegen im Tagesverlauf beträchtlichen Schwankungen. In den Abendstunden zwischen 19.00 und 22.00 Uhr werden nicht nur die meisten Klicks und Orders im Online Retail beobachtet, sondern auch die erwarteten Erfolgsraten der Werbemittel sind in der Regel signifikant erhöht. Neben der Nutzeraktivität sinken in den Stunden nach Mitternacht auch die geschätzten Conversions beträchtlich. Die Vermutung von höheren Conversions zur Mittagszeit kann hingegen nicht bestätigt werden.
Weniger bedeutend werden die Erfolgsraten davon beeinflusst, um welchen Zeitpunkt im Monatsverlauf es sich handelt. Die beobachteten Effekte sind hier schwach und es empfiehlt sich, nach den einzelnen Partnerschaften zu differenzieren. Eine allgemeingültige Aussage über das Vorliegen eines Monatseffekts lässt sich nicht treffen.
In den Analysen hat sich gezeigt, dass sich der Erfolg von Online-Werbekampagnen sehr stark an den Kenngrößen der einzelnen Partnerschaften festmacht. Die erzeugten Rates sind über die Partnerschaften hinweg durchaus heterogen und vermutlich stark von den beteiligten Affiliates abhängig.
Die Erfolgsraten im Affiliate Marketing unterliegen zunächst moderaten jahreszeitlichen Schwankungen. Eine deutliche Steigerung von Orders im Weihnachtsgeschäft geht im Gesamtmarkt mit einer Erhöhung der Rates einher. Auch bei den saisonalen Effekten kann von branchenspezifischen Einflüssen ausgegangen werden. Zum Beispiel bleibt bei den Advertisern für Kinderprodukte der positive Effekt auf die Rates in der Vorweihnachtszeit aus. In manchen Branchen sind darüber hinaus vermehrte Orders im Frühjahr zu erkennen.
Der Erfolg von Produktwerbung im Online-Marketing wird nach den hier erzielten Erkenntnissen vom Wochentag der Internetnutzung beeinflusst. An Sonntagen lässt sich ein deutlicher Überhang an Orders und Klicks im Vergleich zu den übrigen Wochentagen erkennen. Auch die Kaufbereitschaft, gemessen an den Erfolgsraten, ist sonntags signifikant gesteigert. Gegensätzliche Effekte wurden für die Effekte des Online-Marketings an den Wochentagen Freitag und Samstag identifiziert. An diesen Tagen bleiben die erwarteten Klicks und Orders, resultierend aus Ad-Impressions, deutlich hinter dem Wochendurchschnitt zurück.

Die Analysen lassen des weiteren darauf schließen, dass die durchschnittlichen Warenkorbwerte bei den Advertisern im Affiliate Marketing nicht in einen zeitlichen Kontext zu setzen sind. Sowohl im Tages- als auch Monatsverlauf verhalten sich die beobachteten Warenkorbwerte relativ konstant, es gibt allenfalls advertiser- oder branchenspezifische Effekte im Zeitverlauf.
Kaufentscheidungen fallen im Online-Marketing offenbar recht kurzfristig. Die meisten Käufer bestellen bereits innerhalb von 24 Stunden nach dem Klick auf ein Werbemittel, sehr häufig sogar innerhalb von zwei Stunden. Es ist also davon auszugehen, dass die Wahrscheinlichkeit für einen Kauf nach dem Klick auf ein Werbemittel mit verstreichender Zeit sinkt. Bei diesem Ergebnis ist jedoch zu berücksichtigen, dass nur User mit einmaligem Klick auf ein Werbemittel analysiert werden konnten. Bei mehrfachen vorangegangenen Klicks muss wohl von längeren Entscheidungsprozessen ausgegangen werden.
In den deskriptiven Analysen haben sich Unterschiede der zeitlichen Aspekte nach Partnerschaften gezeigt. Um diesen Diskrepanzen Rechnung zu tragen, wurden die zeitlichen Verläufe jeweils nach den aussagekräftigen Advertiser-Branchen differenziert. Es zeigen sich nicht selten branchenspezifische Eigenschaften und Abweichungen vom Gesamtmarkt. Alle branchenspezifischen Besonderheiten herauszuarbeiten, würde wohl den Rahmen sprengen und sollte vielmehr durch eine gezielte Analyse ausgewählter Partnerschaften erfolgen.
Neben den zeitlichen Einflüssen wurden in den Modellen noch einige andere Einflüsse auf die Erfolgsraten identifiziert. Bei vermehrter Ausbringung von Werbemitteln setzt anscheinend eine Sättigung ein. So lässt sich für eine steigende Anzahl an Ad-Impressions und Klicks eine abnehmende Erfolgsrate beobachten. Während sich die Diskrepanz zwischen selten und häufig gezeigten Werbemitteln mit erhöhten Standardfehlern erklären lässt, ist anzunehmen, dass Spitzenwerte in der Werbemittelschaltung gegenüber mittlerer Ausbringung keine Verbesserung der Rates liefern. Im Allgemeinen schneiden mittlere bis große Advertiser im Bezug auf die Erfolgsraten besser ab als der Long Tail-Bereich. Gesteigerte Erfolgsrate durch Brand Effects im Online-Marketing lassen sich nicht zurückweisen.

6.2 Kritik und Ausblick

Einige Eigenschaften der zur Verfügung stehenden Datengrundlage wirken sich restriktiv auf die Analysen aus. Aufgrund der Datenmenge wurden Stichproben aus der Grundgesamtheit gezogen. In den vorliegenden Analysen wurden nur Advertiser aus dem Bereich Online-Retail untersucht. Natürlich gibt es noch viele andere relevante Advertiser im Affiliate Marketing.
Immer dann, wenn Impressionszahlen für die Analysen unabdingbar waren, sprich bei der Modellierung der Conversions, musste seitens der Publisher auf Werbeträger des Business Models Media mit Key Account-Betreuung eingeschränkt werden. Die Einschränkung auf diese Websites führte zu einer nicht unerheblichen Reduzierung der Datengrundlage. Hier muss mit Implikationen auf die Allgemeingültigkeit der Ergebnisse gerechnet werden. Da sich der Werbeerfolg im Affiliate Marketing kaum losgelöst von der Anzahl der generierten Ad-Impressions bemessen lässt, erscheinen diese Restriktionen jedoch alternativlos. Für eingehendere Untersuchungen ist eine präzise Dokumentation der Ad-Impressions auf Stundenbasis unabdingbar.

Die bereitgestellten Daten wurden zu Abrechnungszwecken dokumentiert und wurden nicht mit dem Hintergrund einer späteren Datenanalyse erhoben. Daher erfolgte die Aufzeichnung nach dem Verursachungsprinzip. In der Datenbasis wird nur dann ein Eintrag erzeugt, wenn tatsächlich ein Klick oder eine Order stattgefunden hat. Unklar bleibt bei dieser Zählweise, woraus 0-Beobachtungen resultieren. Es ist fraglich, ob in der betrachteten Partnerschaft Ad-Impressions erzeugt wurden und die Werbekampagne schlichtweg erfolglos war oder ein Werbemittel im betreffenden Zeitraum überhaupt nicht ausgebracht wurde.
Für die Untersuchung von Kausalzusammenhängen wäre eine personenbasierte Erhebung der Daten von großem Vorteil. Somit wäre nachvollziehbar, an welchem Punkt der Customer Journey im Affiliate Marketing der User verloren geht und es könnten unter Umständen Hintergründe für den Drop-out beleuchtet werden. Die erhobenen Sekundärdaten liegen in Teildatensätzen vor, über die sich der Weg eines Nutzers in der Wertschöpfungskette nicht nachvollziehen lässt. Aus diesem Grund musste stets die vereinfachende Annahme getroffen werden, dass die Zeitpunkte von Impression, Klick und Order eines Users identisch sind. Diese Annahme führt zu Verzerrungen in der Analyse von zeitlichen Aspekten.
Die Erfolgsraten im Affiliate Marketing werden höchstwahrscheinlich außer von den beschriebenen Einflussgrößen noch von einer ganzen Reihe anderer Faktoren gesteuert. Teilweise liegen diese nicht vor und können auch überhaupt nicht dokumentiert werden, zum Beispiel aus Datenschutzgründen, wenn es sich um personenbezogene Daten handelt. In anderen Fällen wurden zwar relevante Einflussgrößen erhoben, aber durch vorzeitige Löschung oder Aggregierung für die Analysen unbrauchbar gemacht. Zum Beispiel wird die wertvolle Information, ob es sich um einen Internetzugriff über Desktop oder Mobile Devices handelt, nur kurzfristig gespeichert. So lassen sich die Zielgrößen in den Modellen nicht immer ausreichend durch die zur Verfügung stehenden Einflussgrößen beschreiben.
In der vorliegenden Arbeit stellen die zeitlichen Aspekte den primären Untersuchungsgegenstand dar. Informationen zeitlicher Natur waren dabei in ausreichendem Maße zugänglich. Trotz getroffener Annahmen und Einschränkung der Datengrundlage können allgemeine und branchenspezifische zeitliche Effekte im Online-Marketing aufgezeigt werden. Für weiterführende Untersuchungen empfiehlt sich eine Anpassung der Datenerhebung in Richtung einer verstärkten Erfassung von Ad-Impressions und Nutzer-basierter Datendokumentation. Auf diese Weise kann der Weg eines Users im Affiliate Marketing genauer nachvollzogen werden und es können, basierend auf der vorliegenden Arbeit, weitere Fragestellungen analysiert werden.

Literaturverzeichnis

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

BVDW (2013). OVK Online Report. Website. Erhältlich auf http://www.bvdw.org/medien/ovk-online-report-2013-01—kostenfreier-download?media=4611; besucht am 05.08.2013.

Cox, D. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society (B)* **34**, 187–220.

Cox, D. und Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Eilers, P. H. C. und Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science 11*(2), 89–121.

Fahrmeir, L., Kneib, T., und Lang, S. (2009). *Regression*. Berlin: Springer.

Fahrmeir, L., Künstler, R., Pigeot, I., und Tutz, G. (2010). *Statistik*. Berlin: Springer.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. London: Sage.

Hastie, T. und Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Internet World Business (2010). Klickraten stabilisieren sich. *Internet World Business* **24**, 2.

Internet World Business (2013). Das ist wie Kindergarten - Gutscheinmarketing hat in Deutschland noch mit Schwierigkeiten zu kämpfen. *Internet World Business* **7**, 10.

Kleinbaum, D. G. und Klein, M. (2005). *Survival Analysis*. New York: Springer.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Madsen, H. und Thyregod, P. (2011). *Introduction to General and Generalized Linear Models*. London: Chapman & Hall.

McCullagh, P. und Nelder, J. A. (1989). *Generalized Linear Models* (zweite Aufl.). New York: Chapman and Hall.

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

Pinheiro, J. und Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Ramsay, J. O., Wickham, H., Graves, S., und Hooker, G. (2013). *fda: Functional Data Analysis*. R package version 2.3.8.

Rigby, R. A. und Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics* **54**, 507–554.

Ruppert, D., Wand, M. P., und Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Therneau, T. M. (2013). *A Package for Survival Analysis in S*. R package version 2.37-4.

Therneau, T. M. und Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Webel, K. und Wied, D. (2012). *Stochastische Prozesse*. Wiesbaden: Gabler.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society (B) 62*(2), 413–428.

Wood, S. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B) 65*(1), 95–114.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99**, 673–686.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73*(1), 3–36.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A

Schätzer GAMM

A.1 Parameterschätzer GAM Impression-to-Order

Kategoriale Variablen (Log-Link)				
Kovariable/Kategorie	$\hat{\beta}_i$	Standardabw.	t-Value	p-Value
(Intercept)	-5.89570	0.300310	-19.6321	6.052e-85
Wochentag (Referenz: Montag)				
Dienstag	-0.11720	0.026054	-4.4981	6.898e-06
Mittwoch	-0.11924	0.026249	-4.5427	5.589e-06
Donnerstag	-0.13185	0.026616	-4.9540	7.336e-07
Freitag	-0.17032	0.026752	-6.3666	1.979e-10
Samstag	-0.21313	0.027293	-7.8090	6.072e-15
Sonntag	0.04755	0.025162	1.8896	5.882e-02
Branche (Referenz: Elektro)				
Fashion	-0.02151	0.037915	-0.5673	5.705e-01
Home & Acc.	0.05442	0.050456	1.0786	2.808e-01
Kinder	0.67489	0.099871	6.7576	1.445e-11
Sonstige	2.06594	0.031404	65.7849	0.000e+00
Vollversender	-0.70266	0.118076	-5.9509	2.715e-09
Advertiser Account Manager (Referenz: Nein)				
Ja	1.03240	0.091180	11.3226	1.270e-29
Advertiser-Größe (Referenz: 100-500 EUR)				
500-1000 EUR	2.19089	0.270831	8.0895	6.358e-16
1000-2000 EUR	3.34815	0.270981	12.3557	6.240e-35
2000-5000 EUR	3.88975	0.269022	14.4589	3.989e-47
5000-10000 EUR	2.73744	0.269765	10.1475	3.933e-24
>10000 EUR	3.50517	0.269789	12.9923	1.996e-38
Glatte Effekte				
Funktion	Schätzer	Standardabw.	t-Value	p-Value
cs(log.Imp, df = 10)	-0.85542	0.007623	-112.2136	0.000e+00
Wahrscheinlichkeit Mischverteilung (Logit-Link)				
Funktion	Schätzer	Standardabw.	t-Value	p-Value
(Intercept)	-1.911	0.07236	-26.41	7.212e-151

Tabelle A.1: Parameterschätzer GAM ITO ohne Interaktionen

A.2 Parameterschätzer GAMM Conversion Rate

Kategoriale Variablen (Log-Link)					
Kovariable/Kategorie	$\hat{\beta}_i$	Standardabw.	t-Value	p-Value	
(Intercept)	-6.882523	8.653543	-0.795	0.426419	
Wochentag (Referenz: Montag)					
Dienstag	-0.040220	0.010401	-3.867	0.000110	***
Mittwoch	-0.035654	0.010442	-3.415	0.000639	***
Donnerstag	0.001583	0.010354	0.153	0.878498	
Freitag	-0.078888	0.010645	-7.411	1.28e-13	***
Samstag	-0.088248	0.010767	-8.196	2.54e-16	***
Sonntag	0.066818	0.010055	6.645	3.06e-11	***
Branche (Referenz: Elektro)					
Fashion	1.556679	5.493647	0.283	0.776902	
Home & Acc.	1.540219	7.218564	0.213	0.831040	
Kinder	2.207252	6.834154	0.323	0.746717	
Sonstige	1.591455	6.690518	0.238	0.811985	
Vollversender	2.025303	6.428920	0.315	0.752740	
Advertiser-Größe (Referenz: 0-100 EUR)					
100-500 EUR	0.288350	7.969350	0.036	0.971137	
500-1000 EUR	0.077454	9.038451	0.009	0.993163	
1000-2000 EUR	1.695003	7.675165	0.221	0.825216	
2000-5000 EUR	2.216425	7.665167	0.289	0.772464	
5000-10000 EUR	2.302752	8.772808	0.262	0.792947	
>10000 EUR	2.025201	11.901991	0.170	0.864888	
Glatte Effekte					
Funktion	edf	Ref.df	F	p-value	
s(Stunde):Elektro	7.635	8.000	981.64	< 2e-16	***
s(Stunde):Fashion	7.961	8.000	9512.06	< 2e-16	***
s(Stunde):Home & Acc.	7.608	8.000	97.16	7.90e-16	***
s(Stunde):Kinder	7.962	8.000	4430.01	< 2e-16	***
s(Stunde):Sonstige	7.893	8.000	9722.38	< 2e-16	***
s(Stunde):Vollversender	7.183	8.000	1739.55	< 2e-16	***
s(Tag):Elektro	2.519	3.000	95.46	< 2e-16	***
s(Tag):Fashion	2.956	3.000	8535.91	< 2e-16	***
s(Tag):Home & Acc.	2.243	3.000	8.19	0.0455	*
s(Tag):Kinder	2.964	3.000	231.53	8.05e-08	***
s(Tag):Sonstige	2.810	3.000	446.54	< 2e-16	***
s(Tag):Vollversender	2.400	3.000	432.84	< 2e-16	***
s(log(Klicks))	8.802	8.984	279.77	< 2e-16	***
s(Advertiser)	61.900	62.000	544.95	< 2e-16	***

Tabelle A.2: Parameterschätzer GAMM Conversion Rate

A.3 Schätzer GAMM CR ausgewählte Partnerschaften

Kategoriale Variablen (Log-Link)					
Kovariable/Kategorie	$\hat{\beta}_i$	Standardabw.	t-Value	p-Value	
(Intercept)	-2.004934	0.593095	-3.380	0.000725	***
Wochentag (Referenz: Montag)					
Dienstag	-0.041752	0.015490	-2.695	0.007036	**
Mittwoch	-0.026934	0.015628	-1.723	0.084816	.
Donnerstag	0.021581	0.015378	1.403	0.160517	
Freitag	-0.073221	0.016234	-4.510	6.50e-06	***
Samstag	-0.069100	0.016315	-4.235	2.29e-05	***
Sonntag	0.088546	0.014795	5.985	2.19e-09	***
Branche (Referenz: Elektro)					
Fashion	0.635747	0.234556	2.710	0.006724	**
Home & Acc.	0.073948	0.255510	0.289	0.772268	
Kinder	1.006317	0.261631	3.846	0.000120	***
Sonstige	1.252437	0.256109	4.890	1.01e-06	***
Vollversender	0.181275	0.241169	0.752	0.452266	
Advertiser-Größe (Referenz: 500-1000 EUR)					
1000-2000 EUR	-0.021983	0.573407	-0.038	0.969420	
2000-5000 EUR	-0.003598	0.531280	-0.007	0.994597	
5000-10000 EUR	-0.523980	0.545042	-0.961	0.336381	
>10000 EUR	-0.203289	0.584012	-0.348	0.727775	
Business Model (Referenz: Cash Back)					
Coupon	-0.809114	0.215129	-3.761	0.000170	***
Media	-0.779000	0.269813	-2.887	0.003890	**
Price Comparison	-1.074640	0.369062	-2.912	0.003596	**
Topic Website	-0.751359	0.370488	-2.028	0.042567	*
Glatte Effekte					
Funktion	edf	Ref.df	F	p-value	
s(Stunde):Elektro	7.273	8.000	25.394	< 2e-16	***
s(Stunde):Fashion	7.657	8.000	27.323	< 2e-16	***
s(Stunde):Home & Acc.	5.976	8.000	6.956	2.82e-08	***
s(Stunde):Kinder	7.705	8.000	100.624	< 2e-16	***
s(Stunde):Sonstige	7.184	8.000	68.298	< 2e-16	***
s(Stunde):Vollversender	6.991	8.000	20.208	< 2e-16	***
s(Tag):Elektro	2.409	3.000	18.888	7.35e-11	***
s(Tag):Fashion	2.987	3.000	449.503	< 2e-16	***
s(Tag):Home & Acc.	1.536	3.000	2.129	0.017247	*
s(Tag):Kinder	2.516	3.000	22.567	2.41e-12	***
s(Tag):Sonstige	2.566	3.000	10.961	0.000700	***
s(Tag):Vollversender	2.788	3.000	7.351	0.000669	***
s(log(Klicks))	8.812	8.983	206.315	< 2e-16	***
s(Partnerschaft,re)	45.453	46.000	156.928	< 2e-16	***

Tabelle A.3: Parameterschätzer GAMM CR ausgewählte Partnerschaften

A.4 Parameterschätzer GAMM Jahresverlauf

Kategoriale Variablen (Log-Link)					
Kovariable/Kategorie	$\hat{\beta}_i$	Standardabw.	t-Value	p-Value	
(Intercept)	-1.253e+01	7.599e-01	-16.490	< 2e-16	***
Wochentag (Referenz: Montag)					
Dienstag	-1.966e-02	1.325e-02	-1.484	0.137803	
Mittwoch	-3.483e-02	1.329e-02	-2.620	0.008805	**
Donnerstag	-8.988e-02	1.354e-02	-6.637	3.33e-11	***
Freitag	-1.673e-01	1.394e-02	-12.002	< 2e-16	***
Samstag	-2.467e-01	1.453e-02	-16.976	< 2e-16	***
Sonntag	-7.614e-04	1.312e-02	-0.058	0.953722	
Branche (Referenz: Elektro)					
Fashion	-2.394e+00	1.079e+00	-2.220	0.026467	*
Home & Acc.	-1.178e+00	8.284e-01	-1.422	0.155072	
Kinder	-4.314e-01	9.421e-01	-0.458	0.647025	
Sonstige	1.303e+00	9.136e-01	1.427	0.153714	
Vollversender	-3.499e+00	1.047e+00	-3.342	0.000833	***
Advertiser-Größe (Referenz: 100-500 EUR)					
500-1000 EUR	4.332e+00	1.306e+00	3.317	0.000913	***
1000-2000 EUR	7.032e+00	1.715e+00	4.099	4.17e-05	***
2000-5000 EUR	5.018e+00	1.027e+00	4.889	1.03e-06	***
5000-10000 EUR	2.132e+00	1.261e+00	1.691	0.090834	.
Feiertag (Referenz: Nein)					
Ja	-7.888e-02	2.562e-02	-3.079	0.002083	**
Glatte Effekte					
Funktion	edf	Ref.df	F	p-value	
s(TagNr):Elektro	7.824	8.00	238.807	< 2e-16	***
s(TagNr):Fashion	7.631	8.00	55.566	< 2e-16	***
s(TagNr):Home & Acc.	6.173	8.00	180.167	3.08e-13	***
s(TagNr):Kinder	7.795	8.00	224.452	< 2e-16	***
s(TagNr):Sonstige	7.852	8.00	224.763	< 2e-16	***
s(TagNr):Vollversender	5.504	8.00	8.513	3.14e-05	***
s(log(Impressions))	8.811	8.97	439.340	< 2e-16	***
s(Partnerschaft,re)	20.739	23.00	512.169	< 2e-16	***

Tabelle A.4: Parameterschätzer GAMM Jahresverlauf

A.5 Parameterschätzer GAMM Warenkorbwert

Kategoriale Variablen (Log-Link)					
Kovariable/Kategorie	$\hat{\beta}_i$	Standardabw.	t-Value	p-Value	
(Intercept)	4.698993	0.623568	7.536	4.87e-14	***
Wochentag (Referenz: Montag)					
Tuesday	0.005952	0.006940	0.858	0.39111	
Wednesday	0.009984	0.007210	1.385	0.16613	
Thursday	0.007642	0.007485	1.021	0.30721	
Friday	0.020153	0.007729	2.607	0.00912	**
Saturday	0.023838	0.007501	3.178	0.00148	**
Sunday	0.020831	0.006806	3.061	0.00221	**
Branche (Referenz: Elektro)					
Fashion	-0.763389	0.248295	-3.075	0.00211	**
Home & Acc.	-0.380968	0.327767	-1.162	0.24511	
Kinder	-1.016535	0.298471	-3.406	0.00066	***
Sonstige	-1.264174	0.297786	-4.245	2.18e-05	***
Vollversender	-0.820265	0.288299	-2.845	0.00444	**
Advertiser-Größe (Referenz: 0 EUR)					
0-100 EUR	0.552368	0.715566	0.772	0.44016	
100-500 EUR	0.621813	0.608552	1.022	0.30688	
500-1000 EUR	0.170980	0.615161	0.278	0.78106	
1000-2000 EUR	0.460755	0.579429	0.795	0.42650	
2000-5000 EUR	0.254802	0.571424	0.446	0.65566	
5000-10000 EUR	0.491555	0.602506	0.816	0.41459	
>10000 EUR	0.260519	0.693861	0.375	0.70732	
Business Model (Referenz: Cash Back)					
Coupon	-0.050168	0.005718	-8.773	< 2e-16	***
Media	-0.065398	0.009692	-6.748	1.50e-11	***
Portal & Communities	-0.017759	0.016000	-1.110	0.26701	
Price Comparison	0.015243	0.010589	1.440	0.15000	
Topic Website	0.204614	0.010211	20.039	< 2e-16	***
Advertiser Account Manager (Referenz: Nein)					
Ja	0.074486	0.152505	0.488	0.62525	
PublisherAccountManager (Referenz: Nein)					
Ja	0.086435	0.007526	11.485	< 2e-16	***
Voucher (Referenz: Nein)					
Ja	0.108364	0.012218	8.869	< 2e-16	***

Tabelle A.5: Parameterschätzer GAMM Warenkorbwert

Glatte Effekte					
Funktion	edf	Ref.df	F	p-value	
s(Tagesminuten):Elektro	4.437e+00	8	3.104	3.76e-05	***
s(Tagesminuten):Fashion	3.001e+00	8	9.509	0.00487	**
s(Tagesminuten):Home & Acc.	9.396e-04	8	0.000	0.75817	
s(Tagesminuten):Kinder	6.364e+00	8	19.315	< 2e-16	***
s(Tagesminuten):Sonstige	5.868e+00	8	8.179	2.18e-07	***
s(Tagesminuten):Vollversender	1.948e+00	8	1.368	0.00667	**
s(Tag):Elektro	7.413e+00	8	57.813	< 2e-16	***
s(Tag):Fashion	6.959e+00	8	78.264	1.55e-10	***
s(Tag):Home & Acc.	1.483e+00	8	0.491	0.06889	.
s(Tag):Kinder	7.135e+00	8	18.961	< 2e-16	***
s(Tag):Sonstige	1.708e+00	8	0.763	0.03122	*
s(Tag):Vollversender	6.187e+00	8	23.898	8.84e-10	***
s(Advertiser,re)	5.179e+01	53	468.520	< 2e-16	***

Tabelle A.6: Glatte Effekte GAMM Warenkorbwert

Anhang B

Residualplots

B.1 GAMM Conversion Rate

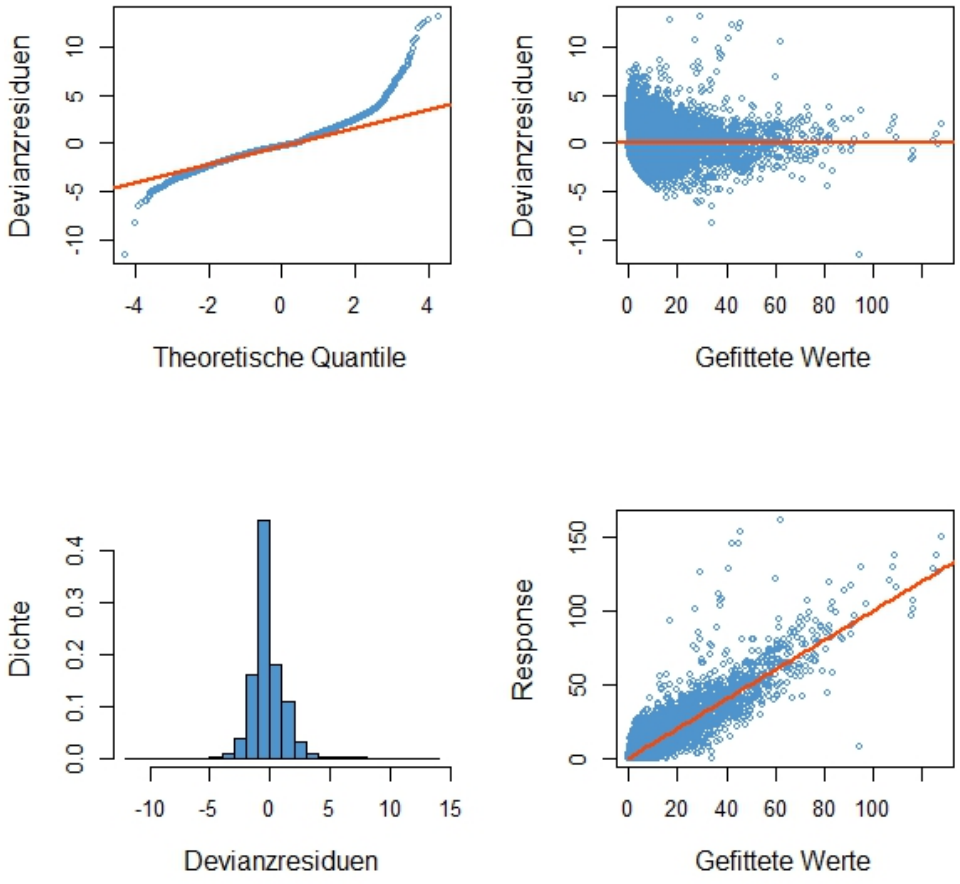

Abbildung B.1: GAMM Conversion Rate: Residualplots

B.2 GAMM CR ausgewählte Partnerschaften

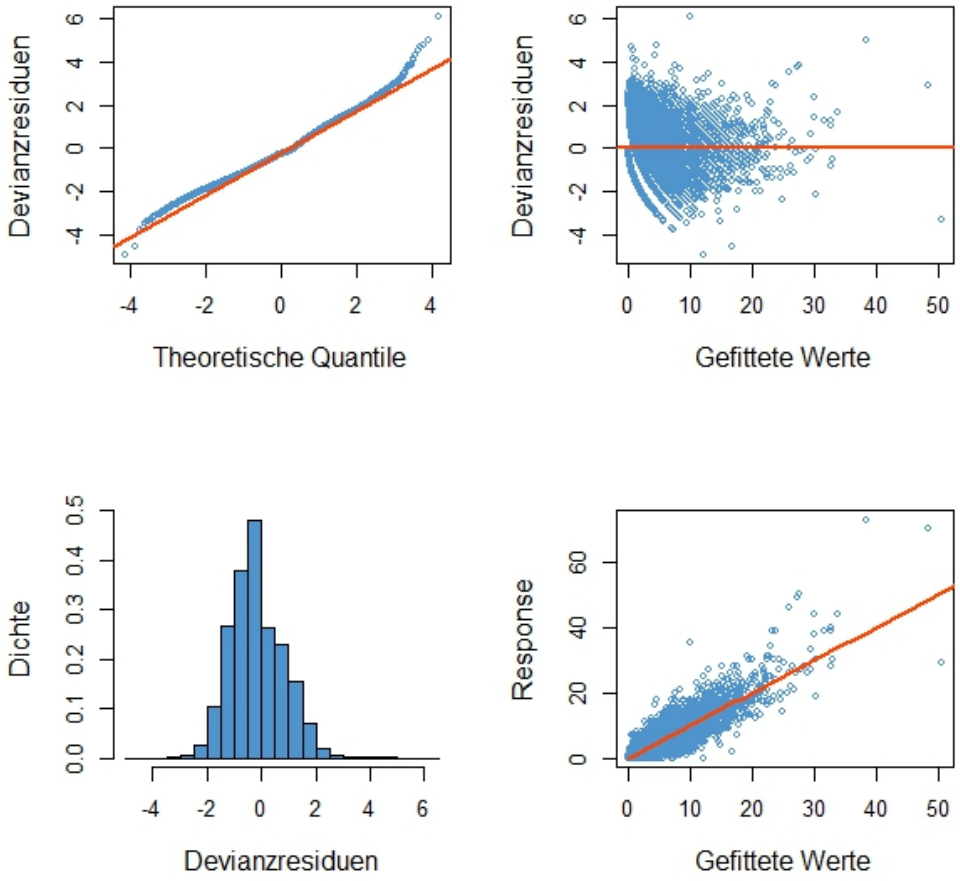

Abbildung B.2: GAMM CR Partnerschaften: Residualplots

B.3 GAMM Jahresverlauf

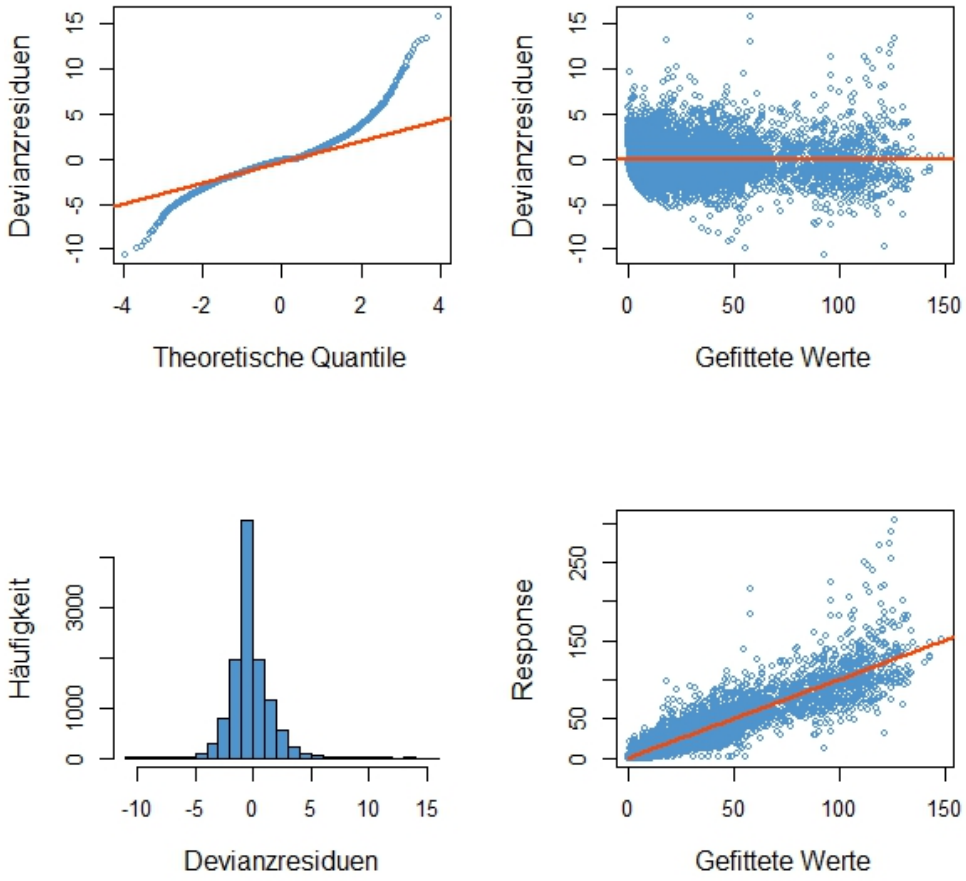

Abbildung B.3: GAMM Jahresdaten: Residualanalyse

B.4 GAMM Warenkorbwert

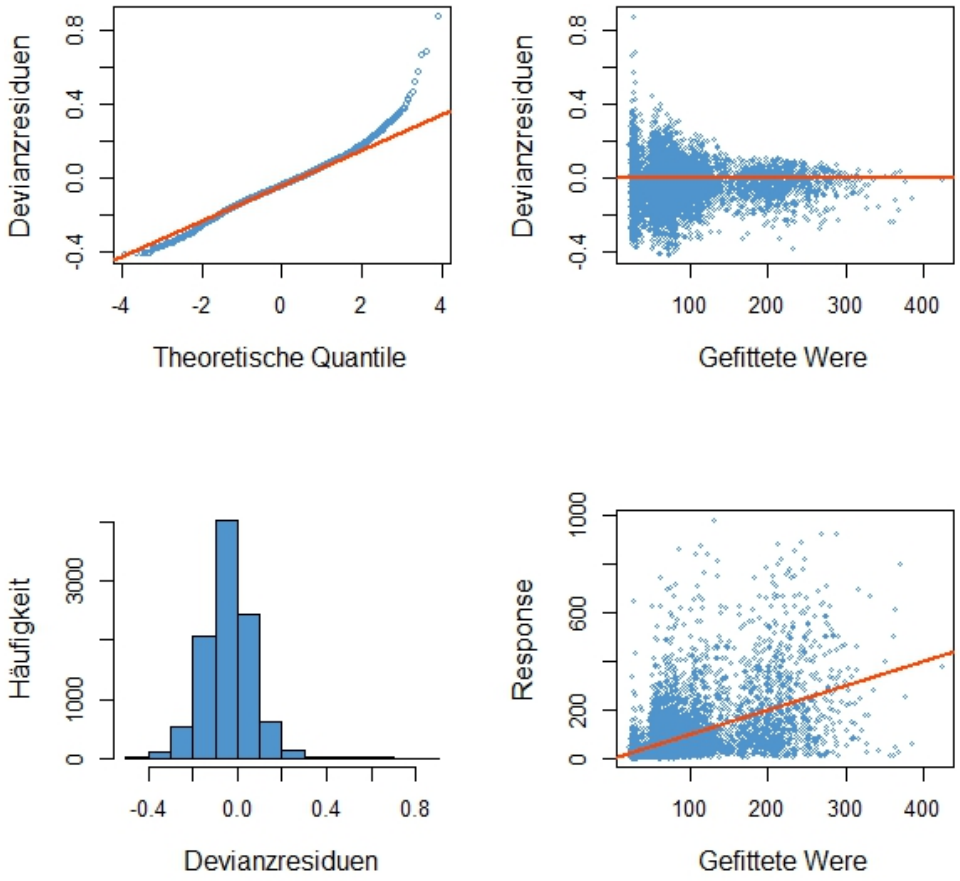

Abbildung B.4: GAMM Warenkorbwert: Residualplots