Ludwig Maximilian University

Department of Statistics

Master Thesis

Development analysis of publishers in online affiliate marketing

Isabel Matheja
August 26th, 2014

Supervised by:
Prof. Dr. Göran Kauermann

Abstract

Affiliate marketing is a marketing-oriented online marketing channel. Via partner programs between advertisers and publishers advertising contacts are established. Resulting transactions are measured and based on performance. The present work uses stored data to gain insights into the underlying structure of customers registered with an affiliate network. A descriptive analysis captures the data structure with publishers as underlying research object. Then the statistical analysis is separated into two parts. First a Generalized Additive Model (GAM) accomplished with a bootstrap analysis is conducted to examine influencing variables for payment. In a second step the height of payment is analysed for those publishers with positive payment in 2013. This is evaluated with a Linear Mixed Model (LMM).

Contents

List of Figures

List of Tables

Preface

This report is written in cooperation with the Department of Statistics of the Ludwig Maximilian University Munich and a leading provider of performance marketing in Europe. The focus of this work is on the description and application of statistical methodology for a data set. In order to understand the content of this work basic statistical knowledge is needed.

I want to thank Prof. Dr. Göran Kauermann from the Department of Statistics for taking over the supervision of the thesis and for supporting me in the planning of the proceedings and in the statistical implementation.

Further, my gratitude goes to the cooperating company for the excellent support during the entire project and for their introduction into affiliate marketing.

In Appendix 6 the abbreviations used in this work are listed. This work was created with LaTeX. All graphics and statistical analysis were generated with the statistical programming language R (R Core Team, 2014) and involved in the LaTeX code with knitr (Xie, 2014). For a detailed description see Xie (2013).

Chapter 1

Introduction

1.1 Introduction into Affiliate Marketing

With the rising use of the Internet and the relevance of social media applications, online marketing gained increasing importance over the last years. Nowadays it is an essential part of marketing and is recording high growth rates. According to a survey of the Federal Statistical Office among German private households, approximately 42.3 million people have bought or ordered goods or services for private use over the Internet in 2012 (Destatis, 2014). This corresponds to a share of 74% of Internet users aged ten years and older. With this large target group for online advertising and campaigns the indispensable part of online marketing becomes apparent. The term online marketing refers to forms of advertisement, which are distributed via the Internet, such as search engine marketing, email marketing, affiliate marketing and social media marketing. According to Statista (2014) the share of online marketing of total German marketing accounted for 11% in the first half of 2014. A study of the Interactive Advertising Bureau Europe (Europe, 2014) stated that in Europe "[...] online advertising grew 11.9% to a market value of €27.3bn in 2013". In Germany online advertising grew 10.7% at the same time. A driving force of this development is the increasing e-commerce business as more and more people are shopping online.

This is a fast moving, innovative and competitive industry, where small improvements in the conversion rate of an advertisement can lead to large improvements in the effectiveness of campaigns. As a media agency, a substantial part of the business activity is to place client advertisements on websites. Given the large amounts spend in online advertising and the competitive nature of the industry, it is increasingly important to make sure each advertising dollar is invested in the right way.

Affiliate marketing is one form of online marketing, where the idea is to have sales and networking partners. The fundamental principle in affiliate marketing is to establish a connection between potential customers and a company, that would like to sell something to these potential customers. The participants of this form of marketing are the advertisers and publishers the affiliate network unites and the customer, who is the object of desire. Advertisers are companies that provide web-based services or products and promote those. So advertisers must find ways to bring potential customers to their website and motivate them to purchase. Publishers are operators of websites and related online services. They can complement their content through

relevant ads for products or services.
The publisher, also known as the 'the affiliate', advertises products or services of the advertiser, also known as the 'merchant', on his (or at a special) homepage or via other distribution channels, as for example via Email. For every transaction or sale made on his page, he receives a provision of the advertiser. The advertisers provide the publishers with promotional material (banners, text links, HTML tags, product data, etc.) that the publishers can integrate on their websites. Affiliate marketing can be conducted independently or via a partner network. Affiliate networks are independent platforms on the Internet. They mediate between advertisers and publishers, bring the right partners together and optimize their business. The mediation through affiliate networks has the advantage, that they include a pool of potential partners and provide technology for performance measurement (tracking) and accounting. So they facilitate cooperation significantly. For this service they receive remuneration. This is based, for example, on a percentage of the transaction value of the order. The affiliate network obtains a fee from the advertisers for their services, based on the advertisers payment to the publishers. For the publishers the services of the network is free of charge.
The underlying affiliate network uses performance marketing, i.e. the advertiser only pays a fee to the publisher if a predetermined action was completed by a visitor. Hence, the advertiser pays only for measurable advertising success. To ensure the measurability of success and accurate accounting between the partners, orders, clicks and impressions are documented. In theory an impression is counted each time an advertisement is shown on a publishers website. However, this is technically not always realizable. It is only to be regarded as a measurable success, if it leads to a click on the ad. Only very few clicks arise for a vast number of impressions. And merely a fraction of these clicks then leads to orders. The conversion rates certainly depend on various factors, e.g. businessmodel or type of banner. Generally it can be assumed, that for every 100,000 impressions, 10 clicks and 1 order result. Technically, this is implemented mostly by means of tracking pixels or use of cookies. Due to the immense amounts of data, most documentation processes run automatically. The commission from the advertiser can be separated into three categories: pay-per-click, pay-per-lead and pay-per-sale remuneration.

PPC (pay-per-click)	fixed commission for each click of the user on the advertising medium (e.g. €0.05). As it is still a challenge to attribute value, it is not uncommon to use the easiest of all models, where all value is attributed to the last click.
PPS (pay-per-sale)	the advertiser pays an affiliate, when the affiliate sends them a customer, who purchases something. The affiliate then receives a percentage of the sale.
PPL (pay-per-lead)	the publisher receives a fixed fee for a particular action of the user, that is, if the customer submits contact data in a lead generation campaign e.g. creates an account or completes a questionnaire.

For combination of the above types, combined programs can be selected. Additionally they offer CPO (Cost per Order) programs, so called postview programs, based on banner impressions. The standard procedure is the following: The customer views an advertisement, which causes an impression, then the customer might click on the ad, which leads to a click. If the customer completes a purchase or registers with the homepage, an order is created. The publishers,

who drive traffic to the advertisers websites are paid for completed transactions. This may, depending on the agreed payment model, include leads, referrals or sales. The action has to get confirmed by the advertiser, only after that the publisher receives his payment. The network is participating with a share of about thirty percent of the publishers revenue, which depends on the advertiser and the network. For more information on online and affiliate marketing see Lammenett (2013).

Tracking Technology

For effective affiliate marketing an efficient tracking mechanism is needed. Through tracking the performance can be enhanced, as publishers can track transactions and identify the best performing products. Cookie tracking is still the foremost and most reliable form of tracking in affiliate marketing. A cookie is a small text file stored in an user's web browser, while browsing a website. When the user revisits the website in the future, the data stored in the cookie is sent back to the website by the browser. Even after a prolonged period, the cookie is then able to attribute the completed sale to the appropriate publisher. Cookies allow smoother interaction on frequently visited sites and permit publishers to enhance and personalise their site experience. Cookies can be deleted by the user, however they are the best and easily-accessible method of storing settings or informations of the user.
A marketing practice based on cookies is retargeting, which can be used to deliver targeted ads and offers to users, based on their previous actions and behaviour. When a potential customer visits a website, retargeting can be used to place display ads of online shops the user previously visited. This makes the shopping experience more personal and relevant, thus significantly increasing the efficiency of campaigns and boosting conversion rates. This form of tracking requires extended technical conditions and is therefore only used by a selected number of customers.
A new form of tracking without cookies is the fingerprint tracking. Based on numerous parameters a digital signature of the user's computer is created, a kind of fingerprint. So computers can be reliably identified, even if cookies are deleted, deactivated or blocked. Large social sites and companies have enormous amounts of data about their users, independent of cookies, that they are eager to further monetize. Usage might therefore increase in upcoming years with technical improvements.

1.2 Subjects and Aims of the Project and the Thesis

This work is written in cooperation with a leading european affiliation network. Within this cooperation this is the second master thesis. The previous thesis dealt with the temporal influences on the success rates in affiliate marketing based on advertisers within online retail. Additionally a survival analysis for the time between clicks and orders was conducted. Then the influence of temporal components on the amount of the mean shopping basket for online orders was analysed (see therefore Meingast (2013)). Other than focusing on the advertisers, this work aims to shed light on the publishers side.
Through the companies position as central aggregator of advertisers and publishers in the affiliate

marketing several terabytes of data from all areas of the online market come together. The data is then checked, corrected and stored in a structured manner in SQL databases. Within the data warehouse the firm uses for data storage and analysis, a vast number of standardized and individual analysis can be created. This can be used both by the customers for their personal activities and internally for controlling and creating reports. While those reports are based on descriptive analysis, this work aims to detect characteristics of the publishers based on statistical model analysis.

In times of big data companies extend their interest in using their stored data to gain information on customers. Especially those companies which are operating in the online business market, with huge data warehouses want to get access to this data and use it for advanced data analysis. Still limited computer power, knowledge and time is a threshold for extensive analysis. Data analysis technologies for marketing research are widely used in the areas of consumer preference analysis, market segmentation, product pricing, sales driver analysis, and sales forecast. Mainly the analysis is conducted with methods such as t-test, ANOVA, regression, conjoint analysis, and factor analysis. While most statistical analysis in marketing adresses the question, which advertisement creates the most value, this work has a diverse approach. It focuses on the publishers of an affiliate network. As counter-party to advertisers, publishers create websites. These websites attract people and publishers charge advertisers for the possibility to show ads to the visitors on the publisher's site.

With growing competition and cost pressure among affiliate networks, the analysis of customer characteristics and development becomes more and more important. High performing affiliates, the so called short term, need to be individually supervised to secure strong revenue and profit growth. However, focusing only on the short term leads to high dependency on those customers and potential vulnerability. Therefore the long term should not be neglected. This work aims to provide insights into the publishers line-up, recent developments and influences on the payment. Using these findings, support measures for (Longtail-, Midtail- and Shorttail-) publishers can be developed. The results might then serve as a benchmark for the analysis of new policies and products.

The Publishers Journey

From registration to the successful long-term work with advertisers within the affiliate network a publisher goes through several stages. The main steps for that development are described in the following.

After a publisher has registered with the network, he will receive a confirmation code via email which has to be activated by clicking on it. After completing the registration form with the required information on the publisher (e.g. website, businessmodel, personal data) a verification check is conducted. After that check the publisher can start and has the internal status *prechecked*. The next step is to select suitable partner programs and apply at the appropriate advertisers. Before the publisher can participate in the program, a release by the advertiser is required. The advertiser can accept but also block the partnership. After the release by the advertiser, the publisher has access to a wide variety of promotional material, that he can easily integrate into the source code of his website. If the correct choice for both the advertiser and

the best converting advertising is made, this leads to high success on both sides.
Ensuring that publishers earn money with their advertisement comes down to understanding the customer journey - where customers are going to research and make decisions, via what devices, time of day etc. This information is invaluable in assuring where affiliates impact consumers and that publishers have the right content and offers. The network actively seeks to give advertisers and publishers information on how they can improve their conversions.

1.3 Structure and Approach

After an introduction into the aim and the topic of the work, chapter 2 explains the dataset in detail. It includes a formal description of the data separated by master, traffic and partnership data. Moreover it clarifies, how specific variables are obtained. Then some additional background information about affiliate marketing is presented to get an understanding of this field of marketing. Chapter 3 presents the descriptive analysis, to grasp the structure of the data and the variables. It starts with the complete data set, then focuses on the payment and traffic of the publishers. In a second step the development throughout the year 2013 and the partnerships are evaluated. Chapter 4 gives the statistical theory for the used analysis and models. It focuses on the Logit Model within the Generalized Linear Model (GLM) framework, bootstrap techniques and the Linear Mixed Model (LMM). Furthermore some additional statistical measurements are explained. The analysis of the data with the main results being summarized and evaluated are presented in chapter 5. The analysis part is structured in the Logit part, analysing how explanatory variables influence payment at all and the LMM part for the height of the payment. It is evaluated for all publishers and in the LMM for each businessmodel category. Finally in chapter 6 the key findings of the work are outlined and an outlook is given.

Chapter 2

Data

The data used for this report was provided by the affiliate network. To integrate data from several sources, store the current and historical data and to produce reports and data analysis the company uses a data warehouse. The data for this thesis is an extract of current and historical data of the data warehouse. In agreement with both the university and the company, the structure and kind of data was determined. Getting to the final data basis has been a long process of adjustments and changes in the data. While the data was recorded from 2001 on, the data warehouse was established in 2012. Thus changes in the publishers attributes are only recorded from that point on. The data consists of several different data extracts. Due to the large size of each extract, they have been merged in `R` to several large data sets. The work focuses on publishers, which have been registered with the German network and have been inserted before the 14th of March 2014. Most of the analysis will focus on the year 2013, as payment and traffic data are only available as of 2012 and for computational reasons.
For the purpose of the work, the aim was to include the majority of publishers. However, some publishers, who joined the network before 2003 had to be excluded, as they showed partially invalid attributes. Some showed no *status*, others no *insertion date* and additionally no values for *age in days*. Those have been excluded from the analysis. Moreover as the data was included subsequently to the data warehouse, publishers, who have registered before 2001, have been assigned to the fixed insertion date ('2001-08-03 12:04:00'). This date was replaced with their insertion day to ensure the right calculation of age. With this procedure the majority of publishers before 2003 could be kept in the analysis.
Steps for achieving the data structure are (in a nutshell) the following:

1. Keep only those publishers before 2001, which had a status and take for those the insertion *day* as insertion *date*
2. Delete publishers without insertion date
    (a) For descriptive analysis in chapter 3: Take only publishers with end date '2999-12-31' - this is the date of the last entry in the data warehouse for each publisher (to have each publisher just once). Exclude publishers, who have been deleted before 2013. Combine them with the traffic data and later the partnership data.
    (b) For models in chapter 5: Keep all entries of a publisher in the data warehouse, except if a publisher has several entries per month. Then keep only the last entry of the

month. Combine them according to Publisher ID and month in 2013 with traffic and partnership data. All characteristics of a publisher are carried forward, such as in every month in 2013 the current characteristics of the publisher are captured.

This scheme also shows the underlying approach of the descriptive and model part of the analysis. While the latter captures the development over time, the descriptive analysis is a current cross sectional analysis of publishers characteristics.

2.1 Data Sets

The data is a combination of several data excerpts and changes during the analysis as stated above. The main variable of interest is the total payment in 2013, either as sum per year as used in the descriptive part or on a monthly basis. More information on the total payment is given in the section for the traffic data in 2.1.2. Several explanatory variables are selected to examine their influence on the total payment per year or month. Those are captured in the master data, traffic data and partnership data. The different data sets and explanatory variables are being described in the following.

2.1.1 Master Data

The master data describes the basic information of each publisher in the network. Publisher information such as personal and business details are obtained through the registration process, every publisher passes. The publisher can be identified through his identification number, the *Publisher ID*. For the analysis the ID is pseudonymized, so no conclusions on the true identity of the publishers are possible. For each change in a publishers master data a new update is made in the data warehouse. While the master data does not change that often, traffic data (mostly) varies each month.
For every publisher the *insertion date* is given, that is the time the publisher has first registered with the network. Then for every change in the publishers master data, for example if the status or the businessmodel changes, a new entry is written. This new entry has the *start date* the current change was undertaken and the *end date* when a new change was introduced. If it is the last entry of a publisher the end date has the date format '2999-12-31', while the first entry of a publisher is marked with the start date '1900-01-01'.
For the master data one of the most important variables presents the so called *status*, which is given internally for the publisher and reflects the current status of each publisher. According to a status a publisher can build partnerships, is actively supervised by a key account manager (KAM) or needs to be checked. After a publishers registration he is automatically checked and then set to 'precheck'. Only after the publisher reached a specified amount (€25) for consecutive months, he will be checked again manually and if all is correct the status is changed to *ok*. As this manually check is time-consuming, many publishers operate with status prechecked, especially if they earn little money. This categorical variable has different levels, which are listed along with their shortcuts in table 2.1.
Another categorical variable is the *businessmodel* of the publisher. Each publisher is allocated into a businessmodel, according to his type of homepage or business. In the analysis for the

Shortcut	Explanation
blocked	blocked
blbypre	blocked by precheck
bl.ref.	blocked and advertisers refunded
ok	ok, checked
oktop	ok, top publisher
notch.	not checked
Pinf.susp	publisher informed of suspicion
prech.	prechecked
susp.	suspicious

Table 2.1: Explanation of status shortcuts

businessmodels short names are used, the original names and their shortcuts can be seen in Table 2.2. Even though Email and Search Engine/Search Engine Marketing differ in their businessmodel, they are combined into one, as they include a limited number of publishers. Before 2012 it was not mandatory to specify ones businessmodel in the registration process, therefore most publishers are assigned to the businessmodel Unknown. As the publishers mostly assign their businessmodel to themselves, it is not guaranteed, that each publisher is allocated in the right businessmodel. For businessmodels Topic this may be more applying than for more specialized models as Coupon or Cashback.

Shortcut	Explanation
CB	Cashback - Online services for customer loyalty and bonus programs
C	Coupon - Websites which offer discount coupons, local deals, live shopping
E	Email Distributor, Search Engine, Search Engine Marketing, PPC
M	Media - Book advertising spaces at high reach quality pages to promote affiliate programs per postview or retargeting
Portal	Portals and Communities - Social networks, forums and blogs
PC	Price Comparison - Price Comparison Portals
T	Topic Website - Websited with thematic focus and specialised services
Unknown	Unknown - The businessmodel is not specified

Table 2.2: Overview of businessmodel shortcuts and their explanation.

The publishers age is given in days in the data warehouse. As most information is evaluated on a monthly basis due to computational reasons this was changed into *age in months*, with reference date 2014-01-01.
Another important variable is the variable *KAM*, which indicates whether a publisher is supervised by a personal key account manager or not. Typically the key account managed publishers are bigger websites, that earn higher revenues. Publishers then get access to individual support, aimed at improving the publishers reach and turnover. Each key account manager is specialized on one or several businessmodels. Sometimes KAM publishers have smaller subpages, which are then marked KAM even though they would not be considered KAM themselves.

2.1.2 Traffic Data

Traffic data is given on an aggregated monthly level. Variables here include total payment, payment generating orders, as well as clicks and impressions. **Impressions** count the number an advertisement is displayed. As this variable is not relevant for billing, it is not validated and adjusted. Moreover impressions are not always available. Traffic corresponding to one special advertiser are excluded, as for this advertiser orders are calculated differently.
Clicks marks the number of clicks that have been registrated on an advertisement. The ratio between clicks and impressions is the clickthrough rate (CTR). The CTR is generally very small. **Orders** count the resulting orders a publisher received over all his advertisers. Depending on the businessmodel the number and height of the orders can vary. While for example travel portals receive fewer orders, those amount to a higher overall value than for advertisers with pet food. **Total payment** is the aggregated payment per given time period for a publisher. It is measured at registration time, i.e. the time point the order and therefore the resulting payment was recorded within the network. This is summarized per month or year for this work. It is calculated by

$$\text{total payment} = \text{total order payment} + \text{click payment} + \text{bonus, where}$$
$$\text{total order payment} = \text{total orders} \cdot \text{avg. total payment per order} - \text{partial cancelation}$$

Bonus are special payments an advertiser pays to a publisher, for example, because they have special agreements over targets to be reached within a month. In general, the bonus offers advertisers the opportunity to pay special payments of any kind to the publisher. Corrections in terms of adjustment for wrong orders or restored orders may reduce the total payment. The total payment is not equal to the payout amount - the confirmed payment. Only after the payment is validated it is paid out to the publisher. The confirmed payment is measured at validation time. In general it takes about three to six month before all payment is validated (as payment can also be canceled again). The confirmed payment is the actual amount the publisher receives. The amount of total and confirmed payment may change but must not. As the validation requires time, the confirmed payment lags the total payment. For this reason the total payment was taken into consideration as dependent variable for this work. It can be possible and is quite common that publishers do not receive a payment in every month. This is especially valid for publishers with less visitors on their homepage und thus less traffic, as the threshold for being paid out is €25 net. If this amount is not reached, the credits are disbursed and paid out as soon as the accumulated credits exhibit this value.

2.1.3 Partnership Data

As the name implies, partnership data provides information about the number of existing, but also deleted and accepted partnerships per publisher. Publishers can enter into a partnership with advertisers. After choosing an advertiser from the advertisers space, the publisher has to apply for the partnership. Only after the advertiser accepts the publisher, the partnership is valid and active and the publisher can include the advertisers promotion into his website. The partnership can be set on hold or be deleted by both sides at any time. Affiliate models and

campaigns will work differently, depending on the brand and market. Moreover the advertising should be used to match the website and its target audience. Therefore it is necessary for publishers to assess which merchants work best with their market and potential customers, before applying for a partnership. Depending on factors like businessmodel, publishers may have different strategies for entering into an affiliate with advertisers. Topic models usually seek advertisers, who fit with their homepage topic, while Coupon models aim on a broad target group with different advertiser types.
The number of *existing Partnerships (PS)* in 2013 is also considered as influential variable on the payment. It has been calculated as

$$\text{existing PS}_t = \text{accepted PS}_{t-1} - \text{deleted PS}_{t-1} + \text{accepted PS}_t - \text{deleted PS}_t$$

where t is the month. Thus, the equation takes the already existing partnerships from the previous month into account. For illustration the calculation for January 2013 is: Existing PS 01.2013 = Accepted before 2013 − deleted before 2013 + accepted 01.2013− deleted 01.2013). Partnerships can additionally be accepted by a key account manager, as this happens without date assignment in the data warehouse, those are not counted as partnership. Therefore the number of existing partnerships as calculated can slightly differ from the actual number.

2.2 Example of Data Structure

Table 2.3 provides a brief insight into the dataset. For application of the models to be desribed in chapter 4, the data set was transformed into the so called long-data format of repeated measures. For computing summary statistics and plotting the wide format was used, where every publisher has one column.

PublisherID	Month	Businessmodel	Age	Status	KAM	Exist.PS	Orders	Payment	...
4050	Jan. 13	Portal	137	ok	0	22	0	0.05	...
4050	Feb. 13	Portal	138	ok	0	22	0	0.06	...
...	...	...	...	...	...	...	...	...	...
4050	Dec. 13	Portal	148	ok	0	21	0	0.11	...
...	...	...	...	...	...	...	...	...	...
2911162	Jan. 13	Topic	98	ok	0	750	41	138	...
2911162	Feb. 13	Topic	99	ok	0	750	33	144.5	...
...	...	...	...	...	...	...	...	...	...
2911162	Nov. 13	Topic	108	prech.	0	713	31	131.7	...
2911162	Dec. 13	Topic	109	susp.	1	707	25	106.6	...
...	...	...	...	...	...	...	...	...	...

Table 2.3: Example of data structure for two selected publishers for payments in 2013.

The publisher with the (pseudonymized) identification number 4050 is one of the oldest publishers, who had payment in 2013. Note that, other than in this example, some publishers have less than twelve existing observations in 2013, when they did not receive payment in every month. Moreover the values of the master data, i.e. for businessmodel, KAM and status may change but must not.

2.3 Data Merging

The relevant data for the analysis are not contained in a single data set from the beginning, but come from different data sets. In our case the above described master data, traffic data and partnership data are stored in different data sets and need to be merged. Therefore feature vectors from the various data sets have to be assigned to each other. This is done based on specific label features, which are contained in each of the data sets. The *Publisher ID* and sometimes *month* serve for this case and the data sets can be concatenated via those variables. Missing data might be generated, if a value of the variable in one data set does not match the values in all other data sets (that is the case if the value is not contained in the other data sets).

Chapter 3

Descriptive Analysis

3.1 Basic Information on Publishers

For the descriptive analysis the data is composed as described in the previous chapter. Here every publisher is counted once by taking only the last status into account (End Date = 2999-12-31), which results in 427,152 publishers. All publishers who have been deleted before 2013 are excluded. This was selected so that publishers deleted during 2013 or in 2014 are still electable for the payment data set for 2013. Then the total number of publishers amounts to 217,339, i.e. almost half of the publishers are marked as deleted in the data warehouse. Most of the plots in this work were created using Hadley Wickham's ggplot2 package for R (Wickham, 2009). Note that if the variables are plotted on a logarithmic scale, the axes mostly show the non-transformed values.

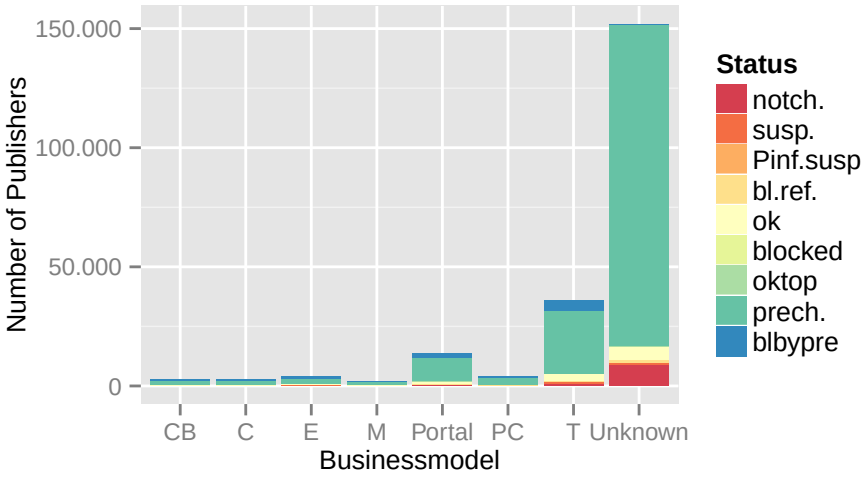


Figure 3.1: Number of existing unique publishers by businessmodel and status.

Figure 3.1 shows the allocation of unique publishers per businessmodel and status for all existing (i.e. nondeleted) publishers as of 2013. A table and a figure of the distribution between deleted and existing publishers can be found in the Appendix at figure A.0.1 and the corresponding table A.1. Most publishers are in businessmodel (BM) Unknown with a total number

of 151,914 publishers resulting in approximately 70% of total existing publishers. Most of those are prechecked, the second biggest group is notchecked, followed by ok and suspicious. 35,803 publishers are assigned to businessmodel Topic, representing 16.5% of nondeleted publishers. Then 6.3% of the total existing publishers have the businessmodel Portal. The other businessmodels do not account for more than 2% each. From the deleted publishers, about 96% came from businessmodel Unkown with a large part of notchecked publishers.

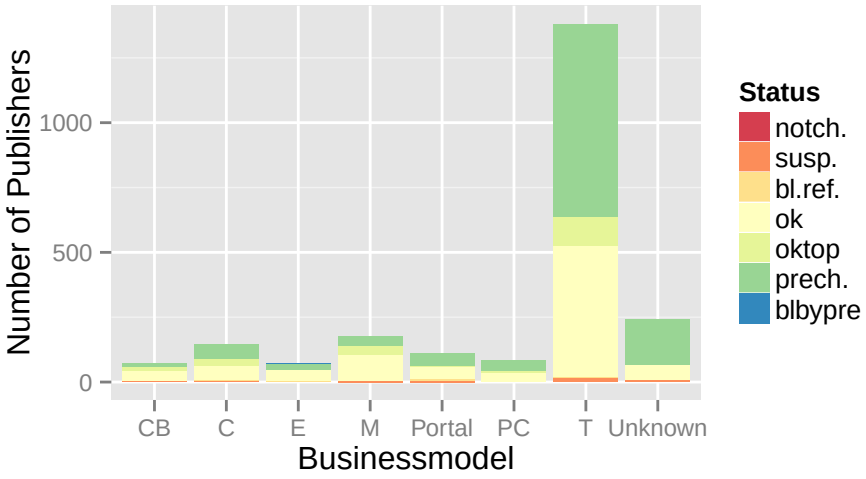


Figure 3.2: Overview of publishers by businessmodel and status for key account managed publishers.

Figure 3.2 gives information about the businessmodel and status allocation of key account managed publishers. With a total number of 2285, key account managed publishers account for only 1.05 percent of all existing publishers. Most of them are assigned to businessmodel Topic with statuses prechecked, ok and oktop. Then next businessmodels with most KAM publishers are Unknown, Media and Coupon. In comparison to all publishers as shown before in 3.1, key account managed publishers do not occur with statuses Pinf.susp and blocked.

3.1.1 Publisher with Insertion Date in 2013

To discover the development of the recently inserted publishers, publishers who have registered with the affiliate network in 2013, are examined in comparison to those inserted beforehand. Those amount to 23,033 publishers. Figure 3.3 shows the classification of businessmodels between 2013 and the time before 2013 (e.g. from 1999 until 2012-12-31). It clearly shows, that most publishers, who signed up before 2013, are in businessmodel Unknown. In 2013 Topic became the most important businessmodel with almost 45 percent of new publishers starting in this category. In 2013 the company changed its registration rules, from that point on every publisher had to insert a businessmodel, this is why the number of unknown businessmodels was naturally reduced. In 2013, the share of other businessmodels than Unknown rose in comparison to previous years. Surely the allocation for all years relates more to the distribution of the publishers before 2013, than to the distribution for the year 2013. Also for the status major

differences can be observed in figure 3.3. While most publishers before 2013 are assigned to the status prechecked, publishers who registered in 2013 are distributed more broadly. Mainly because about 25% each have not been checked manually (i.e. have status notchecked) or are blocked by the precheck.

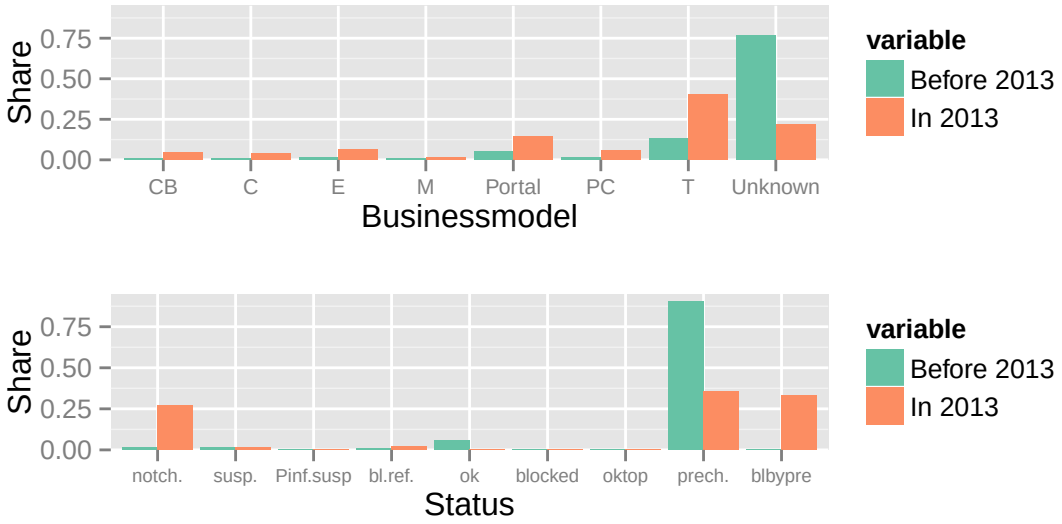


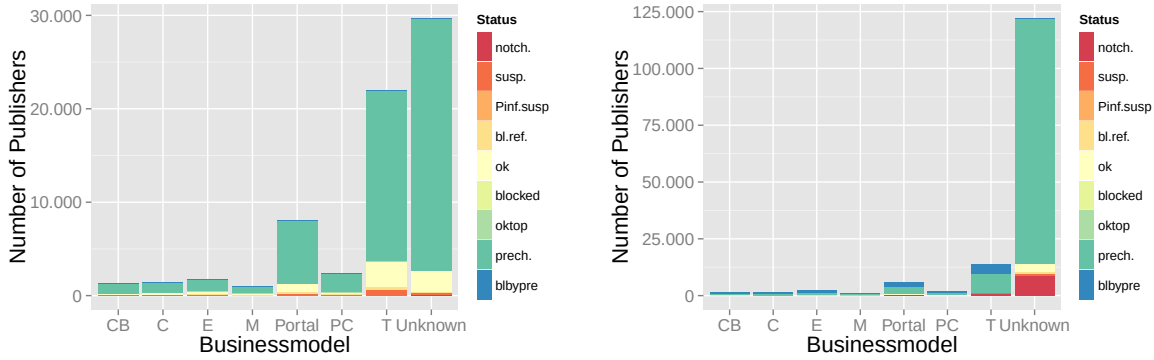
Figure 3.3: Comparison of publishers share with insertion date in 2013 and before 2013 by businessmodel (above) and status (below).

3.2 Payment and Traffic

As the main part of the analysis will focus on the payment of each publisher, the payment and traffic will be further examined. Therefore the traffic data set, which contains impressions, clicks, orders and the total payment is added to the master data set. Only those publishers, who had traffic in 2013 are kept, those not contained are combined into a new data set and can be analysed. Only 67,684 publishers had traffic in 2013 (i.e. had at least one month with either registrated impressions, clicks, orders or total payment), which equals 31.1 percent of existing publishers. Therefore the question arises, why the remaining 68.9 percent of publishers had no payment in 2013.

Evaluation of publishers not contained in the traffic dataset

It might be interesting to understand, why publishers did not earn money in 2013. The two following plots should give an impression about the distribution of both, those with and without payment and traffic in 2013. The differences in businessmodels and statuses for publishers with and without traffic in 2013 are captured in figure 3.4. By far, the largest part of publishers with traffic (3.4a) is operating in businessmodels Unknown, Topic and Portal, with status prechecked, followed by ok and suspicious. Clearly, publishers without traffic are mostly within businessmodel Unknown and status prechecked (3.4b).
The boxplots in figure 3.5 show the difference in age by businessmodel for publishers with and without traffic in 2013. While the overall mean of age in months for those publishers with traffic

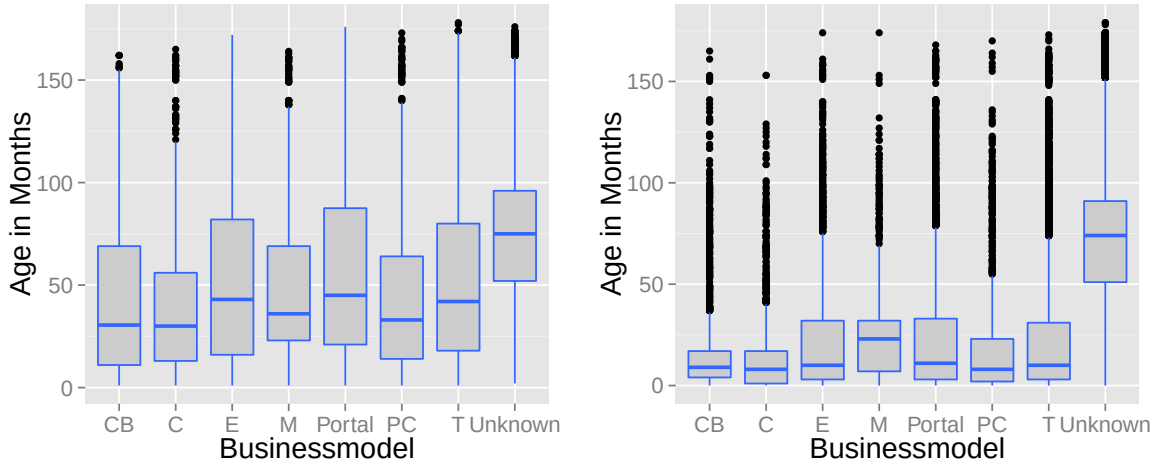


(a) Publishers with traffic in 2013

(b) Publishers without traffic in 2013

Figure 3.4: Comparison of number of publishers by businessmodel and status. The y-axis labels are varying due to the changes in group size.

is about the same than for those without, devoid of the businessmodel Unknown a different picture is given. If Unknown (which accounts for 44% of those publishers with traffic and 82% of those without) is excluded, then the mean age of publishers with traffic is 52 months and for the ones not contained in the traffic dataset 21.2 months. Thus publishers without traffic in 2013 are by average younger than those with traffic, if not contained in businessmodel Unknown. The boxplots in 3.5b indicate the same pattern, however the median age seems even lower. Moreover the spread of values is higher for those without traffic, they include more outliers and there is little change in the median and adjacent quartiles in comparison to the left plot (and without Unknown). Taking a closer look at the publishers with traffic, publishers from Email, Portal and Topic feature the highest median age apart from Unknown. Moreover they exhibit broader hinges, which correspond to the first and third quartiles (the 25th and 75th percentiles). Thus, in those businessmodels the variation is higher.



(a) Publishers with traffic in 2013. The mean age amounts to 63.

(b) Publishers without traffic in 2013. The mean age amounts to 62.2.

Figure 3.5: Boxplot for the distribution of age in months conditional on the businessmodel. The age is calculated with reference date 2014-01-01.

3.2.1 Classification of Active by Payment

To further structure the data and exclude publishers who are not relevant, the following is applied to obtain a classification of the payment in active and inactive. Total payment shall exist and exhibit values greater or equal to zero to be counted as active by payment. The number of publishers active by payment corresponds to the number of publishers, who are contained in the traffic data set, except one publisher, who had negative total payment in 2013.

	Impressions	Clicks	Orders	Total Payment
min	0.0	0.0	0.0	0.0
25%	4.0	0.0	0.0	0.0
med	146.0	10.0	0.0	0.0
mean	329093.2	17439.3	240.9	1370.9
75%	2864.5	104.0	1.0	5.0
max	4710600873.0	156926170.0	1353195.0	7050411.3

Table 3.1: Summaries over traffic variables for publishers active by payment in 2013. The numbers represent sums over the year 2013.

Table 3.1 gives the summaries for several key variables. The median for total payment in 2013 is zero, so at least half of all publishers had no overall payment in 2013, while they must had some traffic to appear in the data set. The maximum of the impressions is quite astonishing with one publisher having more than 4.7bn total impressions in 2013. The publisher who earned most in 2013, received more than seven million euros. Here the relation between the amount of impressions, clicks and resulting orders can be seen. The mean impression number is by far higher than the mean of clicks, which is larger than the mean of orders. The mean of the total payment surpasses the mean of orders, as the payout per order is generally higher than €1.

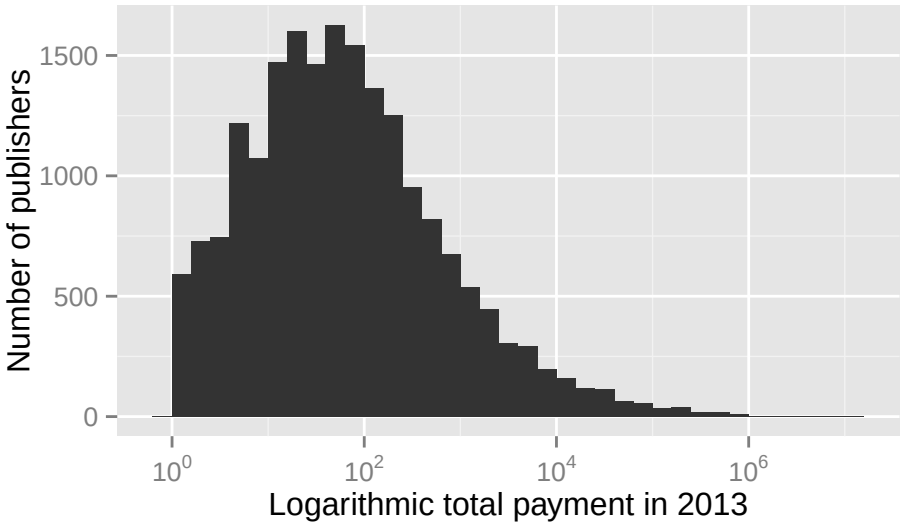


Figure 3.6: Histogram of logarithmic total payment in 2013 for publishers active by payment.

To depict the payment distribution graphically, a histogram of the logarithmic payment is given in 3.6. About 69% of publishers in the payment dataset earned not more than zero euro. Therefore the histogram is restricted to publishers, who received at least €1 in 2013. The

majority of these publishers earned less or about €100 for the total year. This is the so called longtail of publishers. They might not operate on a professional level, as their webpages have few visitors or bad converting advertisements. Moreover they could have registered with the network long ago and are not interested in putting much effort in it. Nevertheless, as the number of those publishers is considerably high, the network cannot neglect them. A simplified calculation shows: If assumed, that the network earns 20% on average on the publishers revenue and 10,000 publishers (i.e. about 50% of all publishers earning at least €1) earn €50 on average per year. Then rising that average income ten percent will increase the revenues for the network ten percent from €100,000 to €110.000. Hence, even publishers with small amounts are valuable, as they come on a large extend. However, the task of the network is to activate this longtail in order to leverage revenues. Returning to the histogram, the higher the total payment rises, the less publishers are included, resulting in a right-skewness.

Revenue share per businessmodel, KAM and status

The following plots provide an assessment of revenues shares relating to variables of the master data and their characteristics.

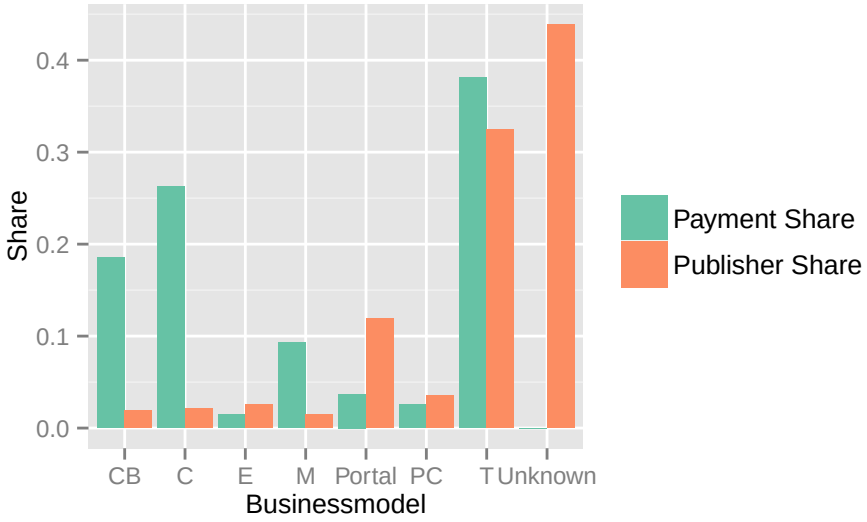
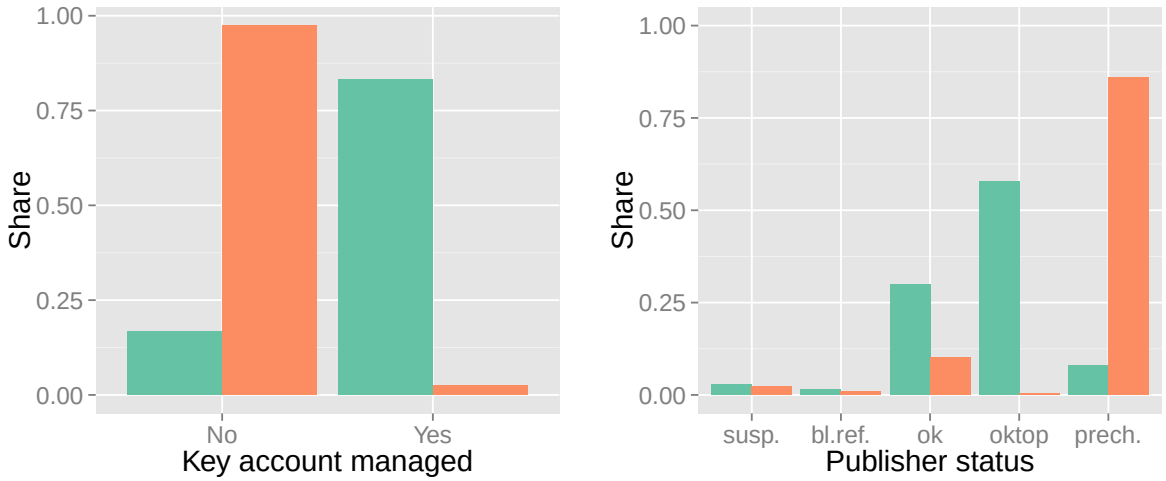


Figure 3.7: Comparison of share of total payment and share of number of publishers per businessmodel for publishers active by payment.

Figure 3.7 shows the comparison between the share of payment per businessmodel versus the share of publishers. As seen before publishers from businessmodel Unknown, which includes 44% of publishers with traffic in 2013, produced zero percent of the total payment in 2013. The biggest part of the payment, 38.1%, came from the Topic publishers, which included the second biggest amount of publishers with 32.5%. A huge part of total payment with 26.3% and 18.5% was produced by publishers from businessmodels Coupon and Cashback, respectively, while their share in the publisher base amounted to only 2% each. Publishers with businessmodel Media considerably contribute to the payment with a small share of publishers, while Portal falls behind despite a broader publisher base. An extended table for this plot with additional

numbers can be found in the Appendix at A.2.



(a) Overview of total payment 2013 by KAM

(b) Overview of total payment 2013 by status

Figure 3.8: Overview of total payment 2013 by KAM (left) and status (right). For both plots the share of the overall 2013 total payment and the share of the number of publishers are included. Green bars correspond to the payment share and red bars to the publishers share as in 3.7.

The share of total revenues in 2013 and share of total publishers for KAM and status are given in figure 3.8. As expected, while only three percent of publishers have a key account manager, they generate 83% of total payment, as can be seen in 3.8a. In 3.8b only the relevant statuses are pictured, excluding blocked, notchecked, blocked by precheck and publisher informed suspicious, as those values were at the utmost 0.3%. From the remaining statuses, oktop leads to the highest share of total payment in 2013 with 58%, followed by ok with 30%. Including about 86% of publishers with traffic in 2013, status prechecked generated eight percent of total 2013 payment. To further structure the data the publishers can be assigned in groups according to their total payment in 2013. The organisation of the publishers is orientated on the average gross margin groups, which are defined by the affiliate network. Those are calculated for the whole year by taking a medium margin of 20%. The resulting seven groups are given in table 3.2:

Group	Total payment range in euro
0	payment ≤ 0
1	$0 <$ payment ≤ 6000
2	$6000 <$ payment $\leq 30{,}000$
3	$30{,}000 <$ payment $\leq 60{,}000$
4	$60{,}000 <$ payment $\leq 120{,}000$
5	$120{,}000 <$ payment $\leq 300{,}000$
6	$300{,}000 <$ payment $\leq 600{,}000$
7	payment $> 600{,}000$

Table 3.2: Assignment of paymentgroups by sum of total payment in 2013.

69% of publishers are assigned to paymentgroup zero, while publishers from paymentgroup four until seven only accounted altogether for about 2% of all publishers.
Figure 3.9 displays the allocation of the share of total payment per paymentgroup for each busi-

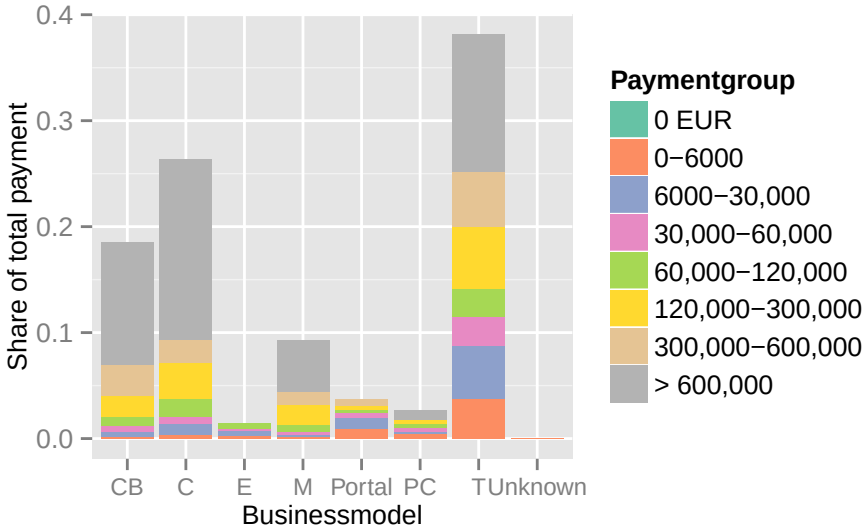
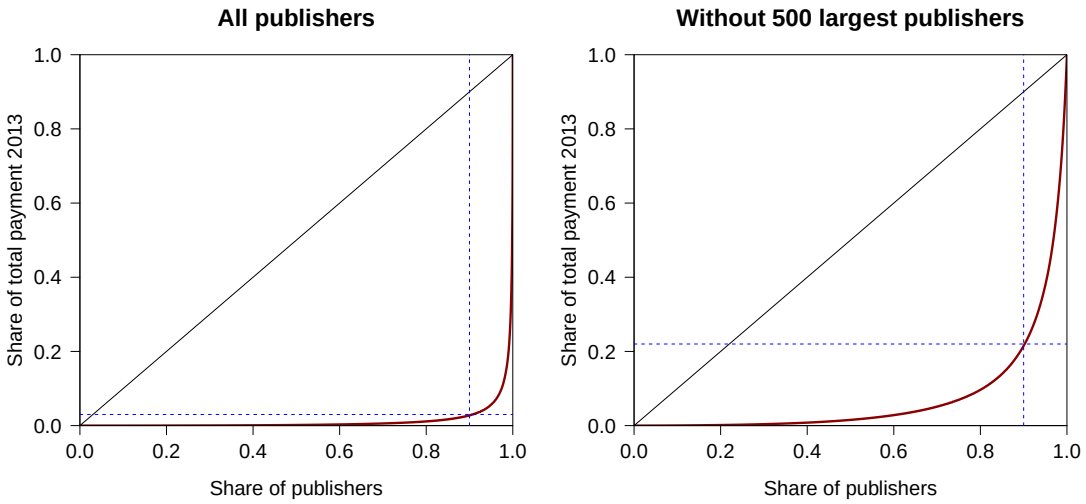


Figure 3.9: Share of total payment in 2013 by businessmodel and paymentgroup.

nessmodel. Table A.3 in the Appendix shows the corresponding values by paymentgroup. The interpretation is the following: businessmodel Cashback produced around 20% of the total payment in 2013. Those are mainly resulting from paymentgroup seven with incomes of more than €600,000 per year (this is only coming from two publishers and is accounting for 11.5% of the total payment in 2013). In the Coupon businessmodel the paymentgroup seven is represented by seven publishers, accounting for 17% of total payment. The segmentation in the businessmodel Topic is broader, thus also paymentgroups one to six contribute to the share of the total payment. In this businessmodel nine publishers are in paymentgroup seven. Unknown is just represented in paymentgroups zero and one, while the share on the total payment is not more than 0.03%. Overall 24 publishers are included in paymentgroup seven, which accounted for more than 47% of the total payment in 2013. All of those publishers are key account managed and while the majority has status oktop, three are labeled ok. Publishers from paymentgroup four to seven, which includes 195 publishers, accounted for about 80% of the generated income in 2013.

To extend the understanding of the distribution of the payment per publisher, it is useful to take a look at the Lorenz curve and the Gini coefficient. The Lorenz curve is a relative concentration measurement and can be used to measure inequality. Figure 3.10 pictures the Lorenz curve for both all publishers and all but the 500 largest publishers. The Lorenz curve is a graph that shows, for the bottom x% of publishers, the percentage y% of the total payment they had in 2013. The percentage of publishers is plotted on the x-axis, the percentage of payment on the y-axis. A perfectly equal payment distribution would be one, in which every subject has the same income. Thus a perfectly equal distribution can be depicted by the bisector, then the concentration is zero. This is clearly not the case here. For all publishers, about 90% of publishers account for just about 3% of the share of total payment in 2013, shown by the blue dotted lines. The Lorenz curve is used to calculate the Gini coefficient, which is the area between the bisector and the Lorenz curve, as a percentage of the area between the bisector and the abscissa. The high Gini coeffient of 0.97 corresponds to the Lorenz curve. The Lorenz curve is also shown for all but the



(a) The vertical line (blue dotted) intersects at 0.9 and the horizontal line intersects at 0.03. The corresponding gini coefficient amounts to 0.97.

(b) The vertical line (blue dotted) intersects at 0.9 and the horizontal line intersects at 0.22. The corresponding gini coefficient amounts to 0.85.

Figure 3.10: Lorenz curve for publishers, who earned at least one euro in 2013 (left) and all but 500 biggest of those publishers by payment (right).

500 biggest publishers (which relates to 99.3% of all publishers). Then the curve in 3.10b is less concentrated and 90% of those publishers account for about 22% of the share of the remaining total payment in 2013. The Gini coefficient here is 0.85.

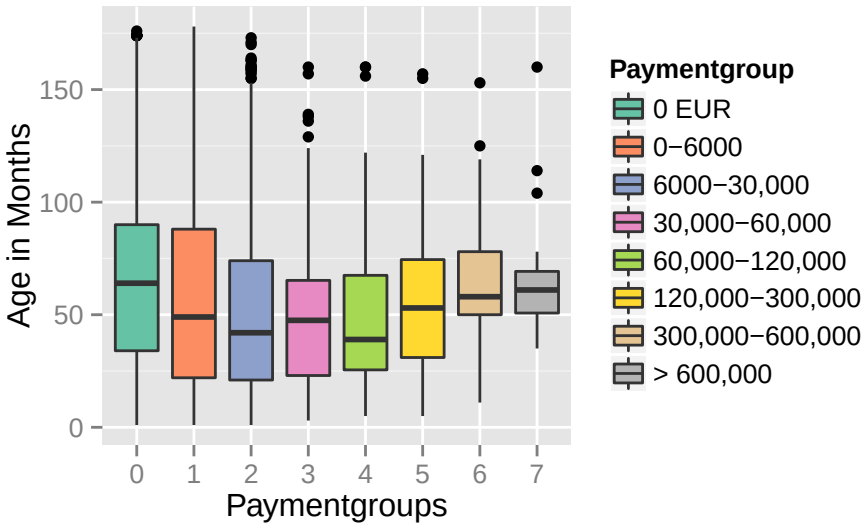


Figure 3.11: Boxplots comparing the distribution of age in months by paymentgroups with reference date 2014-01-01.

Returning to the paymentgroups, figure 3.11 gives an insight into the age of publishers per paymentgroup. The median age for paymentgroup seven is about the same than for paymentgroup zero. For both groups it is about 60 months. Also paymentgroup five and six exhibit median ages higher than fifty months. The boxes for paymentgroups zero to three are wider, as they

include a broader base of publishers.
Now, taking a closer look at the height of the payment, figure 3.12 shows boxplots for the logarithmic payment in 2013 by businessmodel. As the distribution of the data is widely spread, it is useful to use logarithmic terms for plotting and modeling. While the distribution and the median from Unknown is clearly smaller, the other businessmodels seem at a first glance similarly distributed. Their median total payment is about €100, the size of the hinges are alike and all show outliers. However, due to the fact that the plotted values are on a logarithmic scale, small deviations mark greater differences than perceived. Thus the higher values for the 75th percentiles of BM Cashback, Coupon and Media relates to their high number of top publishers. Moreover, as they do not include such a large number of publishers, this gives them greater weight than in businessmodel Topic.

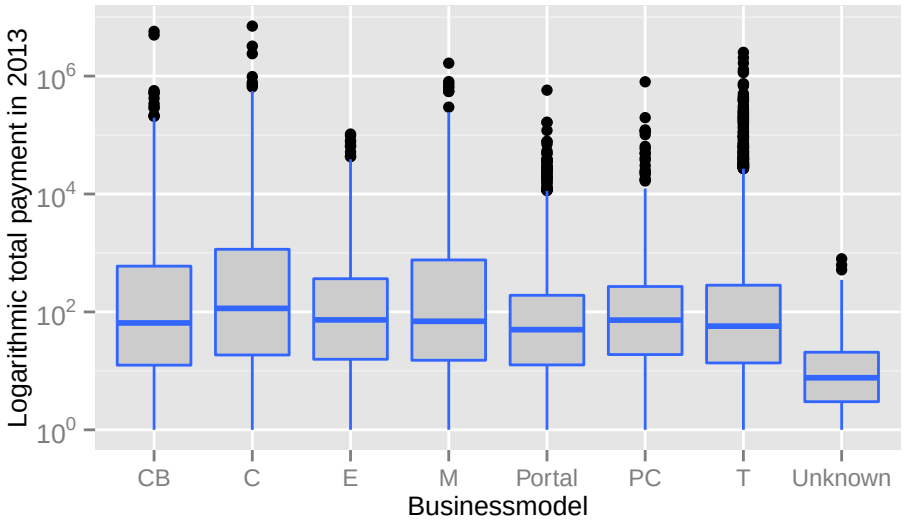
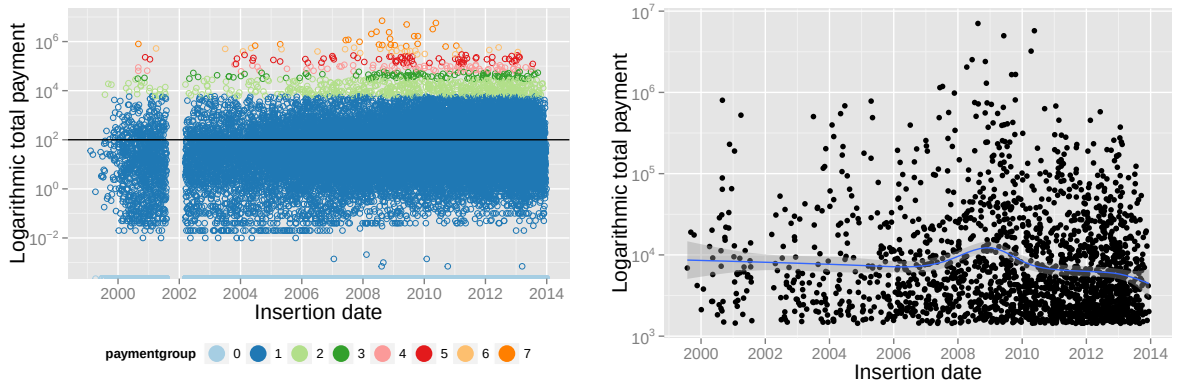


Figure 3.12: Boxplots for the distribution of logarithmic total payment in 2013 by businessmodel for publishers earning at least one euro in 2013. Note that the y-axis is labelled in the original data space and not on a logarithmic scale.

The question arises, when the publishers with the highest payment or number of orders have been registered as publishers in the affiliate network. Figure 3.13a therefore shows the relationship between date of registration and total payment in 2013. It additionally highlights the paymentgroups. The blank spot at the end of 2001 marks those excluded publishers, who had been inserted without insertion date and other variables. The publishers with payment in 2013 have been inserted in the network from February 1999 to December 2013. As can be clearly spotted the paymentgroups are layered evenly, so the distribution over all years seems more or less equal. Paymentgroup three tends to be inserted more frequently from 2008 on. Publishers from paymentgroup seven, with the highest total payment in 2013, have registered with the network from June 2004 to January 2011.
Figure 3.13b shows the logarithmic total payment in 2013 for the 2000 best earning publishers. Therefore it can be seen as a zooming into plot 3.13a, which causes the axis labels to change. As it can be hard to see exactly what trend is shown by the data, a smoothed line was added to the plot. A slight peek can be observed for the year 2009, then for publishers inserted after that



(a) All publishers by paymentgroup. The horizontal line is drawn at 100 Euro total payment.

(b) 2000 best earning publishers with smooth function for logarithmic total payment in 2013. A generalised additive model was used as a smoother.

Figure 3.13: Scatterplot of publishers insertion date and logarithmic total payment in 2013. Note the varying y-labels.

the smoothed line is declining. This seems reasonable as publishers, who have been inserted in 2013 have not been paid for the complete year. The peak is caused by the ten best performing publishers, who have been won as customers between 2008 and 2010.

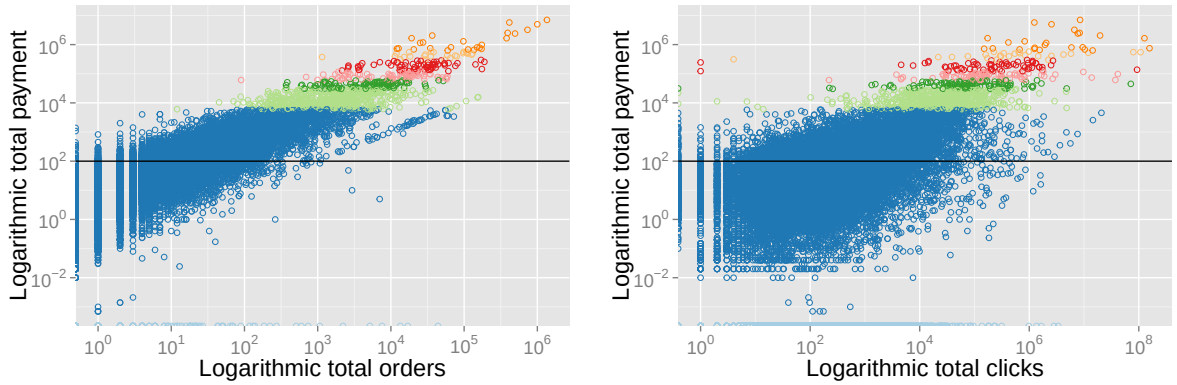


Figure 3.14: Scatterplot of logarithmic total payment with logarithmic total orders (left) and logarithmic total clicks (right). Paymentgroups are coloured as in 3.13. A horizontal line is drawn at €100 total payment in 2013. While the data is displayed on a logarithmic scale, the axis labels refer to the untransformed data. Note that the x-label differs.

Figure 3.14 shows the relationship between logarithmic total payment in 2013 and both logarithmic total orders and total clicks. The pearson correlation between orders and payment is positive and amounts to 0.89. With rising orders the payment also rises. The correlation between payment and clicks is less strong with a pearson coefficient of 0.2. For both plots it is clearly visable that publishers, who earned more money in 2013 tend to have a higher number of orders and clicks. Moreover many publishers with a considerably high number of clicks are unable to turn that into profit. This corresponds, among others, surely to the fact that clicks do not naturally lead to payment as orders do.

3.2.2 Classification of Active by Traffic

In the previous section, publishers who are contained and not contained in the traffic data set have been evaluated. Those have therefore been referred to as publishers with or without traffic. All but one of those publishers are active by payment. However, not all of them are active by traffic. Those will be the focus of the next section and are defined as stated below.

$$\text{Traffic}_i = \begin{cases} 0 & \text{if publisher i has (Clicks+Impressions)} < 20 \\ 1 & \text{if publisher i has (Clicks+Impressions)} \geq 20 \end{cases}$$

trafficgroup	number	share	MeanImpr.	MeanClicks	MeanOrders	MeanPayment
0	18542	0.27	3.6	1.4	1.8	8.7
1	49142	0.73	453256.9	24018.6	331.1	1884.6

Table 3.3: Number and share of publishers, as well as mean of variables per trafficgroup

Table 3.3 shows the assignment to both trafficgroups and the resulting means for the traffic variables. Around one third of all publishers in the data set are assigned to the inactive traffic group. So they are contained in the traffic dataset, but achieved less than twenty impressions and clicks in 2013. They have been included in the payment classification by payment, as they earned at least zero euros in 2013. The summaries for publishers, who are active by traffic are shown in table 3.4.

	Impressions	Clicks	Orders	Total Payment	Age in Months
min	0.0	0.0	0.0	0.0	1.0
25%	82.0	5.0	0.0	0.0	27.0
med	746.0	35.0	0.0	0.0	56.0
mean	453266.0	24019.0	331.1	1884.9	61.0
75%	7256.0	243.0	4.0	21.0	88.0
max	4710600873.0	156926170.0	1353195.0	7050411.3	178.0
nmiss	0.0	0.0	0.0	0.0	0.0

Table 3.4: Summaries of different variables of active publishers by traffic.

Figure 3.15 shows an overview of the shares of clicks, orders and total payment in 2013 in comparison to the share of publishers. The relating table can be found in the Appendix at table A.4. Again it can be easily seen that businessmodel Unknown, with the highest share of publishers, produces basicly no return. Businessmodel Media has the highest share of clicks, but that does not translate to relatively higher orders or payments. With a relatively low rate of clicks, both Cashback and Coupon achieve shares of both orders and payment between eighteen and thirtyone percent. Those businessmodels generate this high share of payment, with a very low share of publishers. Businessmodel Topic's share of the payment is with 38% the highest, achieved with a relatively high number of orders, clicks and publishers.

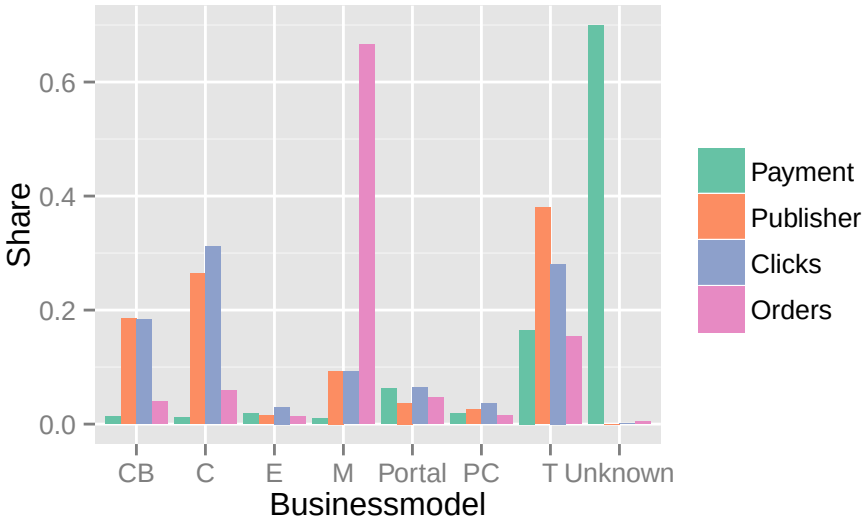


Figure 3.15: Comparison of the share of clicks, orders, payment and number of publishers by businessmodel for 2013.

3.3 Partnerships

Each publisher can have several partnerships with advertisers. In total 185,022 publishers have partnerships. In combination with all publishers, who have been active by payment in 2013, the merged dataset includes 63,814 publishers. The relationship between logarithmic total payment and number of existing partnerships (PS) in 2013 is pictured in 3.16. Additionally the publishers are highlighted by their respective paymentgroup. The assumption that publishers with a higher number of partnerships in 2013 have higher returns in 2013 in relation to those with fewer partnerships, cannot be verified through this plot. The pearson correlation coefficient of 0.11 confirms this. There are some publishers especially in paymentgroup one and zero, who earn money without existing partnerships in 2013. Due to data warehouse issues some publishers are marked to have no partnership (this number amounts to 3378 publishers). The highest amount for a publisher without existing partnerships due to the data in 2013 and of paymentgroup three earned in total €33,827. This publisher had two accepted and also two deleted partnerships before 2013. As taken a closer look at this publisher, the partnerships are still existing. This shows, that the existing partnerships cannot always be calculated correctly. However, due to mostly small deviations, the numbers can be used be taken into account.
To evaluate possible differences in the number of partnerships for each businessmodel, figure 3.17 shows boxplots per businessmodel of the mean number of monthly existing partnerships in 2013. The highest amount of partnerships a publisher reached in 2013 was 1497 in December 2013. Most publishers exhibit about fifty partnerships per month. While Topic and Unknown are below that mark, Coupon and Price Comparison exceed that value. Their businessmodels aim to reach as many people as possible, and their websites are usually not focused on a specific theme as in businessmodel Topic.

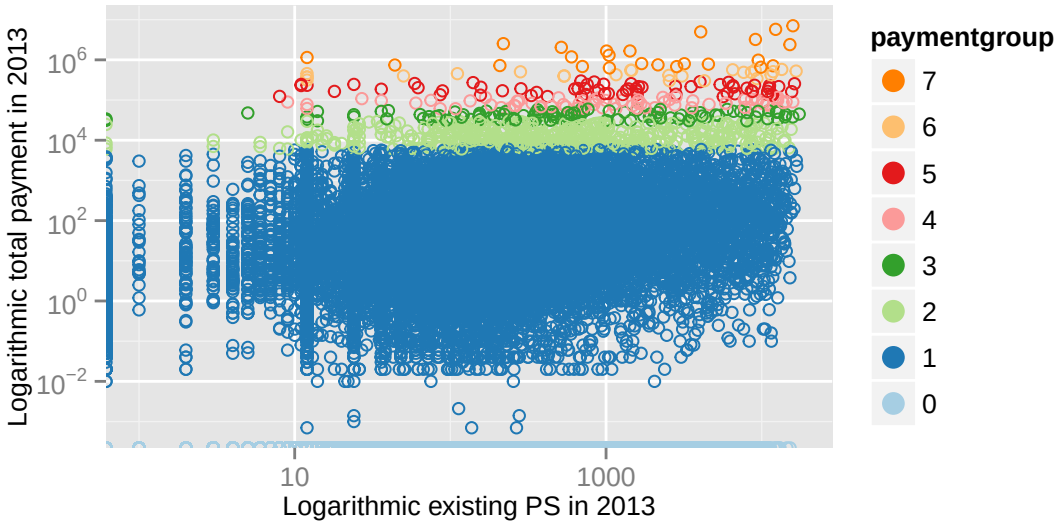


Figure 3.16: Scatterplot of logarithmic total payment versus logarithmic existing partnerships per publisher in 2013 by paymentgroup. While the data is displayed on a logarithmic scale, the axis labels refer to the untransformed data.

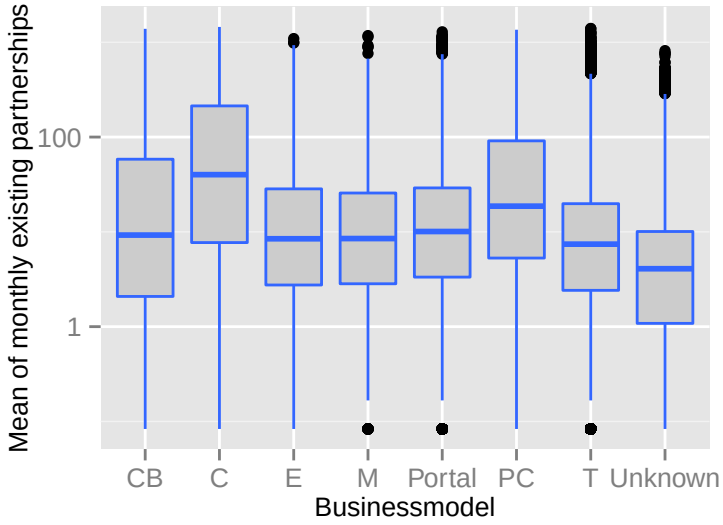


Figure 3.17: Boxplot of mean monthly number of existing partnerships in 2013 per businessmodel. Axes are in untransformed format.

3.4 Change in Variables

Until now, the focus was on the publishers last and therefore current status in the data warehouse. However, it is also of interest to which extend the publisher changed its variables along his membership. Changes within the characteristics of a publisher are recorded in the data warehouse. As this was established in 2012, only changes after that could be recorded. Several characteristics of the publisher can be changed by the publisher himself or by the company. In the following the focus is on changes in the businessmodel, the status and the KAM variable. Figure 3.18 shows the number of changes, which publishers had. The first change recorded in the

data warehouse was on the 16th of december 2011. Most entries in the data warehouse show no change to previous values of publishers, businessmodel, KAM or status. The amount of changes in the ID of publishers equals all 427,178 unique publishers in this data set. Most changes occur for switching the status of a publisher (44,526 changes), followed by changes in businessmodel (39,873). Naturally, as only a small part of all publishers are key account managed, changes in this variable emerge 1,547 times.

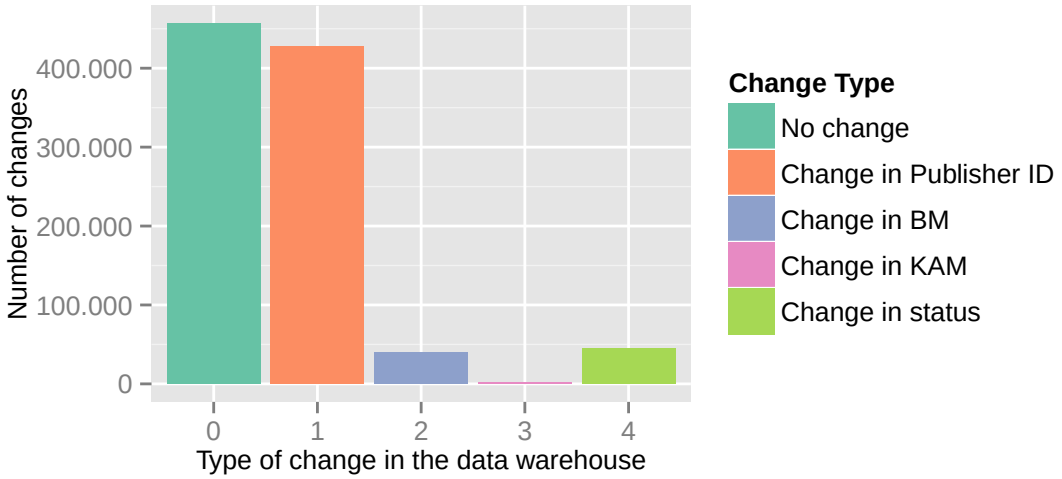


Figure 3.18: Counts of changes in the characteristics of a publisher within the data warehouse. Note that as the data warehouse was established in 2012, only changes after that could be recorded

In a next step the changes of status and businessmodel are analysed in detail. It is to check, from which initial point the change was conducted and to which status or businessmodel this led. Figure 3.19 therefore sheds light into the changes within a publishers status. The x-axis shows the initial status and the bars are marked in the colour, to which the status switches. Most changes, start with notchecked and pass to status prechecked or blocked by precheck. This is the automatic procedure in the registration process. Therefore most publishers change from status notchecked. Moreover statuses migrate from prechecked to suspicious or blocked&refunded and the other way around. Only a slice of changes occur from the other statuses.
Figure 3.20 shows the transition of businessmodel changes in the data warehouse. The first entry of change of a publishers businessmodel was on the 17th December 2011. While there is no change in the businessmodel in 96% of all entries in the data warehouse, the plot shows the cases, when a change occurs. As expected, most publishers change from businessmodel Unknown mostly to businessmodels Portal and Topic, but also to all other businessmodels. The next biggest number of publishers change from businessmodel Email to Topic and Portal. A small number changes from PC and Portal to Topic. Changes starting from businessmodel Coupon, Cashback, Media or Topic are very rare.

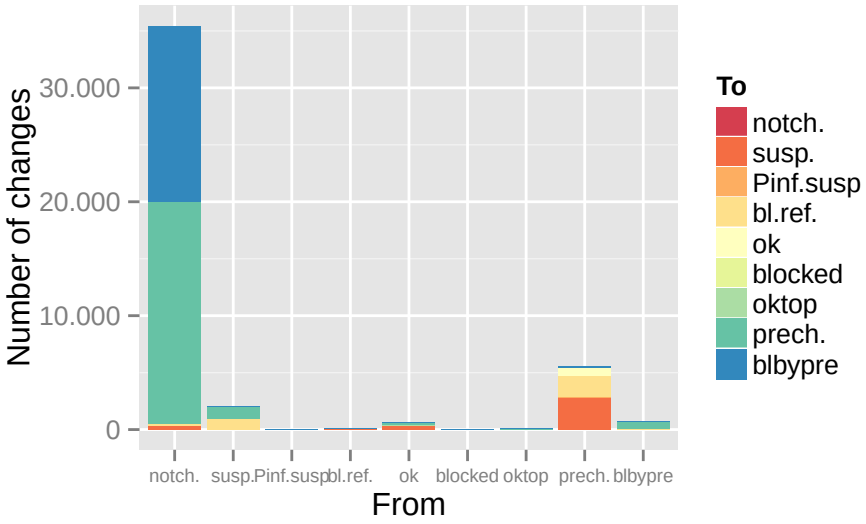


Figure 3.19: Changes in the status of a publisher within the data warehouse. The values on the x-axis show the previous status, then the bars are coloured by the following status. Note that as the data warehouse was established in 2012, only changes after that could be recorded

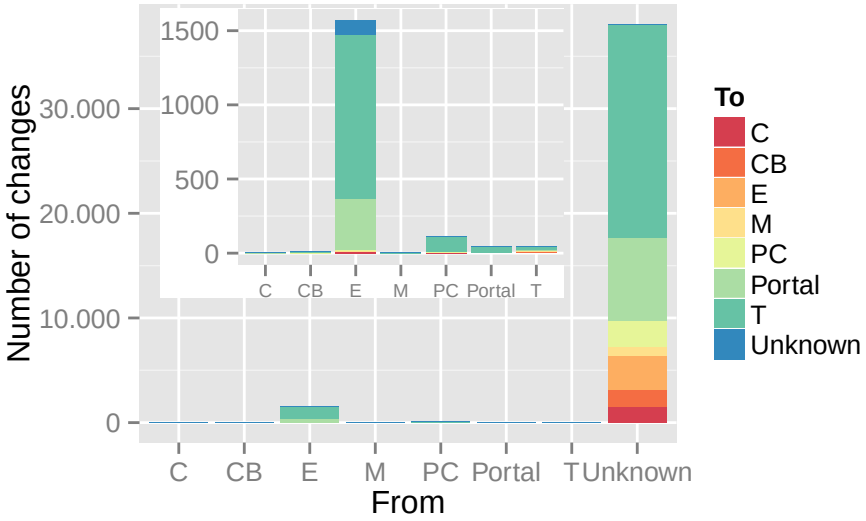


Figure 3.20: Changes in the businessmodel of a publisher within the data warehouse. The big plot shows all businessmodels, while the small plot inside focuses on the allocation for all businessmodels but Unknown.

Chapter 4

Theoretical Background

In this chapter all used methods for the data analysis of chapter 5 are reasoned and described theoretically. Models are presented in the order, in which the data analysis part covers them.

4.1 Generalized Linear Models

Linear models are suitable for regression analysis with a continuous and at least approximately normal response variable. However, in many applications the response is not a continuous variable, but rather categorical, binary or a count variable. Generalized linear models (GLMs), which have been introduced by Nelder und Wedderburn (1972) and further developed by McCullagh und Nelder (1989), allow for different response distributions apart from the normal distribution. They provide a very flexible and powerful framework for the application of regression models to a variety of non-normal response variables, for example the logistic regression for binary responses and the Poisson regression for count data. As this is the main application for GLMs in this work, 4.1.1 starts with a description of regression models for binary responses. The general introduction into GLMs is offered in section 4.1.3. This section is mostly based on Fahrmeir et al. (2013) and Fahrmeir und Tutz (1994).

4.1.1 Binary Regression Models

Categorical regression aims to explain the link between covariables considered as the independent variables and the response as the dependent variable. They thus have the same objectives as metric regression but differ from classical normal regression in several ways. In comparison to classical linear regression, in categorical regression modeling the response variable can only take a limited number of values. Binary regression is the most simple case, where the dependent variable y_i takes only two values. Lets assume that (ungrouped) data on n objects or individuals are given in the form $(y_i,x_{i1},...,x_{im}),i = 1,...,n$ with the binary response y coded by 0 and 1 and covariates denoted by $x_1,...,x_m$. The response variables are assumed to be (conditionally) independent given the covariates. The distribution of the binary random variable is fully characterized by the probability

$$\pi_i = P(y_i = 1) = E(y_i)$$

for the outcome $y_i = 1$ and given values of the covariates $x_{i1},...,x_{im}$. Models for binary and

binomial responses are determined by relating the response probability π_i to the linear predictor η_i via some response function

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}) \tag{4.1}$$

where h is a strictly monotonically increasing cumulative response function. Equation 4.1 can also be expressed as

$$\eta_i = g(\pi_i)$$

with the inverse function $g = h^{-1}$. Within the frameworks of GLMs, h is known as the *response function* and g is called the *link function*. In the context of binary regression models, Logit and Probit models are the most widely used models. In the following, the focus will be on Logit models.

4.1.2 The Logit Model

In our case we want to determine what influences the confirmed payment in total. Therefore we define a binary variable for the monthly payment, with the categories being either "payment" ($y = 1$) or "no payment" ($y = 0$). For dependent variables with two categories, i.e. $y \in \{0,1\}$, the commonly known logistic regresssion model is used. The aim of a regression analysis with binary responses is to model the probability

$$P(y = 1) = P(y = 1|x_1,...,x_m) = \pi$$

in the presence of covariates. The expected value and variance of the binary variable y are given by

$$E(y) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi,$$

$$Var(y) = (1 - \pi)^2 \cdot \pi + (0 - \pi)^2 \cdot (1 - \pi) = \pi \cdot (1 - \pi)$$

The mean of the binary distribution is represented by the response probability π and the variance is fully determined and depends on π with minimal value at $\pi = 0$ and $\pi = 1$ and maximum at $\pi = 0.5$.
The logistic distribution function is given by

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{4.2}$$

The *link function* and the *linear predictor* determine the general form of the parametric binary regression model. With several continuous $(x_i) = (1,x_{i1},...,x_{im})$ covariates the linear predictor can be written as

$$\eta_i = \beta_0 + \beta_1 x_{i1} 1 + ... + \beta_m x_{im},$$

which then yields the logit model

$$\pi_i = P(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \tag{4.3}$$

then equivalently to 4.3 the logit link function is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}$$

This yields a linear model for the logarithmic odds. Transformation with the exponential function gives

$$\frac{\pi_i}{1-\pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot ... \cdot \exp(\beta_m x_{im}).$$

For a continuous variable x_j, β_j describes the additive change in logits of an increase in x_j of one unit on the logit in relation to the reference, if all other variables are kept fixed. $\exp(\beta_j)$ indicates the multiplicative change in odds for $x_j \to x_j + 1$. The reference category is important for model interpretation, changing it results in different estimates for β. For each of the response categories, one linear predictor η_r is estimated. The linear predictor equals 0 for the reference category, which is important for the identifiability of the parameters. For a general introduction into regression for categorical data see Tutz (2012). For the analysis in this work a generalized linear model with the logit link is used.

4.1.3 Generalized Linear Models

The basic structure of the generalized linear model is

$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta}, \tag{4.4}$$

where $\mu \equiv \mathbb{E}(Y)$, g is a smooth monotonic link function, $\boldsymbol{X}$ is a model matrix, and $\boldsymbol{\beta}$ is the corresponding vector of unknown parameters. Additionally, a GLM typically makes the distributional assumptions, that the Y follow some exponential family distribution and are mutually independent. The exponential family of distributions contains many distributions, such as the Binomial, Poisson, Gamma and Normal distribution. Generalized linear models are specified in terms of the linear predictor, $\boldsymbol{X}\boldsymbol{\beta}$, thus many of the general concepts of linear modeling can be transferred with some modification. In terms of model formulation GLM's differentiate to that effect, that they need a link function and a distribution to specify the model. The linear model is a special case of GLM's, when the identity link is selected in combination with the normal distribution. A distribution belongs to the exponential family of distributions, if its probability density function can be written as

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

where θ is the natural parameter of the distributions, ϕ an additional scale or dispersion parameter, $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of exponential family, and ω is a known value (usually a weight). The expected value and variance are given by:

$$\mathbb{E}(y|\boldsymbol{x}) = \mu = b'(\theta)$$
$$Var(y|\boldsymbol{x}) = \phi b''(\theta)/\omega$$

Thus, a specific GLM is completely determined by the type of the exponential family, the choice of the link or response function, and the definition and selection of covariates.

4.1.4 Maximum likelihood estimation

In generalized linear models regression analysis is based on likelihoods. The likelihood for the parameter vector β for i.i.d distributed random observations y_i is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) \tag{4.5}$$

In equation 4.5 the likelihood $L(\boldsymbol{\beta})$ equals the product of the densities of y_i, which depend on the unknown parameter $\boldsymbol{\beta}$ through $\pi_i = E(y_i) = h(x_i'\beta)$. With maximization of the log-likelihood $l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta}))$ the ML estimator $\hat{\boldsymbol{\beta}}$ can be obtained. Unlike as for linear models, the ML estimator has no closed form. Thus, the solution is carried out numerically, using Fisher scoring or the Newton-Raphson Approach. Therefore the score function $\boldsymbol{s(\beta)}$ and the observed or expected Fisher matrix $\boldsymbol{F(\beta)}$ are required. The ML estimator is the solution of

$$s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

for the score function given by

$$s(\boldsymbol{\beta}) = \sum x_i \frac{h'(\eta_i)}{\sigma_i^2}(y_i - \mu_i) = \boldsymbol{X'D\Sigma^{-1}(y - \mu)},$$

where $\boldsymbol{D} = diag(h'(\eta_i),...,h'(\eta_n))$, $\boldsymbol{\Sigma} = diag(\sigma_1^2,...,\sigma_n^2))$ and $\boldsymbol{\mu} = (\mu_i,...,\mu_n)'$. The Fisher matrix is given by

$$\boldsymbol{F(\beta)} = \sum x_i x_i' \tilde{\omega_i} = \boldsymbol{X'WX}$$

Here, $\boldsymbol{W} = diag(\tilde{\omega_1},...,\tilde{\omega_n})$ is the diagonal matrix of working weights $\tilde{\omega_i} = (h'(\eta_i))^2/\sigma_i^2$. The ML estimator $\hat{\boldsymbol{\beta}}$ is obtained iteratively using Fisher scoring in form of iteratively reweighted least squares estimates.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X'W^{(t)}X)^{-1}X'W^{(t)}\tilde{y}^{(t)}}, \quad \text{for iterations} \quad t = 0,1,...$$

4.1.5 Generalized Additive Models

A generalized additive model (GAM) is a special case of the GLM, in which the linear predictor is given by a sum of smooth functions of the covariates, plus a conventional parametric component of the linear predictor. The basic structure of the GAM is an extension of the GLM with smooth functions

$$g(\mu_i) = \boldsymbol{X_i\beta} + f_1(x_{i1}) + ... + f_m(x_{im}) + \epsilon_i \tag{4.6}$$

where the description is analogous to 4.4 and the f_j's are unspecified smooth functions of the covariates x_j. Model estimation is by penalized versions of the least squares or maximum likelihood/IRLS methods, by which the generalized linear models are fitted. This is possible, since the GAM is simply a GLM, with associated penalties. However, the penalization of the fitting process has to be chosen. This is captured in the section about splines in 4.1.6.

Additive Logistic Regression

For two-class classification, recall the logistic regression model for binary data, discussed in section 4.1.2. The mean of the binary response $\mu(x) = P(Y = 1|x)$ is related to the predictors via a linear regression model and the logit link function. Then the generalized additive logistic model has the form

$$log\left(\frac{P(Y=1|x)}{P(Y=0|x)}\right) = \boldsymbol{X\beta} + f_1(X_1) + ... + f_m(X_m).$$

A simple example is:

$$logit(\mathbb{E}(y_i)) = f_i(x_{1i}) + f_2(x_{2i})$$

where the (independent) response variables y_i is binary, and f_1 and f_2 are smooth functions of covariates x_1 and x_2.

4.1.6 Splines

The smooth components of the GAM model are estimated by penalized regression smoothers, which are based on splines. A spline curve is a is piecewise polynomial curve, i.e. it joins several polynomial curves. The *knots* of the spline are the points at which the sections join. Each f_j can be represented using a linear basis expansion:

$$f(x) = \sum_{k=1}^{q} b_k(x)\beta_k,$$

where f is an element of a space of functions, $b_k(x)$ is the k^{th} basis function for some values of the unknown parameters β_k. A broad type of penalized regression smoothers can be used. Examples are regression splines, cubic splines or p-splines. The penalized regression spline fitting problem is to miminize

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

with respect to $\boldsymbol{\beta}$ and matrix of unknown coefficients $\boldsymbol{S}$. The penalized least squares estimator of β is then given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X^T X + \lambda S)^{-1} X^T y}.$$

Choosing the smoothing parameter λ can be critical. When λ is too small, then the data will be underfitted, and if it is too high then the data will be overfitted. In both cases the spline

estimate $\hat{f}$ will not be close to the true function f. A possible approach is the use of cross validation, where the model is chosen in order to maximize the ability to predict data to which the model was not fitted. This can be done using the *generalized cross-validation* (GCV) score

$$V_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{A})]^2}$$

where $\hat{f}$ is the estimate from fitting to all the data, $\boldsymbol{A}$ ist the corresponding influence matrix, and $tr(\boldsymbol{I} - \boldsymbol{A})/n$ is the mean weight, chosen to arrive at the GCV score. For further information on splines in the GAM context see Wood (2006).

4.2 Bootstrap

For both the GLM and for GAM, problems might occur, if the data inherits a dependent structure. Then the assumption of independency is violated. To correctly assess the variation of the mean parameters, some adjustment is necessary to account for the correlation present in the data. To avoid this a bootstrap appraoch can be helpful. With correlated response data, we can still use the usual models to fit the GAM to estimate $\hat{f}_j$'s. However, the standard error function $SE(\hat{f}_j)$ of each $\hat{f}_j$ is not valid any more due to the nature of the correlated data. A valid estimate of $SE(\hat{f}_j)$ can be obtained from the bootstrap estimates of f_j. Bootstrap is a resampling method, which involves repeatedly drawing samples from a training set. Then the model of interest can be refitted on each sample, in order to obtain additional information about the fitted model. The most simple case is the one-sample problem with $X = (X_1,...,X_n)$ where X_i follows an underlying unknown distribution F. We are interested in some statistic $T(X)$. With the observed data $x = (x_1,...,x_n)$, an estimate of the corresponding statistic can be obtained. Therefore a bootstrap sample is received from drawing n times *with* randomly drawing from $x = (x_1,...,x_n)$, leading to

$$x^* = (x_1^*,x_2^*,...,x_n^*) \rightarrow T(x^*).$$

With the calculated statistics $T(x^{*1}),...,T(x^{*B})$ from each bootstrap replication B, statements about the distribution of T can be made. For example about the average of the bootstrapped statistics of T^*

$$\bar{T}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} T(x^{*b}).$$

and the estimated bootstrap variance of T^*

$$\text{Var}_F(T) \approx \widehat{\text{Var}}_{\text{boot}}(T) = \frac{1}{B-1} \sum_{b=1}^{B} (T(x^{*b}) - \bar{T}_{\text{boot}})^2$$

which estimates the sampling variance of T.
Through such an approach information can be obtained, that would not be available from fitting the model only once using the original training sample. Resampling approaches can be computationally expensive, because they involve fitting the same statistical method several

times using different subsets of the training set. However, with rising computer power, the computational requirements of resampling methods can be met. The underlying idea of the bootstrap is to receive distinct data sets by repeatedly sampling observations from the original data set, instead of repeatedly obtaining independent data sets from the population (Hastie et al., 2009).

The boostrap was introduced by Efron (1979). The method described above, in which the sampling is with replacement from the training data, is called the *nonparametric bootstrap*. Efron's non-parametric bootstrap treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation. The nonparametric bootstrap allows to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly. As suggested by Efron und Tibshirani (1994) usually it does not need a very large bootstrap replication number, if the goal is to estimate standard errors. Of course, the larger B, the more accurate the resulting bootstrap standard error estimate. From theoretical considerations it follows, that $B = 200$ in the one-sample problem is usually sufficient to estimate a standard error. For confidence intervals significantly more replications are needed ($B \sim 2000$).

Bootstrap intervals

Several intervals in the bootstrap context can be selected, for example normal or percentile intervals. The bootstrap *percentile interval*, uses the empirical distributions of the estimates $\hat{\theta}^*$ from the B bootstrap replications. After drawing the bootstrap replications, the $\hat{\theta}^*(b)$'s are sorted according to size $\hat{\theta}^*_{(1)}$,...,$\hat{\theta}^*_{(B)}$. Then $B\alpha$ and $B(1-\alpha)$ are calculated, yielding ${\hat{\theta}^*_{(B)}}^{\alpha}$ and ${\hat{\theta}^*_{(B)}}^{(1-\alpha)}$. Then an approximate confidence interval for $(1-2\alpha)$ is given by

$$[\hat{\theta}_{(\text{lower})},\hat{\theta}_{(\text{upper})}] = [{\hat{\theta}^*_{(B)}}^{\alpha},{\hat{\theta}^*_{(B)}}^{(1-\alpha)}]$$

The percentile method is invariant to (strictly monotonic) transformations and is range-persaving, i.e. the percentile interval lies in the permitted range of the parameter. However, intervals tend to be over-optimistic.

Leave One Out Bootstrap

The bootstrap error estimator tends to be upward-biased as the training sets contain only 63.2% of the observations on average. An improvement to the normal bootstrap method is the leave one out bootstrap (.632+ estimator) which was proposed by Efron und Tibshirani (1997). The idea is to include only those cases in the estimation of the prediction error, which are not included in the respective bootstrap sample. The probability that a case is in the bootstrap sample is given by

$$1 - \left(1 - \frac{1}{n}\right)^n \approx 0.632.$$

The leave one out bootstrap estimator offers an improvement by mimicking cross-validation and

is defined as:

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i,\hat{f}^{*b}(x_i)) \tag{4.7}$$

where $C^{-i} = \{b : (y_i,x_i) \notin S_b\}$ and $S_b,b = 1,...,B$ are the bootstrap samples.
For every observation i use only bootstrap samples C^{-i}, which do not contain this observation. The average number of distinct elements in the $S_b's$ retained in $\widehat{Err_i}^{(1)}$ is about $0.632 \times N$. To correct the upward bias in $\widehat{Err_i}^{(1)}$ the .632 estimator can be used and is defined as

$$\widehat{Err}^{(.632)} = .368 \times \overline{err} + .632 \times \widehat{Err}^{(1)} \tag{4.8}$$

where $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\hat{f}(x_i))$ is the training error.

4.3 Linear Mixed-Effects Models

A common problem in modeling data is the presence of correlation among subjects or units. Typical examples for this kind of data are clustered data, longitudinal and repeated measurements. When the dependent variable is measured repeatedly for each subject or unit, the analysis of this data must take the dependence among a subjects multiple measurements into account. As for the same individual several observations are made, the measurements might be correlated rather than independent, as it is supposed for most models. Therefore repeated measurement data requires a special data handling. Models for those data need to include explanatory variables like in the usual multiple regression model, but in addition parameters that account for the correlational structure of the repeated measurements. While the latter are often regarded as nuisance parameters, the explanatory variables are generally of most interest. To avoid misleading inferences about these parameters, an appropriate model for the correlational structure of the repeated measures is necessary. A comprehensive overview of methods for the analysis of repeated measurements data can be found in Davis (2002).
A general modeling framework for many of these problems are mixed-effect models. Those describe, like other types of statistical models, a relationship between a response and some of the explanatory variables, that have been measured or observed along with the response variable. Mixed-effects models or, shorter, mixed models are statistical models that include both fixed-effects parameters and random effects. While the former are indeed parameters in the statistical model, random effects are unobserved random variables. Those can be regarded as additional error terms, which account for correlation among observations within the same cluster. A particular class of mixed-models is the linear mixed-effects model or equivalently linear mixed model (LMM).

4.3.1 General Linear Mixed Model

A general linear mixed model is given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \tag{4.9}$$

with the distributional assumption

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{pmatrix} \right) \tag{4.10}$$

In this model, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices, $\boldsymbol{\beta}$ is a vector of fixed effects, and $\boldsymbol{b}$ is a vector of random effects. The covariance matrices for $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are assumed to be nonsingular, and therefore positive definite, and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent.

4.3.2 Linear Mixed Models for Longitudinal and Clustered Data

The LMM for longitudinal and clustered data is a special case of the general LMM, where repeated measurements of the response variable y_{ij} occur. Those are suitable for modeling and analyzing data structures with one grouping level given by the individual or the cluster to which the observation belongs. The repeated measurements for each subject can be regarded as clusters or groups. Let $\boldsymbol{y_i} = (y_{i1},..,y_{in_i})$ be the vector of responses for individual or cluster $i = 1,...,n$.The linear mixed-effects model for longitudinal and clustered data expresses the n_i-dimensional response vector $\boldsymbol{y_i}$ for individual i:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}, \quad i = 1,...,n \tag{4.11}$$

where $\boldsymbol{\beta}$ is the m-dimensional vector of fixed-effects, $\boldsymbol{b_i}$ the $(q+1)$-dimensional vector of individual- or cluster-specific effects. $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are design matrices constructed from known covariates of dimension $n_i \times m$ and $n_i \times (q+1)$, respectively, and ϵ_i is a n_i-dimensional within-group error vector. The random-effects $\boldsymbol{b_i}$ and within-group errors $\boldsymbol{\epsilon_i}$ are assumed to be independent for different groups and from each other for the same group. The component $\boldsymbol{X_i} \times \boldsymbol{\beta}$ is the overall or fixed component and $\boldsymbol{Z_i} \times \boldsymbol{b_i}$ is the subject specific or random effect. The matrices $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ may, or may not, contain the same explanatory variables.
For $\boldsymbol{b_i}$ and $\boldsymbol{\epsilon_i}$, the following distributional assumptions hold:

$$\boldsymbol{b_i} \sim N(\mathbf{0},\boldsymbol{D}), \quad \boldsymbol{\epsilon_i} \sim N(\mathbf{0},\boldsymbol{\Sigma_i})$$

For i.i.d. errors, $\boldsymbol{\Sigma}$ simplifies to $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. The covariance matrix of $\boldsymbol{D}$ for $\boldsymbol{b_i}$ and $\boldsymbol{\Sigma}$ for $\boldsymbol{\epsilon_i}$ are assumed to be nonsingular and therefore positiv definite, which is to say that all its eigenvalues must be strictly positive. $\boldsymbol{Z}$, $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ define the covariance structure for $\boldsymbol{y}$.

$$\boldsymbol{y_i} \sim N(\boldsymbol{X_i}\boldsymbol{\beta},\boldsymbol{V_i}), \quad \text{where} \quad \boldsymbol{V_i} = Cov(\boldsymbol{y_i}) = \boldsymbol{Z_i}\boldsymbol{D}\boldsymbol{Z_i'} + \boldsymbol{\Sigma_i}. \tag{4.12}$$

So including random-effects has an effect on the structure of the covariance matrix $\boldsymbol{V_i}$. When $\boldsymbol{\Sigma_i} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{Z_i} = 0$, the mixed model reduces to the standard linear model. In the model terms of the LMM the covariance matrices of the random effects $\boldsymbol{D} = Cov(\boldsymbol{b_i})$ and the error terms $\boldsymbol{\Sigma} = Cov(\boldsymbol{\epsilon_i})$ are initially arbitrary. Correlation between repeated observations within individuals is caused by the common vector b_i of random effects, as is illustrated for simple random intercept and slope models in the following.

Random Intercept Model

The random intercept model is among the most simple mixed models. In this model, the fact that we have repeated measurements $j = 1,...,n_i$ on the same individual or cluster i is taken into account. Through partitioning the total residuals, that are present in the usual linear regression model, into a subject-specific random component b_i, that is constant over time, plus an error term ϵ_{ij}, that varies randomly over time, some correlational structure for the repeated measures is introduced. The random intercept model assumes, that all variability in subject specific slopes can be attributed to treatment differences. Then the intercepts are subject specific, but the slopes are the same. If the design vector is $\boldsymbol{z'_{ij}} = 1$ and the design matrix is respectively $\boldsymbol{Z_i} = \boldsymbol{1_i} = (1,1,...,1)'$ then the random intercept model results as

$$y_{ij} = \boldsymbol{x'_{ij}}\boldsymbol{\beta} + b_{0i} + \epsilon_{ij}, \quad b_i \overset{i.i.d}{\sim} N(0,\tau_0^2) \tag{4.13}$$

where individuals differ through their specific intercepts. In combination with $\boldsymbol{\epsilon_i} \sim N(\boldsymbol{0},\sigma^2\boldsymbol{I_{n_i}})$ the model leads to a marginal covariance structure, which implies a constant correlation structure of the target variable. The repeated measurements y_{ij} for subject i are then correlated with the within-subject correlation coefficient

$$Corr(y_{ij},y_{il}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} =: \rho \geq 0, \ j \neq l \tag{4.14}$$

The higher the random effects variance τ_0^2 in relation to the error variance σ^2, the stronger is the within-subject correlation.

Random Intercept and Slope model

The random intercept and slope model allows for random slopes in addition to random intercepts. Then subjects vary not only in their baseline level of response, but in terms of the changes in their response over time. To cope with such individual slopes, the random intercept model 4.13 is extended to obtain

$$y_{ij} = \boldsymbol{x}'_{ij}\beta + b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d}{\sim} N(0,\sigma^2) \tag{4.15}$$

where b_{1i} is the individual-specific deviation for the slope.
For the individual-specific parameters, the bivariate normal random effects distribution is defined as

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{i.i.d}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix}\right) \tag{4.16}$$

The parameters τ_0^2 and τ_1^2 determine the variability of the individual-specific intercepts and slopes, respectively. The covariance $\tau_{01} = \tau_{10}$ can capture correlations between random intercepts and slopes. Such a correlation can occur, e.g. when individuals with larger slopes tend to have smaller intercepts, leading to negatively correlated random intercepts and slopes. The marginal variances of t_{ij} and the covariance between t_{ij} and t_{il} can be shown to be

$$Var(y_{ij}) = \tau_0^2 + 2\tau_{01}t_{ij} + \tau_1^2 t_{ij}^2 + \sigma^2 \quad \text{and}$$
$$Cov(y_{ij},y_{il}) = \tau_0^2 + \tau_{01}t_{ij} + \tau_{01}t_{il} + \tau_1^2 t_{ij}t_{il}, \quad j \neq l$$

resulting in an intraclass correlation coefficient, which depends in a rather complicated way on the observed covariate values and is not easy to interpret.

$$Corr(y_{ij},y_{il}) = \frac{Cov(y_{ij},y_{il})}{\sqrt{Var(y_{ij})}\sqrt{Var(y_{il})}}$$

Conditional and Marginal Formulation

The model 4.11 implies the *conditional* perspective

$$y_i|b_i \sim N(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \sigma^2\boldsymbol{I})$$

for the response vector y_i, given the random effect b_i. Here the individual- or cluster-specific effects b_i are interpreted similarly as the usual regression effects, with the difference that they only apply to individual or cluster i.
The *marginal* perspective is given by

$$\boldsymbol{y_i} \sim N(\boldsymbol{X_i\beta}, \boldsymbol{V_i}), \quad \text{where} \quad \boldsymbol{V_i} = Cov(\boldsymbol{y_i}) = \boldsymbol{Z_i D Z_i'} + \boldsymbol{\Sigma_i}. \tag{4.17}$$

Here, the random effects induce a correlation structure and therefore allow a valid statistical analysis of correlated data. In the marginal formulation of the LMM, the marginal, population-averaged expected value of y_i is modeled as a function of population effects.

4.3.3 Likelihood Estimation

LMMs can be estimated by maximum likelihood. However, this method tends to underestimate the variance components. A modified version of maximum likelihood, known as restricted (or residual) maximum likelihood (REML), is therefore often recommended. This ensures consistent estimates of the variance components. Details are given in Longford (1993) and Skrondal und Rabe-Hesketh (2004).
The vector of unkown parameters $\boldsymbol{\theta}$ in $\boldsymbol{\Sigma} = \boldsymbol{\Sigma(\theta)}, \boldsymbol{Z} = \boldsymbol{Z(\theta)}$ and $\boldsymbol{V} = \boldsymbol{V(\theta)}$ can be estimated via a maximum likelihood (ML) and a restricted maximum likelihood (REML) approach. These estimators for unknown variance and covariance parameters are the most commonly used.

Maximum Likelihood Estimation of θ

Maximum likelihood estimation of the unknown parameter θ is based on the likelihood of the marginal model

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V(\theta)})$$

The log-likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is, up to additive constants, given by

$$\log L(\boldsymbol{\beta},\boldsymbol{\theta}) = l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\{log|\boldsymbol{V}(\boldsymbol{\theta})| + (\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\}. \tag{4.18}$$

Maximizing $l(\boldsymbol{\beta},\boldsymbol{\theta})$ for fixed $\boldsymbol{\theta}$ with regard to $\boldsymbol{\beta}$, results in

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{y}$$

Inserting $\hat{\beta}(\theta)$ in $l(\beta,\theta)$ gives the *profile log-likelihood*

$$l_P(\boldsymbol{\theta}) = l(\hat{\boldsymbol{\beta}},\boldsymbol{\theta}) = -\frac{1}{2}\{\log(|\boldsymbol{V}(\boldsymbol{\theta}) + (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}.$$

which is only dependent on $\boldsymbol{\theta}$. The maximization of $l_P(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives the ML estimator $\hat{\boldsymbol{\theta}}_{ML}$.

Restricted Maximum Likelihood Estimation of θ

The restricted maximum likelihood (REML) approach. Rather than using $l_p(\theta)$, estimation of θ is often based on the *marginal or restricted log-likelihood*

$$l_R(\boldsymbol{\theta}) = log\left(\int L(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}\right),$$

It can be shown that the restricted log-likelihood is

$$l_R(\boldsymbol{\theta}) = l_p(\boldsymbol{\theta}) - \frac{1}{2}\{\log(\boldsymbol{X})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\},$$

and maximization of $l_R(\boldsymbol{\theta})$ provides the REML estimator $\hat{\boldsymbol{\theta}}_{REML}$.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}})^2 \tag{4.19}$$

The REML estimator $\hat{\boldsymbol{\theta}}_{REML}$ is preferred over the ML estimator in LMMs as an estimator for $\boldsymbol{\theta}$, as it reduces the bias of $\hat{\boldsymbol{\theta}}_{ML}$. Both $\hat{\boldsymbol{\theta}}_{REML}$ and $\hat{\boldsymbol{\theta}}_{ML}$ are computed numerically through iterative algorithms e.g. using Newton-Raphson or Fisher scoring algorithms. The estimated covariance matrices can be obtained by plugging in $\hat{\boldsymbol{\theta}}$ after convergence, leading to

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}),\ \hat{\boldsymbol{Z}} = \boldsymbol{Z}(\hat{\boldsymbol{\theta}}),\ \hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}}). \tag{4.20}$$

Estimation of Fixed and Random Effects

The estimators for the fixed and random effects $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$
$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{D}}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

which is equivalent to

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{pmatrix} = (\boldsymbol{C'\Sigma^{-1}C} + \boldsymbol{B})^{-1}\boldsymbol{C'\Sigma^{-1}y}.$$

by defining $\boldsymbol{C = (X,Z)}$ and $\hat{\boldsymbol{B}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{D}}^{-1} \end{pmatrix}$. $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}})$ can also be derived as the best linear unbiased predictor (BLUP) by omitting the normality assumption. For further information see McCulloch und Searle (2001).

Two linear mixed-effect models can be compared through a likelihood ratio test (LRT). This is only working if the models have been estimated by maximum likelihood or if the REML estimation models have the same set of fixed effects (Longford, 1993). The likelihood ratio statistic is given by

$$LRT = 2[l(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\theta}})]$$

where $l(\boldsymbol{\beta},\boldsymbol{\theta})$ is the log-likelihood from equation 4.18 of the marginal model. Here $l(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})$ are the unrestricted ML estimates and $l(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\theta}})$ the estimates under H_0. For $H_0 : \beta_j = 0$ for a component β_j of $\boldsymbol{\beta}$, this is a significance test of the jth covariate. Analogous the hypotheses on random effects variances can be testet.

4.3.4 Heterogeneous Variance

One assumption of a basic linear model is homogeneity of variance, meaning that the standard deviation of the error term is constant and does not depend on explanatory variables. Consequently, each probability distribution for the response variable has the same standard deviation regardless of the x-value. Especially in data sets containing repeated measurements for different groups, this assumption often fails. Therefore it is necessary to account for heterogeneity in the data. There are several methods for dealing with heterogeneity. The easiest solution is data transformation. However, in case that heterogeneity is an important extra information, that should not be thrown away, incorporating heterogeneity into the model is a better way to deal with it (Zuur, 2009). For a short introduction in different variance structures, see Zuur (2009).

4.3.5 Correlation Structure

The remaining autocorrelation, which is not explained by $\boldsymbol{Zb}$, as well as possible measuring errors are modelled by $\boldsymbol{\Sigma}$. Considering the definition of repeated measurements, it appears obvious, that closely spaced observations are more alike than measures lying far apart. Hence, there is a certain amount of correlation between the measurements of one subject, publishers respectively. If such correlation is ignored, inferences such as statistical tests or confidence intervals can be invalid. Thinking of the present data structure, it is very likely that payments for a particular publisher are correlated. Therefore it is necessary to include a correlation structure into the model. Before considering the correlation in the model, an appropriate correlation structure is needed. According to Fahrmeir et al. (2013) the most simple and commonly used approach is to include an autoregressive process of order 1 (AR(1)), meaning that only the previous value has a direct effect on the current value. In general, this process is represented by the following:

$$\epsilon_i = \rho\epsilon_{i-1} + u_i \tag{4.21}$$

where $-1 < \rho < 1$ is the AR(1) parameter and for u_i it is assumed:

1. $E(u_i) = 0$
2. $Var(u_i) = E(u_i^2) = \sigma_u^2,\ i = 1,...,n$
3. $Cov(u_i,u_j) = E(u_i,u_j) = 0,\ i \neq j$

4.3.6 Model Diagnostics

To check whether the model assumptions are met, a number of model examinations should be done. The main assumptions underlying the mixed-effects model are:

- Within group errors are independent and identically normally distributed, with zero mean and variance σ^2. Moreover they are independent of the random effects.
- The random effects are normally distributed, with mean zero and covariance matrix Σ and are independent for different subjects.

The main tools for checking the assumptions are based on the estimated residual errors or, simply, residuals and the estimated random effects.

4.3.7 Prediction of Linear Mixed Models

More than the results of the model parameters one is often interested in future prediction, based on the obtained model. To compare those predictions one often uses out of sample prediction, e.g. the model is not evaluated with the complete data set but with a training data set. The training data set must contain a certain amount of data points to ensure model validity. The obtained predictions for the data points, which are not in the training set, can then be compared with the real data. To evaluate the performance of the prediction, the mean squared error (MSE) is used. In Welham et al. (2004) the question is risen, whether random effects should be included in the prediction framework or not.

4.4 Additional Statistical Background

4.4.1 Multicategorical Factors

If a covariate has several categories, it is measured on a nominal scale level. Those variables are often called factor variables. With an intercept in the model and k faktor values, only $k-1$ values can be used, as there would be too many parameters otherwise. One can use dummy variables to model this. Then one dummy variable is omitted and the corresponding category is considered the reference category. The interpretation is the following: β_0 is the mean for the reference category k and $p_{(j)}$ is the increase or decrease of the mean response in comparison to the reference category k. Thus, for a factor variable $k-1$ functionally independent dummy variables are included.

4.4.2 Deviance

The deviance measures the discrepancy between the observations and the fitted model for models, which are estimated by maximum likelihood. It is based on the likelihood ratio statistic, which can be written as

$$\lambda = -2 \log \left(\frac{L(submodel)}{L(model)} \right).$$

Alternative test statistics are the Wald test and the score statistic.

4.4.3 AIC and BIC

To compare models and evaluate the best model out of a set of possible models, the *Akaike Information Criteria* (AIC) (Sakamoto et al., 1986) or the *Bayesian Information Criterion* (BIC, also called "Schwarz's Bayesian Criterion") (Schwarz, 1978) can be used.
The AIC is a measure of the relative quality of a statistical model for a given set of data and defined as:

$$AIC = -2\log L + 2d,$$

where L is the the maximized value of the likelihood function for the estimated model and d is the number of parameters. The term $2d$ penalizes complex models with a large number of parameters. An alternative is the BIC, which is defined as

$$BIC = -2\log L + \log(n)d.$$

It takes the parameters into account and puts a heavier penalty on more parameters than AIC. Therefore the BIC generally selects less complex models, as penalization of the number of parameters is stronger. Both AIC and BIC provide a means for model selection, where the preferred model is the one with the minimum AIC or BIC.

4.5 Analysis in R

For the fit of the logistic regression model the glm function is used. The distribution of the response is defined by the family argument, a binomial distribution in our case. The logistic function is the default link function, when the binomial family is requested. Most grafics were created with the *ggplot2* package by Wickham (2009).

Fitting the GAM models

For modeling generalized linear models the `mgcv` package was used (Wood, 2011). The mgcv implementation of gam represents the smooth functions using penalized regression splines. The smooth terms can be functions of any number of covariates and the smoothness of the functions can be controlled. Generally gam first constructs basis functions and one or more quadratic penalty coefficient matrices for each smooth term in the model, obtaining a model matrix for the strictly parametric part of the model formula. Then these are combined to obtain a complete model matrix and a set of penalty matrices for the smooth terms. In the analysis a "cr" bases is used, which is a penalized cubic regression spline. Cubic regressions splines, as used in the

model, are computational efficient for large data sets. To determine the choice of λ a generalized cross-validation (GCV)is applied.

Fitting the Bootstrap models

The nonparamentric bootstrap was conducted via the package `boot` (Canty und Ripley, 2014). For further information on the package see Davison und Hinkley (1997).

Fitting the LMM with `lme`

There are some packages for modeling linear mixed models in R, of which `nlme` and `lme4` are the most common. Those have several important differences between their functions. In the following, I will concentrate on the `nlme` (Pinheiro et al., 2012) package with its main model fitting function `lme`. Lme assumes that the data is grouped by the levels of some factor(s), and that the same random effects structure is needed for each group, with random effects independent between groups. The within-group errors are allowed to be correlated and/or have unequal variances. The model is fit by maximizing the restricted log-likelihood (REML). The parameter estimates from the REML analysis are in general preferable, because they avoid or reduce the biases of maximum likelihood, as seen above. estimates.lme() uses a mixed EM (expectation-maximization) algorithm and Newton-Raphson iterations for estimation. The package provides several commands for allowing heterogeneity. For examples of applications in R see Pinheiro und Bates (2000), Zuur (2009) and Everitt und Hothorn (2011).

Chapter 5

Analysis of the Data

5.1 Data and Approach

After providing an overview of the relevant statistical background, the presented models are now applied to the data. In comparison to the descriptive analysis part, where only the last data warehouse input of each publisher was taken into account, the data structure is now adjusted. All changes of explanatory variables of a publisher (e.g. change in businessmodel, status) are taken into account. Therefore these variables are now time-varying covariates in the model, as they may take different values for the time units. To adjust for several entries within one month per publisher only the last observation per month is taken. The aim is now to model the dependent variable total payment through suitable covariables. For the new data set only publishers were included which are in the traffic data set, so had at least once a payment. Moreover publishers with status blocked, Pinf.susp and bl.ref were removed from the data set, as they were only represented by 28 publishers altogether. A two step approach is used for modeling the data. First a logit model seeks to determine what influences total payment per month at all. In a second step, section 5.3 accounts for the asymmetry by excluding all payment in each month per publisher, which is zero.

Choice of Reference

In the following models, businessmodel *Topic* is chosen as reference category. This businessmodel incorporates different kinds of publishers and is thus very diversified. The reference category for the factor status is *prechecked*, as this is the status (mostly) every publisher passes in his development process. Moreover it is the starting status (after notchecked) for publishers in the network and includes the highest number of publishers. All statements made upon the status or the businessmodel are then made with respect to the reference. The interpretation of the coefficients depends on the reference categories. Changing it would yield other coefficients. The reference category for KAM is naturally KAM = 0, that is those publishers without key account manager.

5.2 Logit Model

As can be seen in figure 5.1, the boxplot and the histogram of the logarithmic total payment both show an asymetric distribution of the data, as most publishers have a very low logarithmic total payment. Indeed the median of the logarithmic payment is 0 and the mean is 0.6. This is calculated after adding 1 to the total payment per month to conclude a logarithmic transformation, which accounts for the zeros as $log(0) = -\text{Inf}$.

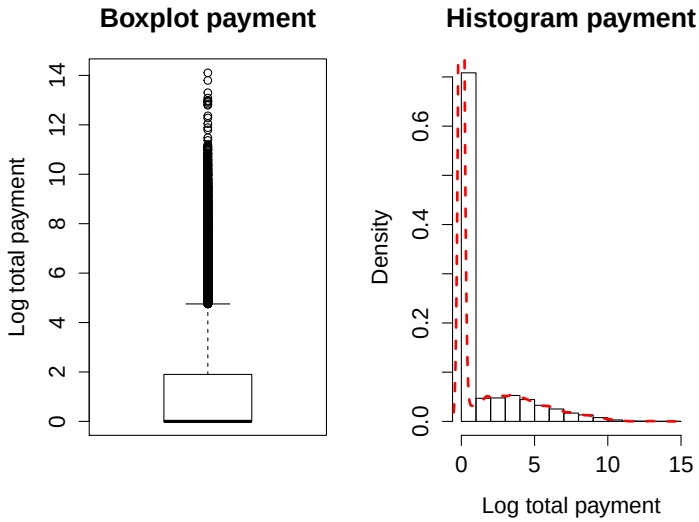


Figure 5.1: Boxplot and histogram of all publishers

First, the aim is to determine which variables have an influence on the payment. Therefore a categorical variable for the monthly payment is defined, with the categories being either "payment" ($y = 1$) or "no payment" ($y = 0$). This is checked for every publisher in each month. Payment_bin is the dummy variable with values

$$\text{Payment\_bin} = \begin{cases} 1 & \text{if monthly total payment for publisher i} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Moreover the dummy variable, which identifies if a publisher has a key account manager, is given by

$$\text{KAM}_i = \begin{cases} 1 & \text{if publisher i has a key account manager} \\ 0 & \text{if publisher i has no key account manager} \end{cases}$$

Note that some publishers might earn payment again after receiving no payment for a while. The initial logit model is:

$$\begin{aligned} \text{logit}(p_{it}) = log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 \text{Age}_{it} + \beta_2 \text{Month}_i + \\ \beta_3 \text{KAM}_{it} + \beta_4 \text{Status}_{it} + \beta_5 \text{BM}_{it} + \beta_6 \text{PS}_{it} \end{aligned} \tag{5.1}$$

where $p_{it} = \mathbb{E}(\text{payment\_bin}_{it}) = P(\text{payment\_bin}_{it} = 1)$ is the probability, that publisher i has payment in month t. Based on the linear predictor, the odds

$$\frac{\pi_{it}}{1-\pi_{it}} = \frac{P(p_{it}=1|x_i)}{P(p_{it}=0|x_i)} = \frac{P(\text{payment}_{it})}{P(\text{no payment}_{it})}$$

follow the multiplicative model

$$\begin{aligned}\frac{P(\text{payment}_{it})}{P(\text{no payment}_{it})} &= \exp(\beta_0)\cdot\exp(\beta_1\text{Age}_{it})\cdot\exp(\beta_2\text{Month}_i)\cdot\exp(\beta_3\text{KAM}_{it})\\ &\quad\cdot\exp(\beta_4\text{Status}_{it})\cdot\exp(\beta_5\text{BM}_{it})\cdot\exp(\beta_6\text{PS}_{it})\end{aligned}$$

Thus, the logit model can be interpreted as a linear model for log-odds, as well as a multiplicative model for the odds $\pi_i/(1-\pi_i)$. For the application of the GAM model the coefficients β_1 and β_2 are replaced by smoothing functions f_1 and f_2. An overview of all variables and their meaning is given in 5.1

Age_{it}	the age of the publisher measured in months (with reference date 2013-12-31). For the analysis the age was centered by substracting the mean of 54.6 years from every age.
Month_i	the month of the measurement by publisher.
KAM_{it}	the indicator for key account management, which equals 1 for key account managed publishers and 0 otherwise.
BM_{it}	shows the businessmodel of publisher i in month t.
PS_{it}	the number of existing partnerships for each publisher and month.

Table 5.1: Overview of explanatory variables.

5.2.1 GLM and GAM Results

To compare the bootstrap results with the original models of GLM and GAM, those were also computed. In the second model a smoothing spline is included using a generalized additive model. The smoothness is controlled by a parameter, which is specified through the degrees of freedom (df). Estimations with GAM have been made with the package `mgcv`. The degree of smoothness for month and age is estimated by generalized cross-validation (GCV). The method will find and fit the curve with a smoothing parameter that minimizes the GCV score. Here knot based cubic regression splines are applied. As the response variable is binary, a logit link function is used.

For the interpretation it is convenient to exponentiate the coefficients and interpret them as odds-ratios. The coeffcients then show the multiplicative effect on the chance of having payment in comparison to not having payment. Then it is possible to say, that for a one unit increase in a variable, the odds of having payment (versus not having payment) changes by a factor of $exp(\beta)$. Note that the odds ratio for the intercept is not generally interpreted. Stars represent the significance of the coefficients according to the p-value. The unit for the age is month, therefore small coefficients do not necessarily imply no impact. For the KAM it is the transition from 0 to 1 of the dummy variable, i.e. the transition from no KAM to KAM. The interpretation of the status is, that each status marks the multiplicative effect in the increase or decrease of the mean response in comparison to the reference category prechecked. The same applies to the

	GLM	GAM
(Intercept)	−1.28 (0.02)***	−1.47 (0.01)***
Age in Months	0.00 (0.00)***	
Month	−0.07 (0.01)***	
orders	0.43 (0.00)***	0.26 (0.00)***
$Month^2$	0.00 (0.00)***	
KAM1	0.23 (0.02)***	0.24 (0.02)***
Status notch.	−14.08 (234.43)	−134.40 (7034920.45)
Status susp.	−0.02 (0.03)	−0.03 (0.03)
Status ok	0.63 (0.02)***	0.61 (0.02)***
Status oktop	0.54 (0.09)***	0.52 (0.09)***
Status bl.bypre	−0.42 (0.27)	−0.48 (0.27)
Cashback	−0.23 (0.03)***	−0.23 (0.03)***
Coupon	−0.01 (0.03)	−0.02 (0.03)
Email	−0.23 (0.02)***	−0.23 (0.02)***
Media	−0.51 (0.03)***	−0.51 (0.03)***
Portal	−0.13 (0.01)***	−0.13 (0.01)***
Price Comparison	0.01 (0.02)	0.01 (0.02)
Unknown	−1.69 (0.01)***	−1.69 (0.01)***
Exist. PS	0.00 (0.00)***	0.00 (0.00)***
Month:orders	−0.02 (0.00)***	
EDF: Age in Months		7.95 (8.70)***
EDF: Month		4.24 (5.19)***
AIC	294268.20	296055.66
BIC	294490.40	296368.83
Log Likelihood	-147114.10	-147999.64
Deviance	294228.20	295999.28
Num. obs.	493680	493680
Deviance explained		0.38
Dispersion		1.00
R^2		0.44
GCV score		-0.40
Num. smooth terms		2

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.2: GLM and GAM Model presented in the log odds notation. Standard errors are given in brackets.

businessmodel with reference category Topic.

Table 5.2 shows the output for the multiple logistic regressions for both GLM and GAM. Additionally to the model of equation 5.1 the GLM model incorporates the squared influence of months and the interaction of month and orders, and the GAM model the orders as explanatory variables. The intercepts are the log odds of payment for no KAM publishers with status prechecked and BM Topic. Here not the multiplicative effects but the log chance is shown. According to the exponential function, log values above 0 indicate a multiplicative factor more than 1 and log values below 0 imply a smaller impact. The values are the coefficients associated with the variable listed to the left and the standard errors are in brackets. The coefficient is the estimated amount by which the log odds of the payment would increase, if the continuous coefficients were one unit higher (for example age). The intercept reveals the log odds for the payment in the reference categories, if the continous variables are zero. The effective degrees of freedom (EDF) for the GAM model are 7.95 for the age and 4.24 for the month. Both smoothing terms are highly significant. As the GAM model faces a penalty due to the need for more df, the residual deviance and the AIC is slightly lower for the GLM model. The models explain only a part of the variance in the data, as can be seen by the R^2. However, for such a complex model this seems reasonable.

Month and the squared month are highly significant. However, the influence is small and slightly negative for month. Both models show similar values for most variables. The multiplicative effect of $\exp(\hat{\beta}_{\text{KAM}})$ implies that the odds of a payment is $\exp(0.23) = 1.3$ times higher when a key account manager for a publisher is present, for fixed levels of the other factors. The status notchecked is not significant and shows huge values for the standard deviation. This can be explained by the low number of only 91 publishers being notchecked, which corresponds to 0.09% of publishers. The statuses ok and oktop increase the chance of a payment in comparison to prechecked. When a publisher has status ok, the chance increases by the factor of $exp(0.63) = 1.9$. All businessmodels have a negative coefficient, thus the multiplicative effect on the mean response is reduced, in comparison to businessmodel Topic. Businessmodels Coupon and PC are not significant. The effect of existing partnerships is significant in both models. However the effect of an increase in the partnerships of one, does not have a high effect on the outcome. Orders show a positive significant influence, thus an increase of one order increases the chance for payment by a factor of $exp(0.43) = 1.5$ in the GLM model.

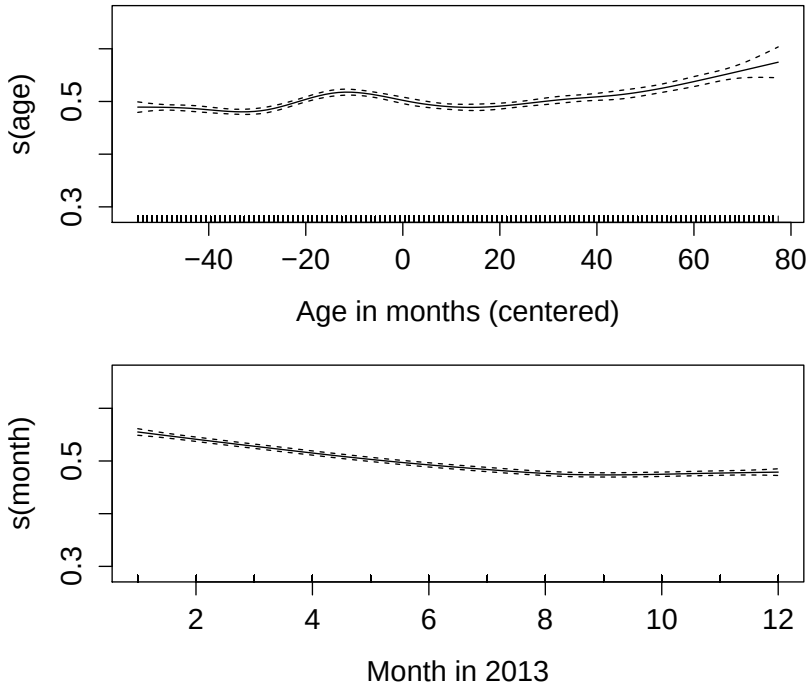


Figure 5.2: Partial contributions of explanatory variables for the GAM model. Solid curves are the function estimates, and dashed curves delimit the 95 percent confidence regions for each function. Smoothing parameter estimation was by GCV. The age was centered and thus shows negative values.

As the coefficients for month and age alone are hardly interpretable, when taking a look at the splines plot the influence over time becomes clear. The partial contributions of each covariate to the conditional probability of payment with bayesian confidence intervals are shown in Figure 5.2. The y-axis shows the predicted model on an inverse logit function scale, such that it returns the scale from 0 to 1. The peak at about 44 months shows a higher change for publishers in that age for payment compared to younger or older publishers. Thus a publisher being about ten years younger than the average publisher, has an increases chance of payment. For publishers

older than 40 month compared to the mean age, the chance is slightly increasing. However, confidence intervals are widening too. The changc for a payment is slightly decreasing for rising months in 2013, nevertheless the effect is small.

5.2.2 Bootstrap Results

As outlined in the theoretical part, due to the nature of the correlated data the validity of the standard error function for the coefficients in the GLM and GAM model is questionable. Therefore the bootstrap estimation is carrried out. For the modelation in R the `boot()` function, which is part of the boot package, is used to perform the bootstrap by repeatedly sampling observations from the dataset with replacement.

	original	bias	std.error	95%-CI Lower	95%-CI Upper
Intercept	-1.283	0.004	0.118	-1.53	-1.06
Age in Months	0.001	0.000	0.000	0.00	0.00
Month	-0.068	-0.004	0.010	-0.10	-0.05
Orders	0.428	0.017	0.186	0.16	0.87
$Month^2$	0.005	0.000	0.001	0.00	0.01
KAM1	0.230	-0.003	0.029	0.17	0.29
Status notch.	-49.083	26.794	15.253	-49.19	-12.04
Status susp.	-0.024	-0.003	0.030	-0.08	0.03
Status ok	0.629	-0.003	0.042	0.55	0.71
Status oktop	0.540	0.000	0.100	0.32	0.73
Status bl.bypre	-0.423	-0.019	0.271	-1.06	0.04
Cashback	-0.229	0.002	0.027	-0.29	-0.18
Coupon	-0.013	-0.004	0.025	-0.06	0.04
Email	-0.234	-0.001	0.021	-0.28	-0.19
Media	-0.509	-0.001	0.039	-0.58	-0.43
Portal	-0.131	0.001	0.013	-0.16	-0.10
Price Comparison	0.012	-0.004	0.017	-0.02	0.04
Unknown	-1.686	0.006	0.042	-1.76	-1.59
Exist. PS	0.001	-0.000	0.000	0.00	0.00
Month:orders	-0.023	0.001	0.021	-0.06	0.02

Table 5.3: Output of GLM with ordinary nonparametric bootstrap. For each statistic calculated in the bootstrap the original value and the bootstrap estimates of its bias and standard errors are printed. Moreover confidence intervals are provided.

Table 5.3 shows the results for the GLM bootstrap. The column "original" corresponds to the original estimates, the same as in the previous GLM model. The difference between the mean of the bootstrap estimates and the original estimates is what is called "bias" in the output. The "std. error" is the standard deviation of the bootstrapped estimates. As the original estimates imply, those are equal to the GLM output. However, the status notchecked is different to the previous model. This shows how unstable the estimation of this factor is. The difference between the bootstrap and the ordinary GLM model is by far the greatest for this factor. For the other variables the bias is considerably low. Except for the statuses notchecked and bl.bypre, the intercept and orders the standard errors are small.
Plotting a boot object draws a histogram and normal quantile-comparison plot of the bootstrap

replications for the coefficients. For the GLM model this can be seen in 5.3. The left plot shows a histogram of the bootstrap replicates. A vertical dotted line indicates the position of the coefficient. The second plot is a Q-Q plot of the bootstrap replicates. The order statistics of the replicates is plotted against the normal quantiles. The expected line is also plotted, which has the intercept mean(t) and slope sqrt(var(t)). The plot can be generated for every coefficient. Here, the KAM coefficient is taken, as an example. The mean value here is 0.23 with slightly negative bias. The quantiles of the standard normal seem reasonable. Significant deviations from the normal distribution are visible only in the extreme left and right tails. This can be confirmed for all other coefficients, except for month and orders, where more deviation can be observed. The plots for status notchecked, show large deviations. As seen in the output, the factor is unstable and should thus be not interpretated.

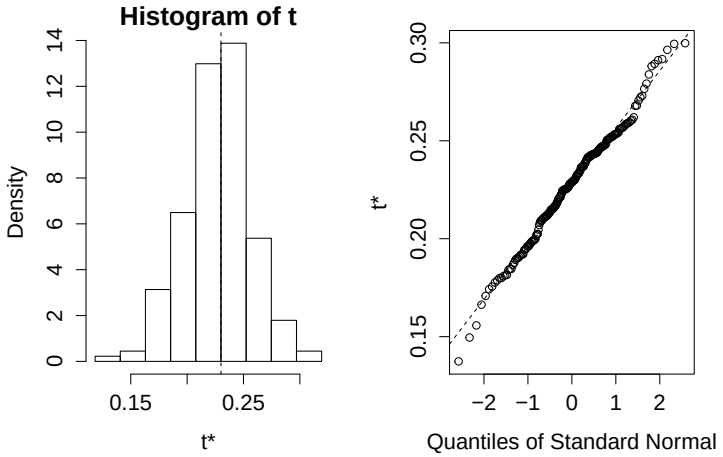


Figure 5.3: Histogram and normal quantile-comparison plot for the bootstrap replications of the KAM coefficient from the bootstrap fit with GLM. The broken vertical line in the histogram shows the location of the regression coefficient for the model to the original sample.

For the gam model the bootstrap results are listed in 5.4. Again, the value for the status notchecked provided in the original is different to the value here. Even on a the logarithmic scale the confidence intervals and the standard error are very large. For the remaining coefficients the bias is small and as in the GLM model the standard errors and confidence intervals are the largest for status bl.bypre, orders and the intercept.
Figure 5.4 pictures the histogram and normal quantile plot for the GAM models. Here the orders coefficient was seleced. The histogram is balanced over the value for the regression coefficient for existing PS for the model fit to the original sample, which lies at 0.001423. Again the plots for orders, show deviations from the standard normal quantiles.

Bootstrap Prediction Error

Bootstrapping tends to reduce the variance but gives more biased results to estimate the prediction error of a model. However, extended bootstrapping methods have been adapted to deal with the bootstrap bias, such as the 632 and 632+ rules, as seen in 4.2.
The .632 bootstrap prediction error can be calculated by boot.632 = 0.368·training error+0.632· bootstrap estimator according to equation 4.8. With 200 replications this results in 0.3680.9236+

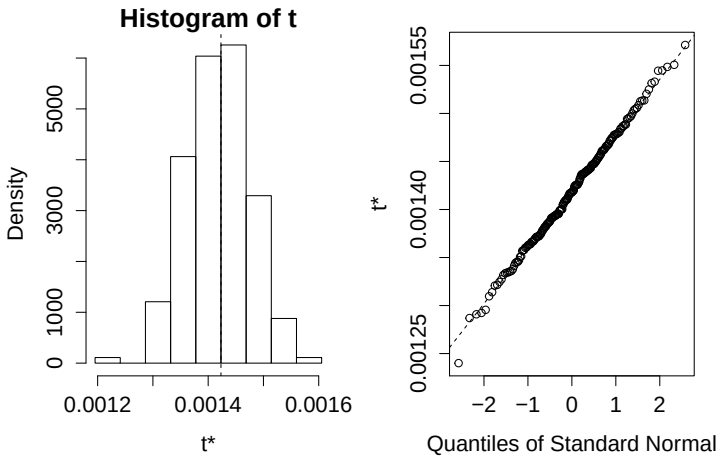


Figure 5.4: Histograms and normal quantile-comparison plots for the bootstrap replications of the existing PS coefficient in the GAM model. The broken vertical line in each histogram shows the location of the regression coefficient for the model fit to the original sample.

$0.632 \cdot 0.9222 = 0.9227$ for the GLM model and 0.9241 for the GAM model. Thus, the .632 bootstrap prediction error is slightly smaller for the GLM model. In this section, influencing coefficients on the height of the payment has been analysed. As expected key account managed publishers and publishers with status ok and oktop have shown an increasing multiplicative on the chance of payment compared to the reference groups. Publishers with status ok, had even a higher chance to oktop publishers. Moreover orders had a positive effect on the chance for payment. It could be shown, that the coefficient for notchecked is unstable and not reliable.

	original	bias	std.error	95%-CI Lower	95%-CI Upper
Intercept	-1.467	-0.008	0.064	-1.62	-1.37
KAM1	0.235	-0.002	0.028	0.18	0.29
Status notch.	-18.114	-14.064	43.916	-180.97	-16.95
Status susp.	-0.033	-0.002	0.031	-0.09	0.02
Status ok	0.609	-0.003	0.045	0.52	0.70
Status oktop	0.518	0.003	0.102	0.31	0.72
Status bl.bypre	-0.486	-0.012	0.274	-1.15	-0.00
Cashback	-0.234	0.001	0.028	-0.29	-0.18
Coupon	-0.016	-0.003	0.025	-0.07	0.03
Email	-0.230	-0.001	0.022	-0.28	-0.19
Media	-0.507	-0.001	0.039	-0.58	-0.43
Portal	-0.132	0.000	0.013	-0.15	-0.11
Price Comparison	0.012	-0.004	0.017	-0.03	0.04
Unknown	-1.691	0.004	0.044	-1.76	-1.60
Exist. PS	0.001	-0.000	0.000	0.00	0.00
Orders	0.259	0.015	0.078	0.16	0.47
Spline(Age):1	-0.038	-0.002	0.017	-0.07	-0.00
Spline(Age):2	-0.039	0.002	0.016	-0.06	-0.00
Spline(Age):3	0.100	-0.006	0.021	0.05	0.13
Spline(Age):4	-0.009	0.002	0.016	-0.04	0.02
Spline(Age):5	-0.020	-0.000	0.017	-0.05	0.02
Spline(Age):6	0.025	-0.001	0.017	-0.01	0.06
Spline(Age):7	0.072	0.002	0.023	0.03	0.12
Spline(Age):8	0.176	-0.001	0.029	0.12	0.24
Spline(Age):9	0.305	-0.000	0.072	0.18	0.45
Spline(Month):1	0.095	-0.001	0.011	0.07	0.12
Spline(Month):2	0.035	0.003	0.010	0.02	0.06
Spline(Month):3	-0.028	-0.002	0.010	-0.05	-0.01
Spline(Month):4	-0.083	-0.001	0.010	-0.10	-0.06
Spline(Month):5	-0.124	0.011	0.013	-0.14	-0.08
Spline(Month):6	-0.157	-0.015	0.020	-0.21	-0.14
Spline(Month):7	-0.158	0.002	0.009	-0.18	-0.14
Spline(Month):8	-0.150	0.002	0.011	-0.17	-0.12
Spline(Month):9	-0.133	-0.004	0.013	-0.16	-0.11

Table 5.4: Output of GAM with ordinary nonparametric bootstrap for 200 replications. For each statistic calculated in the bootstrap, the original value and the bootstrap estimates of its bias, standard error and confidence intervals are printed. The number of cubic splines was estimated by the model.

5.3 Analysis via LMM

Before turning to the specifics of the underlying data set, the LMM approach is described from a general perspective and transferred to the data. Subjects vary in the size of effects and this variability is treated as error or noise in standard analysis of variance models. But mostly this variability is also indicative of reliable individual differences in the effects. In experimental research, statistical analyses emphasize the significance of main effects and their interactions, the fixed effects. As seen in the theoretical part of the LMM's the random effects can be regarded as additional error terms, which account for correlation among observations within the same cluster or group. In this case one publisher is a group, while the several payments in 2013 are the observations. In our case the data is in the form of longitudinal data with repeated observations of the dependent variable y_{ij} for publisher i at time t_{ij}.
Having seen the influence of variables on the presence of payment, the question is how those variables work on the height of payment. Therefore the linear mixed-effects model is used to quantify the variability in total payment between publishers. The extent to which one particular publisher tends to increase or decrease the mean payment, i.e. the "effect" of that particular publisher on the total payment, is less interesting than the extent of the variability between publishers. Therefore the effects of the publishers are modeled as random effect parameters.
The dependent variable now is the payment per month, for those publishers with payment. Thus, the aim is to describe how the explanatory variables influence the *height* of the payment, not the payment at all, as in the GLM and GAM model. As many publishers have months with zero total payment, a commonly recommended transformation for count data with zeros is applied. This is the logarithmic transformation $log(y + 1)$, which leaves the data equal to 0 unchanged since $log(0+1) = 0$. The transformation moves statistical inferences into a multiplicative frame. The advantage of using a logarithmic transformation is that the parameter estimates obtained on a logarithmic scale, can be exponentiated and directly interpreted as multiplicative effects on the original scale. This transformation does not change the direction of effects and they rarely affect the significance of main effects. Nevertheless for interactions this does not hold always. The data is unbalanced in a way, that not every publisher has a total payment for exactly twelve months in 2013. However, this is no problem as LMM does produce sensible parameter estimation even if the data is not balanced (Pinheiro und Bates, 2000).
The boxplot of the logarithmic payment per month in figure 5.5 now includes only publishers with payment. Thus, publishers are reduced to those, where the binary variable for the payment is 1. In comparison to all publishers, the median rises from zero to a logarithmic payment 2.86 (which is equivalent to €17), while the mean rises to a logarithmic payment of 3.18 euro, equivalent 24 euro. Still, as can be seen on the left side at the histogram, there is a considerably high number of publishers, who earn less than 100 euro ($\sim \log(4.6)$) per year, while having months with zero logarithmic payment.

5.3.1 The LMM for the Publisher Data

In our case the publishers payment is tracked over time, therefore the measurements on an individual publisher are correlated. An adequate solution for this problem is a two components

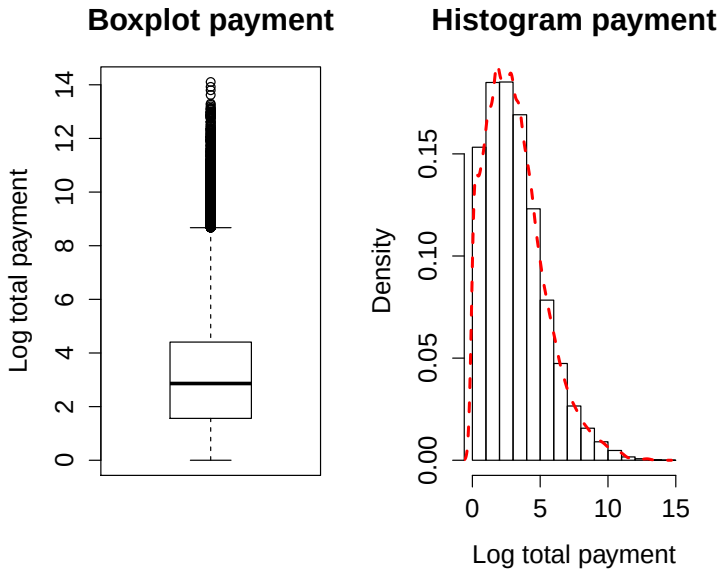


Figure 5.5: Boxplot and histogram for monthly payment with positive payment unequal to zero.

model with trend and random effects. Correlated errors are modelled through random effects, which account for the correlation in the data. Note that not all publishers have measurements for the entire year of 2013.

In the data set are several sources of variability. First between the publishers, that is the deviation from the mean of all publishers. Second there is variation within the publishers, that is deviation of a measurement, i.e. a publisher's total payment per month. The aim is therefore to estimate the publisher's specific effects, the effects of the population and the correlation structure. Random effects are used to model subject specific deviations from the population specific effects. Subjects and month are specified as random factors, varying in mean payment. The LMM assumes that month's mean payment as well as publisher's mean payments, are normally distributed around the respective fixed effects (i.e., the overall mean payment).

The LMM for clustered and longitudinal data as given in matrix notation in 4.11 is given by

$$y_{ij} = \boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i} + \epsilon_{ij}$$

for individuals $i = 1,...,m$ observed at occasions $t_{i1} < ... < t_{ij} < ... < t_{in_i}$

5.3.2 Random Intercept and Slope Model

The random intercept ans slope model allows for random slopes (over time) in addition to random intercepts. This is a more realistic structure of the covariances, where heterogeneity is ensured in both the slopes and the intercept. Suppose that the relationship between a publisher's payment and the month is different for each publisher. If there is any between publisher variation and a month-publisher interaction, this cannot be ignored. Otherwise this systematic variation ends up in the residuals, which leads to potentially biased inference. For more efficient estimation, a model with a random intercept *and* a random slope can be applied. This provides an individual slope and intercept for each publisher. The random intercept and slope model for the logarithmic

total payment of the ith publisher at the tth time point can then be written as:

$$\begin{aligned} \text{log payment}_{it} &= \beta_0 + \beta_1 \text{Age}_{it} + \beta_2 \text{Month}_i + \beta_3 \text{KAM}_{it} \\ &+ \beta_4 \text{Status}_{it} + \beta_5 \text{BM}_{it} + \beta_6 \text{PS}_{it} \\ &+ b_{0i} + b_{1i} \text{Month}_i + \epsilon_{it} \end{aligned} \tag{5.2}$$

The first two lines of 5.2 show the fixed-effects part of the model, where log payment$_{it}$ denotes the value of monthly logarithmic total payment for publisher i ($i = 1,...,m$), at time t ($t = 1,...,12$) i.e. each month in 2013. β_{0t} is the overall mean value of logarithmic payment, the overall intercept. The explanatory variables are analogous to 5.1. The third line shows the random effects. b_{0i} is the publisher-specific (random) deviation from the overall intercept β_0. $(\beta_0 + b_{0i})$ is the random intercept for publisher i. β_2 is here the "fixed" publishers slope of the effect of month, while b_{1i} describes the publisher specific deviation for the slope. Then $(\beta_2 \text{Month}_i)$ is the overall publishers effect for month and $\beta_2 \text{Month}_i + b_{1i} \text{Month}_i$ is the publisher-specific effect of month. ϵ_{it} is an independent error term distributed normally with mean 0 and σ^2.
To test whether the random intercept and slope model is superior to the random intercept model, a likelihood ratio test was conducted. The p-value associated with the LRT was significant (pvalue $< .0001$) and the AIC has dropped. Thus both citeria argue for preferring the model where both intercept and slope are random over the more simple random intercept model. Moreover the model with the interaction of month and orders was selected due to the LR test.

Random Intercept and Slope Model with AR1

The simple linear mixed model assumes no within-group correlations, so the repeated measurements are uncorrelated. As this seems unrealistic as publishers total payment tend to be correlated to previous months, a correlation structure was incorporated in the LMM. According to the LR test, a model which accounts for the correlation structure is superior. Several correlation structures have been compared using LR test and the ACF plot. The ACF is the empirical autocorrelation function of within-group residuals and can be helpful to examine the model. The ACF plot is provided in the appendix in B.0.1. Finally an AR(1) correlation structure for the residuals was selected.
Table 5.5 shows the model summaries for both the random intercept and slope model (RIaS) and the model with additional correlation structure. Most coeffiencents are significant on a 0.001 level, indicating that the p-value is below this significance level. Thus one would conclude that there is a relationship between the covariates and response. Then the null hypothesis that there is no relationship can be rejected. At the end of the table summary statistics about the fit (AIC, BIC and Log-likelihood) as well as the number of observations and groups (e.g. how many unique publishers) are listed. Note that all coefficients relate to the height of the *logarithmic* payment. The regression coefficients for the month in 2013 and the quadratic month are both highly significant. Both the month and the centered age of the publisher show a slightly negative effect. Thus the older the publisher gets and the closer we come to the end of 2013, the logarithmic total payment decreases. If a publisher is key account managed it

	RIaS Model	Model with AR
Intercept	2.54 (0.02)***	2.53 (0.02)***
Age in Months	−0.01 (0.00)***	−0.01 (0.00)***
Month	−0.05 (0.00)***	−0.05 (0.00)***
Orders	0.00 (0.00)***	0.00 (0.00)***
$Month^2$	0.00 (0.00)***	0.00 (0.00)***
KAM1	1.22 (0.04)***	1.25 (0.04)***
Status susp.	0.25 (0.04)***	0.28 (0.04)***
Status ok	0.90 (0.03)***	0.96 (0.03)***
Status oktop	2.58 (0.11)***	2.61 (0.11)***
Status blbypre.	0.27 (0.70)	0.28 (0.70)
Cashback	0.13 (0.06)*	0.14 (0.06)*
Coupon	0.17 (0.06)**	0.18 (0.06)**
Email	0.03 (0.04)	0.04 (0.04)
Media	0.20 (0.08)*	0.19 (0.08)*
Portal	−0.07 (0.03)*	−0.07 (0.03)*
Price Comparison	−0.01 (0.04)	−0.01 (0.04)
Unknown	−0.42 (0.02)***	−0.44 (0.02)***
exist.PS	0.00 (0.00)***	0.00 (0.00)***
Month:orders	0.00 (0.00)***	0.00 (0.00)***
AIC	286399.10	283948.81
BIC	286616.39	284175.54
Log Likelihood	-143176.55	-141950.40
Num. obs.	93677	93677
Num. groups	18469	18469

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.5: Random intercept and slope model without (left) and with (right) correlation structure. Values in brackets show the standard error.

increases the monthly payment in comparison to publishers who have no key account manager. In relation to the reference category prechecked for the publishers status suspicious, ok and oktop are significant, which all increase the monthly total logarithmic payment. While the latter two indications are obvious, it is to identify why the status suspicious in comparison to prechecked leads to higher payment. In comparison to the previous section, publishers with status notchecked are naturally not included, as for those the payment was zero. Publishers with businessmodel Cashback, Coupon and Media achieve higher monthly payments than the reference Topic. Portal and Unknown publishers receive smaller payments. Existing publishers are significant, nevertheless the impact of an increase in one partership is vanishingly small. Also the interaction between month and orders (incoperated by the idea that orders might increase or decrease in the course of the year) shows just a small negative impact.
The fits of the models can also be compared on a publisher level, which is shown in 5.6. More than the difference between both models, this plotpictures the underlying idea of the model. For all publishers, the coefficients of the fixed effects are the same. For the random effects - intercept and the month - the values vary for each publisher.

	Variance	Standard error	Correlation
(Intercept)	2.198871735	1.48285931	(Intr)
Month	0.008881181	0.09424002	-0.39
Residual	0.820188688	0.90564269	

Table 5.6: Variance and correlation components of the RIaS model with correlation structure.

Table 5.6 shows the variance components of the model with random intercept and slope and

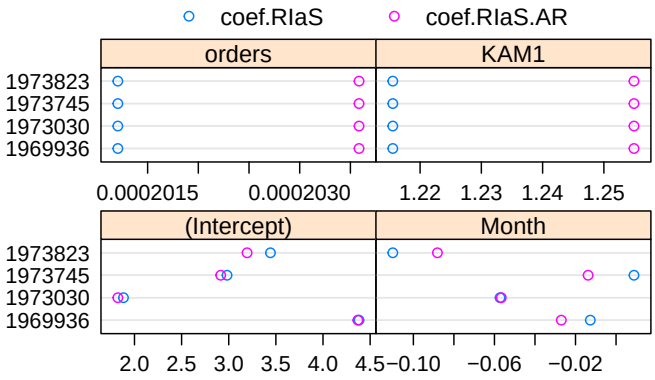


Figure 5.6: Comparison of selected coefficients between slope and correlation model for six publishers. While the fixed coefficients do not change per publisher, the intercept and the month coefficients vary over publishers. Values for the random intercept and slope model (RIaS) are in blue and coefficients for the model with correlation structure in pink.

the correlation structure for the random effects. A LMM estimates the variance-covariance parameters and the fixed effects simultaneously. The matrix gives the estimated variances and standard deviations for the random intercept and slope in the first two columns and rows. The third column gives the estimated correlation among random effects for the same publisher. The within-group error variance and standard deviation are included as the last row of the table. The random intercept variance is $\tau_0^2 = 2.199$, this marks the amount of variability of publisher-specific deviations from the overall intercept. $\tau_1^2 = 0.009$ is the amount of variability of publisher-specific deviations from the overall slope, which is very small. This could indicate that the simpler random intercept model may be sufficient. However, according to the LRT the random slopes were supported. The output indicates that there is a sizable negative correlation, -0.39, between intercept and slope, i.e. publishers with larger slopes tend to have smaller intercepts. The within-group variance is $\sigma^2 = 0.82$. Thus, estimated random effects variance of is quite large compared to the overall error variance. This indicates strong publisher-specific heterogeneity. Additionally it can be derived, how much the publisher effect accounts for the total variance. This can be calculated by taken the values of the variance into account. For this model, the publisher effect accounts for about 73% of the variance in the logarithmic total payment.

	Lower	Estimated	Upper
sd(Intercept)	1.455	1.483	1.51
sd(Month)	0.009	0.094	0.098
sd(Residual)	0.897	0.906	0.915
cor(Intercept,Month)	-0.418	-0.39	-0.362

Table 5.7: Intervals for the standard deviance of variance components and the correlation structure.

The intervals for the standard deviations of the variance and correlation components in 5.7 mark tight intervals for the estimated standard deviance and correlation.

5.3.3 Examination of Fitted Model

For the fitted mixed effects model it should be checked, whether the underlying distributional assumptions seem valid for the data. As seen in the theoretical part, two assumptions are important:

1. Within-group errors are independent and identically normally distributed, with mean zero and variance σ^2. Moreover they are independent of the random effects.

2. The random effects are normally distributed, with mean zero and covariance matrix $\mathbf{\Sigma}$ (not depending on the publisher) and are independent for different publishers.

Check Assumptions on the Within-Group Error

To check assumption 1 the within-group residuals, which are the difference between the observed response and the within-group fitted values, need to be examined. The within-group residuals are the estimated BLUPs of the within-group errors, as the random-effects variance-covariance matrix is replaced with their estimates.

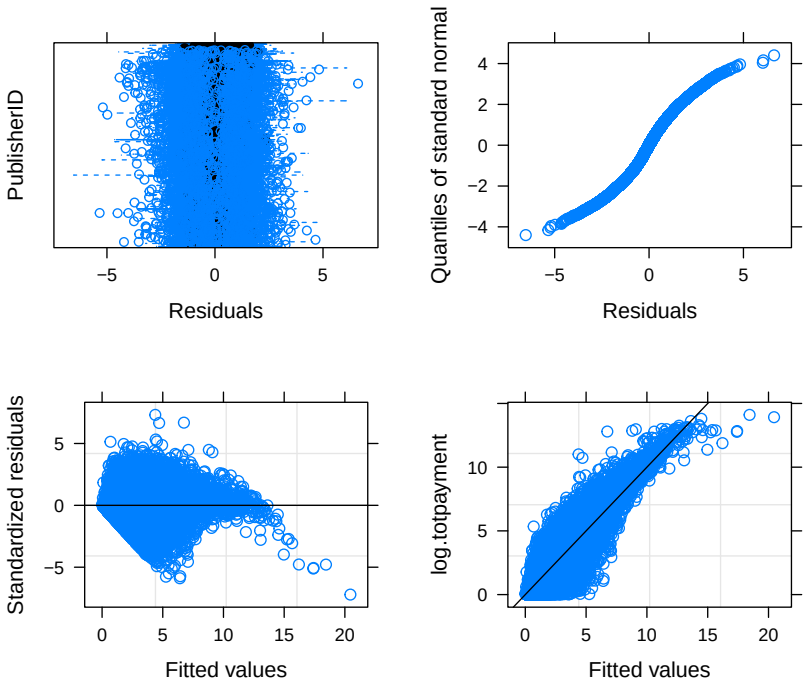


Figure 5.7: Model checking for for LMM with AR(1) structure. Clockwise from top left: Plot of residuals by publisher, normal probability plot, Scatterplot of standardized residuals versus fitted values, and observed versus predicted values.

The residuals by publisher are shown in figure 5.7. It shows that the errors are centered at zero ($E(\epsilon) = 0$), have more or less constant variance by group ($Var(\epsilon_{ij} = \sigma^2)$), and are independent from the group levels. As for several publishers only some observations are given, the individual residuals are less reliable. The plot of standardized residuals versus fitted values shows indication of within-group heteroscedasticity. Several outliers can be identified within the plot. Normal probability plot of the residuals provides clear evidence about departures from

the normality assumption, especially for very large residuals. Generally, minor violations of the normality assumption are not problematic since estimators are often relatively robust against such departures from normality.

Check Assumptions on the Random Effects

Now assumption 2 is examined. To check the model assumptions of normality for the random effect terms figure 5.8 is provided. For both plots some slight deviation from linearity can be observed. The assumption of normality seems acceptable for both random effects, although there is some asymmetry.

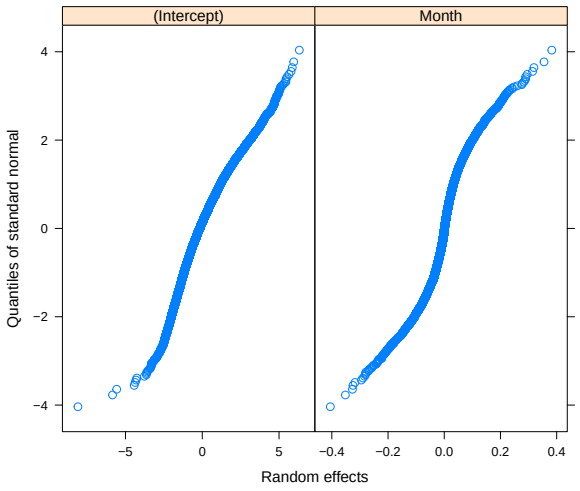


Figure 5.8: Normal plot of estimated random effects from lmm fit with heteroscedastic random intercept and slope model with correlation structure.

5.3.4 In Sample Prediction

When plotting the in sample predicted values with the raw data as in figure 5.9 it can be seen that for each publisher a separate intercept and slope is estimated. All the panels have the same vertical and horizontal scales, which allows to evaluate the pattern over time for the plotted publishers and also to compare patterns between subjects. The plot shows the difference of publishers in both the slope (the typical change in logartihmic total payment per month for those particular publishers) and the intercept (the average logarithmic payment for the publisher). The plot shows 49 publishers, selected as example. Moreover it also reveals that not all publishers contain data for the entire year 2013.
Now, the model estimates are examined for the age and month effects. Figure 5.10 shows the mean of logarithmic payments and fitted values by months and age in 2013. For each month the mean over all publishers was calculated for the observed data, as well as for the in sample predictions from the random intercept and slope model with correlation structure. For both the age and the months the red lines provide a good prediction of the real values. The mean for the logarithmic payment is very similar throughout the year. The mean payment by age is slightly decreasing with rising age.

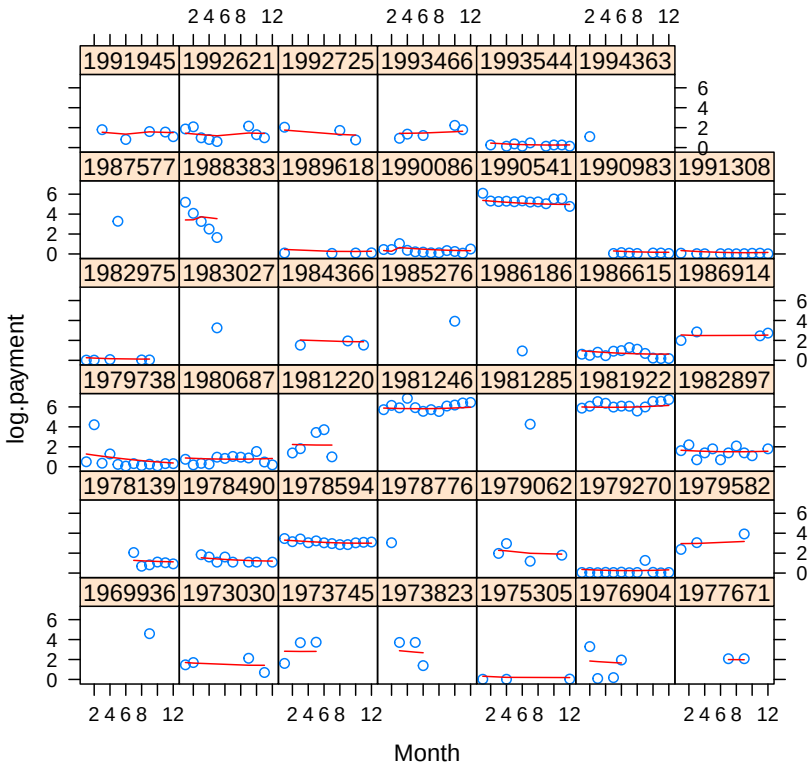


Figure 5.9: Plot of the predicted values of the random intercept model and slope model with correlation structure. The blue dots represent the raw data and while the red line are the predicted values. Each publisher's data are shown in a separate panel, along with the regression line of the predicted values fit to the data in that panel. The publisher number is given in the strip above the panel.

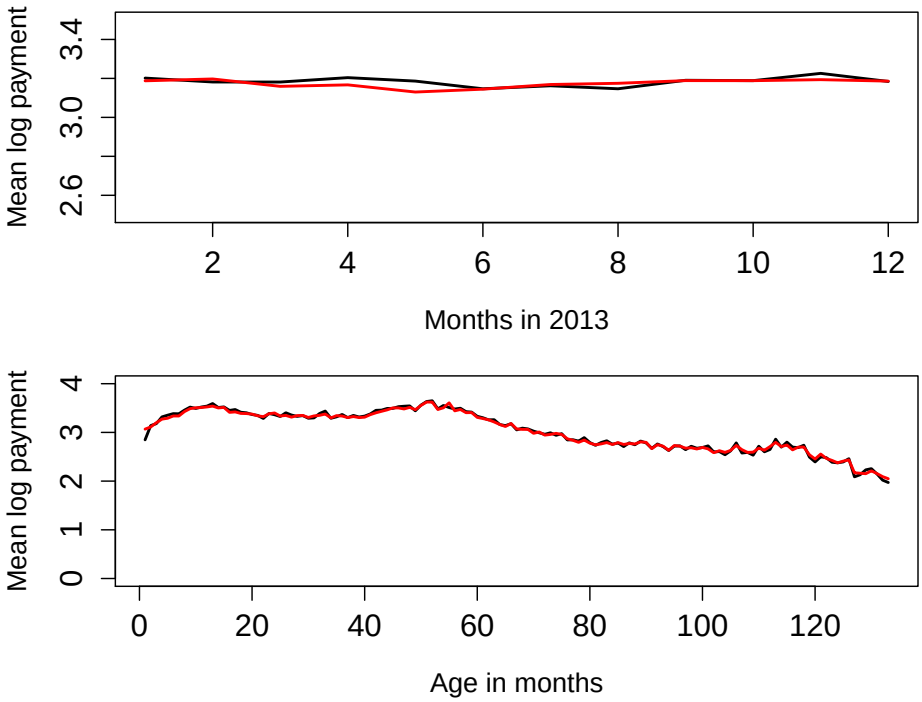


Figure 5.10: Mean of logarithmic payments and fitted values (red line) by months in 2013 (above) and age in months (below). Axes are different due to better readability.

5.3.5 Data Subsets

The previous model with a random slope and intercept with an correlation structure is now executed for different subsets of the data. First, the payment can be divided into the advertiser groups the payment was coming from. The advertisers are separated into three groups according to their payment force. Group A includes the strong advertisers, followed by group B and C. Advertisers from group A are usually top brands with a wide range of ad formats that generate many clicks and high conversion rates. The payment from group B and C was combined, as only a small share of payments came from advertisergroup C. For modeling those, the same coefficients than in model 5.2 are included, while the dependent variable is totalpaymentA$_{it}$ for advertisergroup A and totalpaymentBC$_{it}$ for paymentgroup B and C. Note that publishers can have payments from both groups or from only one of the groups. 88% of publishers received payments from advertisergroup A and 85% received payment from advertisergroups B and C.

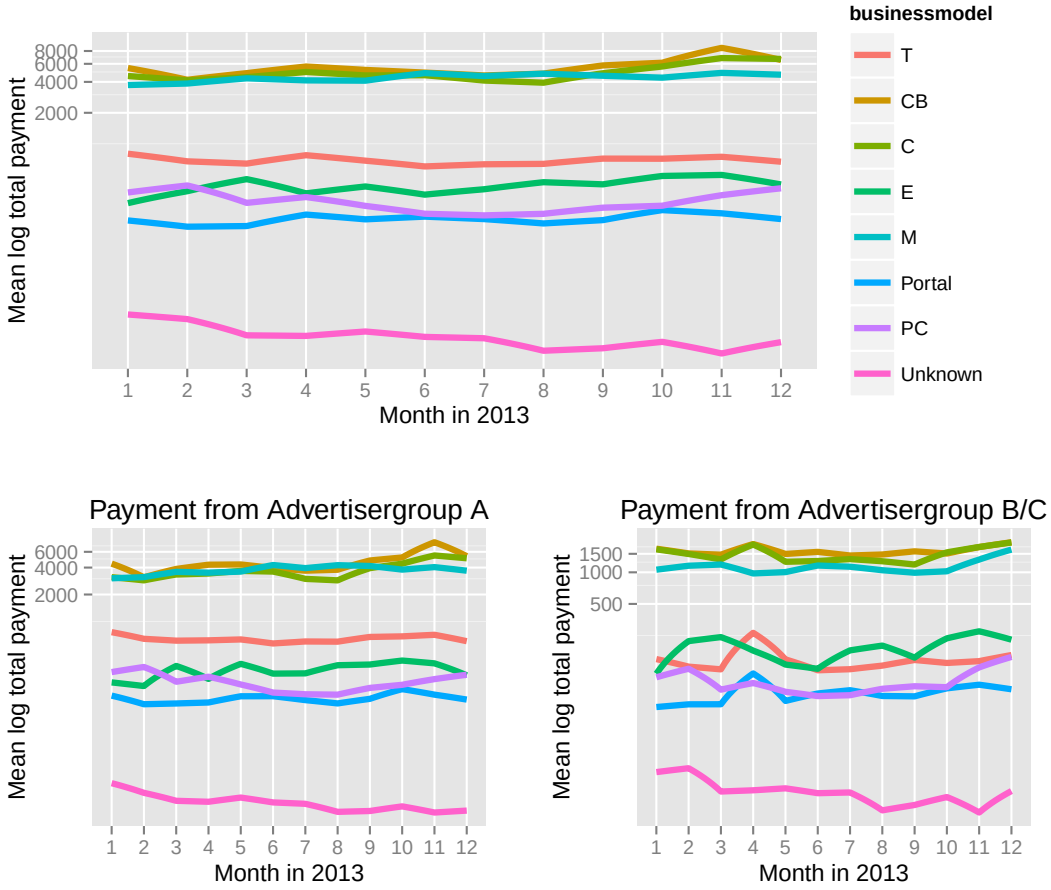


Figure 5.11: Mean of logarithmic total payment over publishers by businessmodel and month in 2013. The upper chart shows the total payment, while the lower charts shows the payment derived from advertisersgroup A (left) and advertisergroup B and C (right). Note that the transforming occured after the statistics have been computed and the axes are untransformed.

The overall mean logarithmic payment along with the payments separated by businessmodel are given in 5.11. The plots clearly structure the paymentgroups by their height of payment. Leading are businessmodels Coupon, Cashback and Media. While their mean payment is about 6000 euro for paymentgroup A, this reduces to about 1,500 euro for paymentgroups B and C. The

next group of publishers with a mean overall payment below €2000 are Topic, Email, PC and Portal. Lastly Unknown is far below this group. Both businessmodels Cashback and Coupon increase their logarithmic mean payment from August 2013 on. This is especially visable in paymentgroup A. As the lines are on a logarthmic scale, this translates to a higher effect for the mean payments. Also PC increases the mean payment from that point on, after the payment was declining in previous months. Businessmodel Unknown has the most decreasing effect over the year. This could be also a reason, why the overall mean payment throughout the year is slightly decreasing as seen for example in 5.10, as BM Unknown includes many publishers.

	Advertisers A	Advertisers BC
Intercept	1.75 (0.02)***	1.43 (0.02)***
Age in Months	−0.01 (0.00)***	−0.01 (0.00)***
Month	−0.03 (0.01)***	−0.07 (0.01)***
Orders	0.00 (0.00)***	0.00 (0.00)***
$Month^2$	0.00 (0.00)***	0.00 (0.00)***
KAM1	1.44 (0.05)***	0.81 (0.05)***
Status susp.	0.36 (0.04)***	0.19 (0.05)***
Status ok	0.83 (0.03)***	0.80 (0.03)***
Status oktop	2.85 (0.13)***	2.06 (0.13)***
Status bl.bypre	−0.47 (0.79)	−0.08 (0.72)
Exist. PS	0.00 (0.00)***	0.00 (0.00)***
Month:orders	0.00 (0.00)***	0.00 (0.00)***
AIC	274425.94	261415.43
BIC	274584.45	261573.28
Log Likelihood	-137195.97	-130690.72
Num. obs.	82780	79654
Num. groups	16048	15724

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.8: Coefficients from random intercept and slope model. The logarithmic total payment is separated by the advertiser group.

	Variance	Standard error	Correlation
(Intercept)	2.89514	1.7015	(Intr)
Month	0.01048	0.1024	-0.412
Residual	1.08275	1.0406	

Table 5.9: Variance and correlation components of model with advertisergroup A

	Variance	Standard error	Correlation
(Intercept)	2.3381	1.5291	(Intr)
Month	0.0109	0.1044	-0.432
Residual	1.0684	1.0336	

Table 5.10: Variance and correlation components of model with advertisergroup B and C.

The coefficients for the models by paymentgroup are stated in 5.8. The intercept for paymentgroup A is higher, this relates to the previous plots. Also the effect of KAM is far higher, thus the payment that comes from advertisers A is much more related to publishers with key account manager. The significance of the coefficients is equal in both models. As the models with the payment from the different advertisergroups includes mostly the same publishers, the variance components are very similar. Nevertheless, the variance in the intercept is higher for the model

for paymentgroup A, whereas the residual deviance is only slightly higher.
Additionally to modeling the payment according to the advertiser groups, the data can be subsetted through the different levels of the factor variables KAM and status of the publisher. Here, the dependent variable is again the total payment for all advertisergroups combined.

KAM

Starting with KAM, the data is subsetted into key account managed publishers and those without. The corresponding table of model estimates is provided in the appendix at B.1. For no KAM publishers businessmodel Cashback increases the overall monthly payment in comparison to Topic, while Portal and Unknown are decreasing it. In the KAM publishers, Media can significantly rises the mean payment, while Unknown reduces it. Other BM coefficients are not significant.

	Variance	Standard error	Correlation
(Intercept)	1.915537	1.38403	(Intr)
Month	0.007297	0.08542	-0.392
Residual	0.813505	0.90195	

Table 5.11: Variance and correlation components of model for no KAM publishers

	Variance	Standard error	Correlation
(Intercept)	5.00788	2.2378	(Intr)
Month	0.01375	0.1173	-0.233
Residual	0.73148	0.8553	

Table 5.12: Variance and correlation components of model for KAM publishers. The subset relates to 9123 publishers.

The variance between publishers in the KAM model is far higher than for publishers without KAM. This is indicated in the variance components tables given in 5.11 for no KAM and 5.12 for KAM. Additionally the standard deviation for KAM publishers is higher. The correlation between random intercept and slope is smaller for KAM publishers. This is reasonable as the standard deviation increases. The interpretation is, that publishers with larger slopes tend to have smaller intercepts. For the KAM publishers this effect is therefore not as pronounced as for no KAM models. While for the model with no key account managed publishers, the publisher effect accounts for 70% of the variance in the payment, in the model for the KAM publishers only, it accounts for 87%.

	Variance	Standard error	Correlation
(Intercept)	1.780181	1.33423	(Intr)
Month	0.007121	0.08438	-0.386
Residual	0.822636	0.90699	

Table 5.13: Variance and correlation components of model for publishers with status prechecked.

	Variance	Standard error	Correlation
(Intercept)	3.14343	1.7730	(Intr)
Month	0.01573	0.1254	-0.419
Residual	1.16461	1.0792	

Table 5.14: Variance and correlation components of model for publishers with status suspicious.

	Variance	Standard error	Correlation
(Intercept)	3.262591	1.80626	(Intr)
Month	0.008821	0.09392	-0.3
Residual	0.681366	0.82545	

Table 5.15: Variance and correlation components of model for publishers with status ok.

Status

Comparing the publisher effect for the different models by statuses, yields the following results. For the prechecked model the variation between publishers accounts for 68%, in the ok model for 83%, in the oktop model for 91% and in the suspicious model for 73%. The respective model estimates are given in the appendix at B.2. While for all statuses (and also in the overall model) the coefficient for the age in months is slightly negative, for status oktop it is positive. Thus, in that status, if the publishers age increases the payment rises. However, this is just relating to 158 publishers, as the group label shows.
Thus, in the KAM and oktop publisher models, the highest percentage of variance in the payment can be explained through the publisher effect. The variance between publishers is very high for those publishers, as their size of payment varies much. The doubling of revenue from €1000 to €2000 is clearly different to an increase from €1,000,000 to €2,000,000. However, the overall residual variance for those models is smaller, increasing the percentage of variability the publishers account for.

5.4 Separate Models per Businessmodel

As the businessmodels have different characteristics, as seen in the course of the work, it is also interesting to look at them separately. The models for businessmodels Cashback, Coupon, Email and Media are presented in 5.17. With the number of groups represented in the model output, the size of each businessmodel is recognisable. The highest mean logarithmic payment, with all other variables being fixed, is obtained in businessmodel Coupon. Only for BM Email an increase in age is significantly leading to a lower logarithmic payment. Additionally, it is the only businessmodel (with Unknown) for which month is not significant. Publishers in BM Cashback are significantly decreasing their logarithmic payment, if the have status suspicious, in comparison to status prechecked. For the other three businessmodels status suspicious has no significant effect. The influence of oktop is higher than the effect for KAM, for all models. As the number of KAM publishers is higher than oktop publishers, thus oktop publishers probably are already higher earning publishers, this seems reasonable.
The remaining models are shown in 5.18. Here status P.blbypre is additionally included in

	Variance	Standard error	Correlation
(Intercept)	5.53345	2.3523	(Intr)
Month	0.01475	0.1214	0.137
Residual	0.51927	0.7206	

Table 5.16: Variance and correlation components of model for publishers with status oktop.

	Cashback	Coupon	Email	Media
(Intercept)	2.64 (0.13)***	3.13 (0.14)***	2.74 (0.09)***	2.98 (0.18)***
Age in Months	0.00 (0.00)	0.00 (0.00)	−0.01 (0.00)***	−0.01 (0.00)
Month	−0.11 (0.03)***	−0.13 (0.02)***	−0.05 (0.03)	−0.09 (0.04)*
Orders	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***
$Month^2$	0.01 (0.00)***	0.01 (0.00)***	0.00 (0.00)	0.00 (0.00)
KAM1	1.21 (0.20)***	1.24 (0.15)***	0.75 (0.23)**	1.52 (0.25)***
Status ok	0.84 (0.14)***	0.66 (0.11)***	0.65 (0.14)***	0.59 (0.21)**
Status oktop	2.92 (0.52)***	2.18 (0.34)***	1.96 (0.86)*	2.02 (0.38)***
Status susp.	0.74 (0.16)***	0.22 (0.16)	0.29 (0.16)	−0.08 (0.27)
Exist. PS	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)*	0.00 (0.00)***
Month:orders	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)	0.00 (0.00)*
AIC	9052.27	15413.69	11828.26	6248.93
BIC	9148.13	15517.21	11926.71	6337.65
Log Likelihood	-4510.14	-7690.84	-5898.13	-3108.47
Num. obs.	2966	4782	3486	1902
Num. groups	550	777	1006	364

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.17: Random intercept and slope model for selected businessmodels I.

	Portal	Price Comparison	Topic	Unknown
Intercept	2.51 (0.05)***	2.78 (0.08)***	2.62 (0.03)***	1.83 (0.03)***
Age in Months	−0.01 (0.00)***	0.00 (0.00)**	−0.01 (0.00)***	−0.01 (0.00)***
Month	−0.06 (0.01)***	−0.04 (0.02)*	−0.06 (0.01)***	−0.02 (0.01)
Orders	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)*
$Month^2$	0.00 (0.00)***	0.00 (0.00)**	0.00 (0.00)***	0.00 (0.00)
KAM1	1.39 (0.16)***	1.24 (0.18)***	1.14 (0.05)***	0.54 (0.15)***
Status ok	0.78 (0.07)***	0.63 (0.13)***	1.05 (0.04)***	0.47 (0.07)***
Status oktop	2.50 (1.54)	1.12 (0.54)*	2.42 (0.16)***	1.37 (0.73)
Status susp.	0.13 (0.09)	0.51 (0.16)**	0.24 (0.05)***	0.04 (0.15)
Exist. PS	0.00 (0.00)***	0.00 (0.00)**	0.00 (0.00)***	0.00 (0.00)*
Month:orders	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)	0.00 (0.00)***
Status bl.bypre		−1.16 (1.49)	1.21 (1.10)	0.74 (0.70)
AIC	45144.44	19581.10	146268.07	27873.56
BIC	45266.18	19695.38	146417.58	27997.25
Log Likelihood	-22556.22	-9773.55	-73117.03	-13919.78
Num. obs.	14909	6149	48789	10694
Num. groups	3025	1230	9264	3533

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.18: Random intercept and slope model for selected businessmodels II.

businessmodels PC, Topic and Unknown, however it shows no significant influence on the payment. For BM Portal and Unknown the status oktop is not significantly rising the payment in comparison to the status prechecked. The highest influence on the height of the payment for the oktop publishers are given in businessmodel Cashback, Coupon and Topic. Those are the publishers who include the highest number of publishers from paymentgroup seven, as seen in the descriptive analysis. In model PC, the influences of KAM and oktop were about the same level.

The percentage to which the publishers effect can account for the total variability in the different models, is the following: Cashback 76.2%, Coupon 78%, Email 61.7%, Media 73.4%, Portal 70.5%, PC 62.8%, Topic 74.2% and Unknown 78.5%. Thus, the effect is the highest for businessmodel Portal and Unknown, while the publisher effect contribute considerably less in BM Email and PC. The individual variance and correlation components are attached in the appendix, starting with B.3. The correlation between the slope and the intercept is on a low negative level for BM Media with -0.154 and exceeds the highest value for -0.552 for Coupon publishers. This is indicating that publishers in businessmodel Coupon with higher slopes, tend to have smaller intercepts than Media publishers.

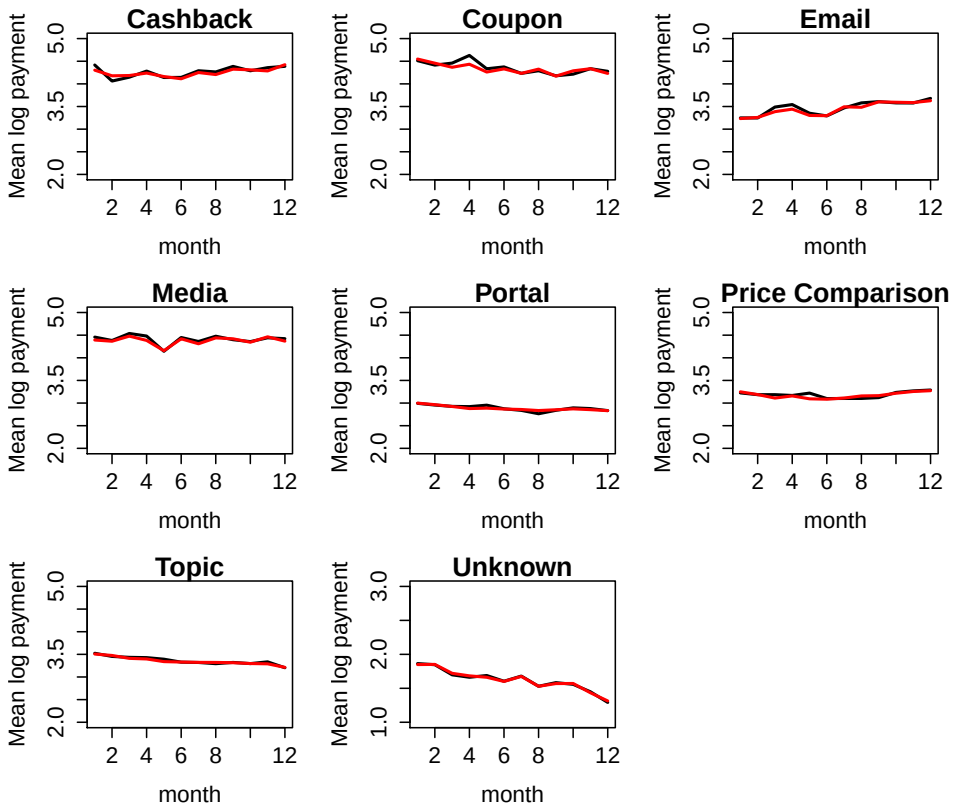


Figure 5.12: Mean of monthly logarithmic payments and mean of monthly fitted values for each businessmodel. Note that the axis labels change for businessmodel Unknown.

Figure 5.12 shows for all publishers the mean of logarithmic payments and fitted values by months and age in 2013. For each month the mean over all publishers of each businessmodel was calculated for the observed data as well as for the in sample predictions, i.e., the model estimates. For all busiessmodels the mean logarithmic payment seems stable throughout the year. However, a slight increase for Cashback and Email, as well as a decrease for Unknown and Topic can be observed. The estimated model values represent the observed values closely.
The development of the mean logarithmic payment by the age of the publishers for all businessmodels is given in figure 5.13 . While the mean seems stable for Portal and Topic, and for Unknown shows only a sharp peek for younger publishers, a different picture is given for other businessmodels. Here, the mean payment per age in months is highly oscillating. This might be due to the reason, that those businessmodels include fewer publishers, so the mean values are

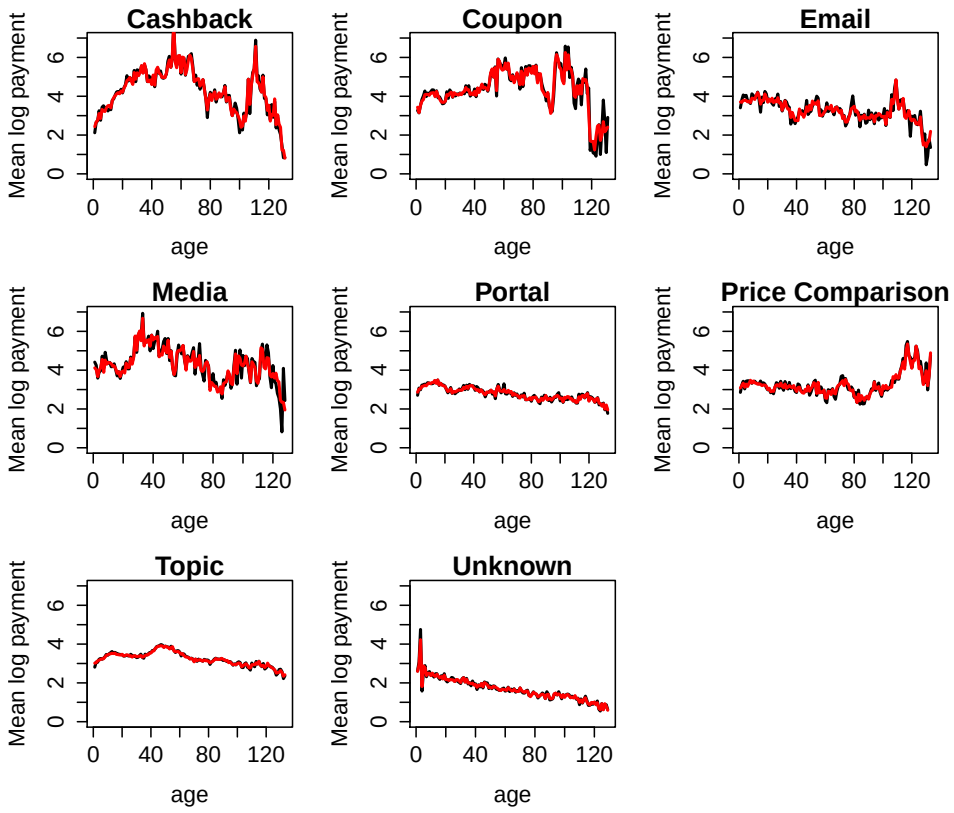


Figure 5.13: Mean of log payments by age in months and mean of fitted values by age in months for each businessmodel.

not that stable. However, conclusions can be drawn with caution. Publishers in Cashback with the highest logarithmic payment are about 60 and 110 months old. Media publishers reach their peak at about 40 months. Publishers with businessmodel Price Comparison show the highest mean payment at age about 115 month, which is thus the businessmodel where "old" publishers exceed the highest mean payment.

This section gave an insight into influencing variables for the height of the logarithmic payment during the course of 2013. All conclusions were drawn on the logarithmic payment, thus the effects on the untransformed payment are even higher. As expected, publishers with KAM and status ok, showed the highest influence on the height of the payment. Several subsets of the data were taken into account, the the payment was seperated due to the advertisergroup.

Chapter 6

Conclusion and Outlook

In the future, all kinds of companies will be confronted with ever growing amounts of data. This has mainly two reasons: First, more and more data is generated by the increasing digitalization of everyday life, such as through social media and sensor technology. Second, the data may be secured economically due to the exponentially decreasing cost of data storage. Companies need to adjust for these changes, as data analysis can secure a competitive edge. Analysing the customer has thus become a new area of gaining information on the behaviour of clients. Especially in the field of marketing, this approach experiences increasing popularity. However, analysing huge data sets can be a challenge, without having the appropriate tools and methods at hand. For this work, through a cooperation between the university and the economy it was made possible, to analyse real life data in the context of online marketing.

Affiliate marketing is a quite recent but already established way of online marketing. The publishers of a network have been analysed in detail. Both their characteristics over time and their influencing variables on the payment have been considered. As we have seen the network is highly dependent on its big publishers, as they produce the biggest part of the overall payment and therefore provision of the firm. With increasing tracking technology of big companies operating online, the chance to separate them from the affiliate partner networks might increase.

After a descriptive analysis was carried out to capture the data, the analysis was devided in two parts. First, a Logit model seeked to determine what influenced the overall chance for payment in 2013. Significant variables here, that increased the chance for payment in comparison to the reference categories, were orders, KAM, ok and oktop. The bootstrap bias and standard deviations were mostly small, thus confirming the previous models. However, the status notchecked was identified as unstable. While the chance of the months was decreasing for 2013, the spline curve indicated a peek of the chance for publishers at about 40 years.

Second, the LMM captured the influencing variables for the height of the payment. Here, as in the Logit model, the variables KAM and oktop had a significant influence on the payment. Mostly the influence of oktop publishers was greater than for KAM publishers. However, deviations from that pattern could be observed when taking subsets of the data. The influence of month was mostly slightly negative throughout the year 2013, confirming the spline curve of the GAM model. Here, the in sample prediction for the age revealed two small peaks at 15 and 40 years. After that age, the mean logarithmic payment decreased. Moreover, order and the

interaction of orders as well as the existing partnerships are affecting the payment. Nevertheless, a increase in one order or one partnerships, only yields small improvements on the logarithmic payment.

To summarize, both models yielded interesting insights into the affecting variables for payment. The main expectations, derived from the descriptive analysis could be confirmed. However focusing on different subgroups incooperated a different view.

As many publishers register with the network but never get active the question arises what causes this inactivity. Therefore the network experiments at the moment with a streamlined process of customer registration and uncomplicated collation of customer details. This is conducted for the UK and Swiss market. Further analysis could reveal if this lean process leads to increased percentage of publisher activity. Moreover the predicitve analysis can be extended, therefore also the year 2012 should be taken into account. For computational reasons, the focus in this work was on the year 2013. However, when specializing on specific paymentgroups of interest, as outlined in the work, this issue can be solved.

To conclude, the current work gives an interesting insight into the publisher base of an affiliate network and reveals the underlying factors for payment. Since online marketing is subject to a constant change and has excess to more and more data, room for future analysis remains.

Bibliography

Canty, A. und Ripley, B. D. (2014): *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-11.

Davis, C. (2002): *Statistical methods for the analysis of repeated measurements*. Springer texts in statistics, Springer, ISBN 9783540953708.

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

Destatis (2014): URL `https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2012/12/PD12_422_63931.html`, german Federal Statistical Office; Accessed: 2014-08-17.

Efron, B. (1979): *Bootstrap Methods: Another Look at the Jackknife*. The Annals of Statistics, 7 (1), 1–26, URL `http://dx.doi.org/10.1214/aos/1176344552`.

Efron, B. und Tibshirani, R. (1994): *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, ISBN 9780412042317, URL `http://books.google.de/books?id=gLlpIUxRntoC`.

Efron, B. und Tibshirani, R. (1997): *Improvements on Cross-Validation: The 632+ Bootstrap Method*. Journal of the American Statistical Association, 92 (438), 548–560.

Europe, I. (2014): *IAB Europe Adex Benchmark 2013*. URL `http://www.iabeurope.eu/news/european-online-advertising-market-records-new-high-273bn`, accessed: 2014-07-10.

Everitt, B. und Hothorn, T. (2011): *An Introduction to Applied Multivariate Analysis with R*. Use R!, Springer, ISBN 9781441996503.

Fahrmeir, L., Kneib, T., Lang, S. und Marx, B. (2013): *Regression: Models, Methods and Applications*. Springer London, Limited, ISBN 9783642343322.

Fahrmeir, L. und Tutz, G. (1994): *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics, Springer New York.

Hastie, T., Tibshirani, R. und Friedman, J. (2009): *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics, Springer, cited on p. 249.

Lammenett, E. (2013): *Praxiswissen Online-Marketing: Affiliate- und E-Mail-Marketing, Suchmaschinenmarketing, Online-Werbung, Social Media, Online-PR*. Springer Fachmedien Wiesbaden, ISBN 9783658033125.

Longford, N. (1993): *Random Coefficient Models*. Oxford science publications, Clarendon Press, ISBN 9780198522645, cited on page 235.

McCullagh, P. und Nelder, J. A. (1989): *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series, Chapman & Hall.

McCulloch, C. E. und Searle, S. R. (2001): *Generalized, Linear, and Mixed Models*. New York: John Wiley and Sons.

Meingast, M. (2013): *Affiliate Marketing: Analyse zeitlicher Aspekte im Online-Shopping*. master thesis, Institut für Statistik, Ludwig-Maximilians-Universität München.

Nelder, J. A. und Wedderburn, R. W. M. (1972): *Generalized Linear Models*. Journal of the Royal Statistical Society. Series A (General), 135 (3), 370–384.

Pinheiro, J. und Bates, D. (2000): *Mixed-Effects Models in S and S-PLUS*. Statistics and computing, U.S. Government Printing Office, ISBN 9780387989570.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. und R Core Team (2012): *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-105.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

Sakamoto, Y., Ishiguro, M. und Kitagawa, G. (1986): *Akaike information criterion statistics*. Mathematics and its application. Japanese series, KTK Scientific Publishers, ISBN 9789027722539.

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

Skrondal, A. und Rabe-Hesketh, S. (2004): *Generalized Latent Variable Modeling*. Chapman & Hall.

Statista (2014): Accessed: 2014-07-10.

Tutz, G. (2012): *Regression for categorical data*. Cambridge series in statistical and probabilistic mathematics, Cambridge and New York: Cambridge University Press, ISBN 9781107009653.

Welham, S., Cullis, B., Gogel, B., Gilmour, A. und Thompson, R. (2004): *Prediction in linear mixed models*. Australian & New Zealand Journal of Statistics, 46, 325–347, doi:10.1111/j.1467-842X.2004.00334.x.

Wickham, H. (2009): *ggplot2: elegant graphics for data analysis*. Springer New York, ISBN 978-0-387-98140-6, URL `http://had.co.nz/ggplot2/book`.

Wood, S. (2006): *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, ISBN 9781584884743.

Wood, S. N. (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.

Xie, Y. (2013): *Dynamic Documents with R and knitr*. Boca Raton, Florida: Chapman and Hall/CRC, URL `http://yihui.name/knitr/`, iSBN 978-1482203530.

Xie, Y. (2014): *knitr: A general-purpose package for dynamic report generation in R*. URL `http://yihui.name/knitr/`, r package version 1.6.

Zuur, A. F. (2009): *Mixed effects models and extensions in ecology with R*. Statistics for Biology and Health, Springer New York, ISBN 978-0-387-87458-6.

Appendix

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BM	Businessmodel
KAM	Key Account Managed (publisher)
LMM	Linear Mixed Models
LRT	Likelihood Ratio Test
PPL	Pay per Lead
PPS	Pay per Sale
REML	REstricted (or "REsidual") Maximum Likelihood

Appendix A

Appendix for the Descriptive Analysis

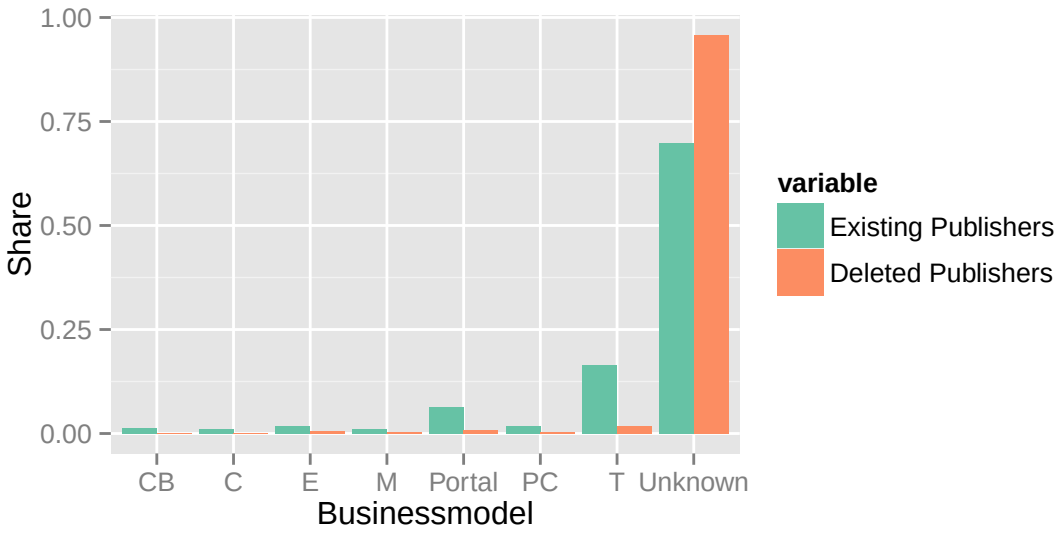


Figure A.0.1: Publishers share by businessmodel, separated by deletion state.

businessmodel	numberP	shareP	numberP_del	shareP_del
CB	2753	0.013	372	0.002
C	2683	0.012	394	0.002
E	4135	0.019	1287	0.006
M	2072	0.010	554	0.003
Portal	13752	0.063	1912	0.009
PC	4227	0.019	565	0.003
T	35803	0.165	3704	0.018
Unknown	151914	0.699	201025	0.958

Table A.1: Comparison of number and share of existing vs. deleted publishersbefore 2013

businessmodel	number	share	sum	Mean	sharePayment	numberP	shareP
CB	1312	0.019	17143978	13067	0.185	2753	0.013
C	1447	0.021	24420779	16877	0.263	2683	0.012
E	1753	0.026	1359609	776	0.015	4135	0.019
M	990	0.015	8600423	8687	0.093	2072	0.010
Portal	8051	0.119	3452801	429	0.037	13752	0.063
PC	2388	0.035	2425366	1016	0.026	4227	0.019
T	22022	0.325	35352005	1605	0.381	35803	0.165
Unknown	29720	0.439	31350	1	0.000	151914	0.699

Table A.2: Number and share of Publishers activ by total payment per BM (left) followed by their sum, mean and share of payment (middle) in comparison to total number and share per BM (right)

paymentgroup	number	share	sum	Mean	sharePayment
0	46615	0.689	0	0	0.000
1	20189	0.298	5783228	286	0.062
2	568	0.008	7706521	13568	0.083
3	116	0.002	4863522	41927	0.052
4	79	0.001	6712625	84970	0.072
5	67	0.001	12803897	191103	0.138
6	25	0.000	11071312	442852	0.119
7	24	0.000	43845205	1826884	0.473

Table A.3: Overview over number of publishers and share by paymentgroup, accomplished by their sum, mean and share of total payment in 2013. Paymentgroup 0 has payments smaller or equal to one, while paymentgroup 7 earned more than €600.000 in 2013.

businessmodel	number	share	shareClicks	shareOrders	sharePayment	numberP	shareP
CB	1071	0.022	0.040	0.183	0.185	2753	0.013
C	1228	0.025	0.059	0.312	0.263	2683	0.012
E	1268	0.026	0.013	0.030	0.015	4135	0.019
M	765	0.016	0.666	0.092	0.093	2072	0.010
Portal	6748	0.137	0.046	0.064	0.037	13752	0.063
PC	2035	0.041	0.015	0.036	0.026	4227	0.019
T	18975	0.386	0.154	0.281	0.380	35803	0.165
Unknown	17052	0.347	0.005	0.001	0.000	151914	0.699

Table A.4: Number and share of Publishers activ by traffic per BM (left) followed by their shares of clicks, oders and payment (middle) in comparison to total number and share per BM (right)

Appendix B

Appendix for the Analysis of the Data

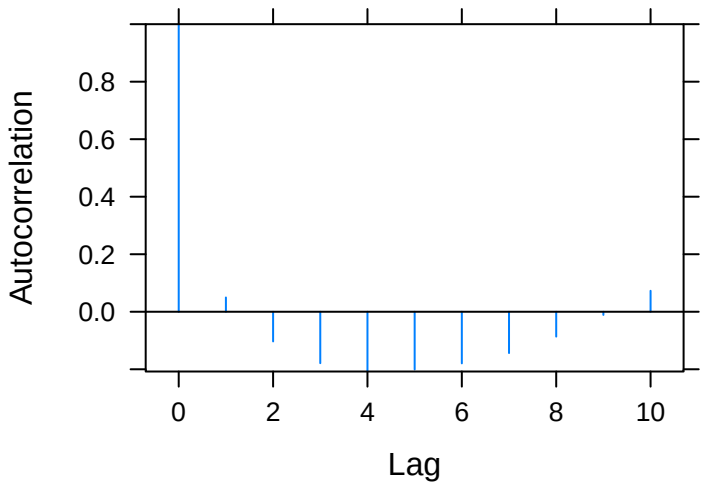


Figure B.0.1: ACF plot for random intercept and slope model with AR(1) correlation structure.

	No KAM	KAM
Intercept	2.50 (0.02)***	4.54 (0.10)***
Age in Months	−0.01 (0.00)***	−0.01 (0.00)**
Month	−0.05 (0.00)***	−0.11 (0.01)***
Orders	0.00 (0.00)***	0.00 (0.00)***
$Month^2$	0.00 (0.00)***	0.01 (0.00)***
Status susp.	0.26 (0.04)***	0.47 (0.15)**
Status ok	0.93 (0.03)***	0.53 (0.08)***
Status oktop	2.53 (0.68)***	1.67 (0.17)***
Status bl.bypre	0.47 (0.72)	−1.19 (2.38)
Exist. PS	0.00 (0.00)***	0.00 (0.00)***
Cashback	0.03 (0.06)	0.37 (0.29)
Coupon	0.14 (0.06)*	0.22 (0.23)
Email	0.06 (0.04)	−0.12 (0.25)
Media	0.13 (0.08)	0.60 (0.28)*
Portal	−0.06 (0.03)*	−0.14 (0.20)
PC	0.03 (0.04)	−0.08 (0.24)
unknown	−0.44 (0.02)***	−0.50 (0.15)***
Month:orders	0.00 (0.00)	0.00 (0.00)***
AIC	255577.33	25855.27
BIC	255792.26	26018.95
Log Likelihood	-127765.66	-12904.63
Num. obs.	84554	9123
Num. groups	17437	1180

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B.1: Coefficients from random intercept and slope model with AR1, separated by KAM factors

	Prechecked	Ok	Oktop	Suspicious
Intercept	2.42 (0.02)***	4.43 (0.06)***	6.87 (0.40)***	3.93 (0.20)***
Age in Months	−0.01 (0.00)***	−0.03 (0.00)***	0.02 (0.01)*	−0.01 (0.00)***
Month	−0.04 (0.01)***	−0.10 (0.01)***	−0.08 (0.03)*	−0.08 (0.05)
Orders	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***
$Month^2$	0.00 (0.00)***	0.01 (0.00)***	0.00 (0.00)	0.00 (0.00)
KAM	1.12 (0.05)***	0.78 (0.07)***	0.08 (0.29)	1.52 (0.31)***
Exist. PS	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)**
Cashback	0.01 (0.07)	0.28 (0.17)	0.24 (0.82)	0.72 (0.35)*
Coupon	0.17 (0.06)**	0.02 (0.14)	0.38 (0.74)	0.14 (0.31)
Email	0.10 (0.04)*	−0.14 (0.11)	−2.36 (1.44)	0.28 (0.27)
Media	0.19 (0.09)*	−0.17 (0.19)	1.08 (0.58)	−0.11 (0.50)
Portal	−0.01 (0.03)	−0.25 (0.09)**	−1.21 (0.65)	−0.46 (0.20)*
PC	0.08 (0.04)	−0.20 (0.14)	−1.49 (0.66)*	−0.13 (0.30)
unknown	−0.38 (0.02)***	−0.63 (0.08)***	−1.36 (0.48)**	−1.26 (0.32)***
Month:orders	0.00 (0.00)***	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AIC	215027.22	55787.95	3582.00	6793.03
BIC	215210.49	55945.66	3688.24	6904.37
Log Likelihood	-107493.61	-27873.98	-1771.00	-3376.52
Num. obs.	70546	19653	1513	1948
Num. groups	15807	2614	158	629

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B.2: Coefficients from random intercept and slope model with AR1, separated by status factors

	Variance	Standard error	Correlation
(Intercept)	2.679612	1.63695	(Intr)
Month	0.008568	0.09257	-0.219
Residual	0.827138	0.90947	

Table B.3: Variance and correlation components of Cashback model.

	Variance	Standard error	Correlation
(Intercept)	3.93237	1.9830	(Intr)
Month	0.02237	0.1496	-0.552
Residual	1.08893	1.0435	

Table B.4: Variance and correlation components of Coupon model.

	Variance	Standard error	Correlation
(Intercept)	1.951244	1.3969	(Intr)
Month	0.007814	0.0884	-0.272
Residual	1.206353	1.0983	

Table B.5: Variance and correlation components of Email model.

	Variance	Standard error	Correlation
(Intercept)	3.3193547	1.82191	(Intr)
Month	0.0008217	0.02867	-0.154
Residual	1.2011169	1.09595	

Table B.6: Variance and correlation components of Media model.

	Variance	Standard error	Correlation
(Intercept)	1.908717	1.38156	(Intr)
Month	0.007377	0.08589	-0.445
Residual	0.792931	0.89047	

Table B.7: Variance and correlation components of Portal model.

	Variance	Standard error	Correlation
(Intercept)	1.749746	1.32278	(Intr)
Month	0.007669	0.08757	-0.326
Residual	1.027818	1.01381	

Table B.8: Variance and correlation components of PC model.

	Variance	Standard error	Correlation
(Intercept)	2.291697	1.51384	(Intr)
Month	0.008267	0.09092	-0.397
Residual	0.786497	0.88685	

Table B.9: Variance and correlation components of Topic model.

	Variance	Standard error	Correlation
(Intercept)	1.378939	1.17428	(Intr)
Month	0.003772	0.06141	-0.442
Residual	0.374451	0.61192	

Table B.10: Variance and correlation components of Unknown model.

Electronical Appendix

The attached electronical appendix (USB Stick) contains the following files:

1. **Data:** The data file contains the data used for the analysis.
2. **R Code:** The R codes finally used in the analysis.
3. **Thesis:** contains the Thesis in PDF-format.

Declaration of Authorship

I hereby confirm that I have authored this master thesis independently and without use of others than the indicated resources.

Isabel Matheja

Munich, August 26th, 2014