Ludwig-Maximilians-Universität
München Institut für Statistik

Analyse von Online-Mietwohnungsanzeigen

Bachelorarbeit

Tobias Kaller

betreut von

Prof. Dr. Göran Kauermann und Dr. Michael Windmann

13. März 2019

Gender-Erklärung

Aus Gründen der besseren Lesbarkeit wird in dieser Arbeit die Sprachform des generischen Maskilinums angewandt. Es wird an dieser Stelle darauf hingewiesen, dass die ausschließlich Verwendung der männlichen Form geschlechtsunabhängig verstanden werden soll!

Abstract

Diese Arbeit beschäftigt sich mit der Struktur der Online-Wohnungsanzeigen in München. Dazu wurden die Webseite „immobilienscout24.de“ über mehrere Jahre, von 2011 bis Ende 2017, täglich mit einem Web-Crawler analysiert. Konkret wurde der These auf den Grund gegangen, ob sich die Anzeigedauer, wie lang eine Wohnungsanzeige online ist, über den evaluierten Zeitraum verändert und inwiefern andere Faktoren, insbesondere die Anzahl der Anzeigen, die gleichzeitig verfügbar sind, aber auch z.B. die Lage, einen signifikanten Einfluss auf die Anzeigedauer haben.

Zu Beginn wird die These der Arbeit genauer definiert und beschrieben. Dieser Definition folgt eine Beschreibung der Datenherkunft, welches Pre-Engeneering nötig war und eine deskriptive Analyse der wichtigsten Variablen. Im Anschluss an die Datenbeschreibung folgt der theoretische Hintergrund zu dem verwendeten „Accelerated Failure Time“-Modell und dem darin enthaltenen Rezessionsmodell. Dabei zeigt die Modellierung, dass das Startdatum der Anzeige einen signifikanten Effekt auf die Anzeigedauer der Wohnungsanzeige hat. Jedes Jahr sind die Anzeigen kürzer verfügbar als noch im Vorjahr. Des weiteren geht das Modell auf Faktoren ein, die die Anzeigedauer beeinflussen. Besonders wird ersichtlich, dass Wohnungen im Stadtkern deutlich kürzer angezeigt werden, als Wohnungen im Randbereich von München. Auch beeinflussen der Preis pro Quadratmeter, die Zimmerzahl und Etage der Wohnung die Anzeigedauer.

Inhaltsverzeichnis

1 Einleitung

„Jede Person, die in gut ausgewählte Immobilien im wachsenden Bereich einer wohlhabenden Gegend investiert, wendet die sicherste Methode an, um unabhängig zu werden, denn Immobilieninvestitionen sind das Fundament des Wohlstands.“
(Sarego GmbH, 2017)

Dieses Worte, frei aus dem Englischen übersetzt, soll einst Theodore Roosevelt[1] geäußert haben. Ob Immobilien wirklich das Fundament des Wohlstands sind, darüber ließe sich bestimmt streiten, jedoch gilt die Grundaussage des Satzes damals wie heute. Immobilen sind für Investoren ein wichtiges Finanzvehikel, um regelmäßige Rendite zu erzielen. Nachdem die Spekulation auf Miet- und Verkaufspreise in Metropolen, wie New York oder London, zu Extrempreisen geführt haben, scheinen Investoren nun, immer auf der Suche nach Objekten mit höherer Rendite, auf deutsche Großstädte gestoßen zu sein. Laut Sven Heinen (02.11.2018) hat sich der Preis pro Quadratmeter zwischen dem Jahr 2007 und 2016, für eine Eigentumswohnung (Neubau) im Schnitt verdoppelt.
Diese Entwicklung wird von vielen kritisch gesehen. Die Stadt München hat sich erst Ende November 2018 dazu entschlossen, ihr Vorkaufsrecht für 300 Wohnungen in Sendling geltend zu machen. Ziel des Ganzen ist es, zu zeigen, dass in München bezahlbarer Wohnraum möglich ist. (Portal München Betriebs-GmbH & Co. KG, n.d.)

Ungeachtet der politischen Entwicklung dieses Themas beschäftigt sich diese Arbeit mit der Veränderung des Immobilienmarktes in München über die letzten Jahre. Im Detail wird untersucht, ob Wohnungen, verursacht durch die Wohnungsknappheit, kürzer online angeboten werden und ob die Anzahl der Wohnungen, die gleichzeitig online sind, kleiner ist, als noch vor einigen Jahren. Anhand von verschiedenen Kennzahlen wird versucht diese Strukturänderung zu zeigen und belegen. Dazu liegen Daten des Internetportals immobilienscout24.de, betrieben durch die Immobilien Scout GmbH, für einen Zeitraum von ca. neun Jahren bis Mitte Februar 2018 vor.

Zu Beginn dieser Arbeit wird der Ursprung, Aufbau und die Struktur des Datensatzes analysiert. Diese Analyse beinhaltet unter anderem auch die Erläuterung der wichtigsten Variablen und das Pre-engineering des Datensatzes. Im Anschluss an die vorausgehende Analyse wird genauer auf die Theorie der verwendeten statistischen Modelle und die Anwendung derer auf die Daten eingegangen. Es folgt eine Zusammenfassung des Modells und der wichtigsten Ergebnisse. Am Ende der Arbeit wird noch ein Ausblick gegeben, welche weiteren Analysen sinnvoll sein könnten.

[1] *1858 bis †1919, 26. Präsident der Vereinigten Staaten von Amerika 1901 bis 1909

2 These der Arbeit und Einführung in die Daten

Dieser Abschnitt widmet sich dem Thema, welche These in dieser Arbeit untersucht wird. Im Anschluss wird auf die Erhebung und Herkunft der Daten eingegangen, sowie eine deskriptive Analyse durchgeführt.

2.1 These der Arbeit

Die Grundthese dieser Arbeit besagt, dass sich der Wohnungsmarkt in München auf Grund der Wohnungsknappheit verändert hat und dass dies quantifizierbar ist. Diese Veränderung wird durch die Anzeigedauer repräsentiert. Aus der Grundthese ergeben sich folgende Leitfragen für diese Arbeit:

- Wie verändert sich die Anzeigedauer der einzelnen Wohnungsanzeigen über den evaluierten Zeitraum?
- Hat die Anzahl der Anzeigen die gleichzeitig online sind, einen Einfluss auf die Anzeigedauer?
- Gibt es neben der Zeit und der Anzahl der Anzeigen, weitere Variablen, die die Anzeigedauer beeinflussen?

2.2 Datenerhebung und -herkunft

Die dieser Arbeit zugrundeliegenden Daten stammen aus dem Internetportal „immobilienscout24.de“. Diese Daten wurden durch ein Verfahren namens Web-Scraping erhoben. Dabei wird ein Computer so programmiert, dass er in vordefinierten Zeitabständen die Daten einer Webseite ausliest. Diese Programme analysieren meistens die HTML und/oder andere Script Dateien einer Webseite. Der Algorithmus wird an Strukturen in den Dateien angepasst und speichert die entsprechenden Werte als Variablen im Datensatz. (Schrenk, 2012, S. 37ff.)

2.2.1 „immobilienscout24.de“

„immobilienscout24.de“ wird betrieben durch die Immobilien Scout GmbH, eine Marke der Scout24 AG. (Scout24 AG, n.d.) Auf diesem Portal werden online Immobilien jeglicher Art, Wohnungen, Häuser, Gewerbehallen, etc., zur Miete und zum Verkauf angeboten. Dabei können Anbieter von Wohnungen, privat oder gewerblich, sehr detailliert offerieren. Daher besteht der daraus entstandene Datensatz aus 170 Variablen, die eine einzige Anzeige beschreiben.

Preismodell Das Preismodell von „immobilienscout24.de“ ist nicht direkt aus der Website erkennbar, muss jedoch für die folgenden Analysen betrachtet werden. Ein genauer Anzeigenpreis wird erst nach Angabe aller Objektdaten angezeigt und schwankt nach eigenen Recherchen zwischen 49, 90 € und 399, 90 €, Stand Dezember 2018. Dies ist abhängig von der Dauer der Anzeige (14 Tage, 1 Monat oder 3 Monate) und dem Anzeigetarif (Basic-, Top- oder Premium-Anzeige). Es gibt Vergünstigungen für gewerbliche Anbieter mit vielen Anzeigen oder Mieter, die selbst Nachmieter suchen. (Immobilien Scout GmbH, n.d.)

Marktanteil Nach einer Studie des Immobilienverband IVD Bundesverband e.V. (2018) werden mittlerweile über 99% der Wohnungsanzeigen online aufgegeben. Dort ist „immobilienscout24.de“ das zweitgrößte deutsche Portal für Anzeigen im Bereich Immobilien im Internet. Des Weiteren geht aus dieser Studie hervor, dass 74,2% der Befragten auf „immobilienscout24.de“ inserieren und 84% der zu vermietenden Wohnungen auf mehr als einem Portal angeboten werden. Somit ist „immobilienscout24.de“ eine gute Datenquelle, um einen Großteil aller angebotenen Wohnungen des Gesamtmarkts zu berücksichtigen.

2.3 Datenbeschreibung und Pre-Engeneering

Es wurden einmal täglich alle Anzeigen auf „immobilienscout24.de“ mittels eines Web-Scraper analysiert. Der Datensatz beinhaltet alle Anzeigen, die vom 24. April 2009 bis zum 31. Dezember 2017 online gegangen sind. Lokal erfolgte die Auswahl über eine Postleitzahl-Liste, sodass alle offerierten Wohnungen in Landeshauptstadt München und Landkreis München-Land erfasst wurden. Dadurch hat der Web-Scraper insgesamt 86.996 Anzeigen mit 170 Variablen für diesen Datensatz indiziert. Dabei hat der Algorithmus in der Variable Startdatum dokumentiert, wann eine Anzeige das erste Mal erkannt wurde und in der Variable Enddatum, wann eine Anzeige nicht mehr online war. Daraus lässt sich errechnen, wie lang eine Anzeige online war und somit wie lang der Vermieter wahrscheinlich benötigt hat ausreichend Interessenten zu finden. Da der Web-Scraper nur einmal alle 24 Stunden die Daten erhoben hat, ist eine Abweichung von bis zu 48 Stunden in den so ermittelten Dauern möglich. Ebenfalls umfasst der Datensatz die Koordinaten der angebotenen Wohnungen und die Postleitzahl, so wie die Adresse, falls vorhanden. Daten, wie Mietpreise (Kalt- und Warmmiete), Größe der Wohnung, Anzahl der Zimmer, Baujahr und viele weitere Merkmale sind ebenfalls indiziert. Da diese unter anderen keine Pflichtangaben sind, sind diese teilweise unvollständig.

2.3.1 Pre-Enegneering

Es wurden einige Anpassungen an dem Datensatz vorgenommen. Im Folgenden werden die Anpassungen und deren Zweck kurz erläutert.

Anzeigedauer Jede Anzeige hat ein Startdatum, aber nicht zwangsläufig ein Enddatum. Eine Anzeige ohne Enddatum war noch online, bevor der Datensatz erstellt wurde. Daher wurde das Enddatum für alle fehlenden Werte auf das maximale Enddatum, den 14. Februar 2018, gesetzt. Alle Anzeigen nach dem 31.12.2017 werden nur noch für die Berechnung der Anzeigedauer benutzt, da keine neuen Anzeigen aufgenommen werden. Dies beutet im Umkehrschluss, dass die Anzahl der Anzeigen, die online sind, nicht richtig ist. Aus den Anfangs- und Enddaten wurde die Anzeigedauer in Tagen und in Wochen berechnet und als neue Variable aufgenommen. Wochen wurden generell aufgerundet, da zum Beispiel die 0,3. Woche umgangssprachlich die „erste Woche“ ist.

Wohnungen mit einer Anzeigedauer von mehr als 90 Tagen wurden aus dem Datensatz entfernt, da die längste Anzeigedauer laut „immobilienscout24.de“ 90 Tage beträgt und eine extrem linkssteile Verteilung vorliegt.

Anzeige Datum Es wurde ein neuer Datensatz erstellt, dieser enthält die Anzahl der Anzeigen die zu einem bestimmten Tag online waren. Dabei stellte sich heraus, dass gerade am Anfang der Messung nur sehr wenige Wohnungen aufgezeichnet wurden. Während der gesamten Periode, mit Ausnahme am Anfang, war die Anzahl der Anzeigen auf einem Mindestwert von 333 Anzeigen. Daher wurden Daten, an denen weniger als 333 Anzeigen online waren, aus dem Datensatz entfernt. Konkret betrifft das alle Daten vor und inkl. dem 27. Dezember 2011. Mit diesem Datensatz lässt sich auch die Anzahl der Wohnungen bestimmen, die gleichzeitig online waren. In dem neuen Datensatz sind des weiteren einige Variablen übernommen, z.B. als prozentualer Anteil, aber auch die Wohnungen nach Stadtteil. Die Anzahl der Anzeigen wurde wiederum in den Originaldatensatz hinzugefügt. Dazu wurde der Mittelwert der Anzahl der Anzeigen über das ganze Intervall, in dem eine Beobachtung angezeigt wurde, für jede einzelne Beobachtung berechnet.

Postleitzahl Als Variable zur Lokalisierung wird die Postleitzahl verwendet. Im Datensatz sind insgesamt 75 Postleitzahlen für München aufgeführt. Dies beinhaltet unter anderem auch den Stadtteil "Haar". Für diesen liegen aber nur 21 Beobachtungen vor, weswegen diese für die alle folgende Analysen entfernt wurden.

Neubauwohnungen Neubauwohnungen sollen nicht Teil der Analyse werden. Daher wurden entsprechende Anzeigen aus dem Datensatz entfernt.

2.3.2 Deskriptive Analyse der relevanten Variablen

Anzeigedauer Die Dauer der Anzeige wurde, wie im Kapitel 2.3.1 dargelegt, aus dem Start- und Enddatum der berechnet. Durch das Entfernen von Anzeigen, die länger als 90 Tage online sind, verläuft diese Variable zwischen 1 und 90 in ganzen Zahlen. Der Median liegt bei 11 und der Durchschnitt bei 19,53 Tagen. Der Minimalwert liegt bei 1, was bedeutet, dass die Anzeige nur an einem Datum vom Web-Crawler indiziert wurde. Das erste Quantil befindet sich bei 4, das dritte Quantil bei 31 und die maximale Anzeigedauer bei 90 Tagen.

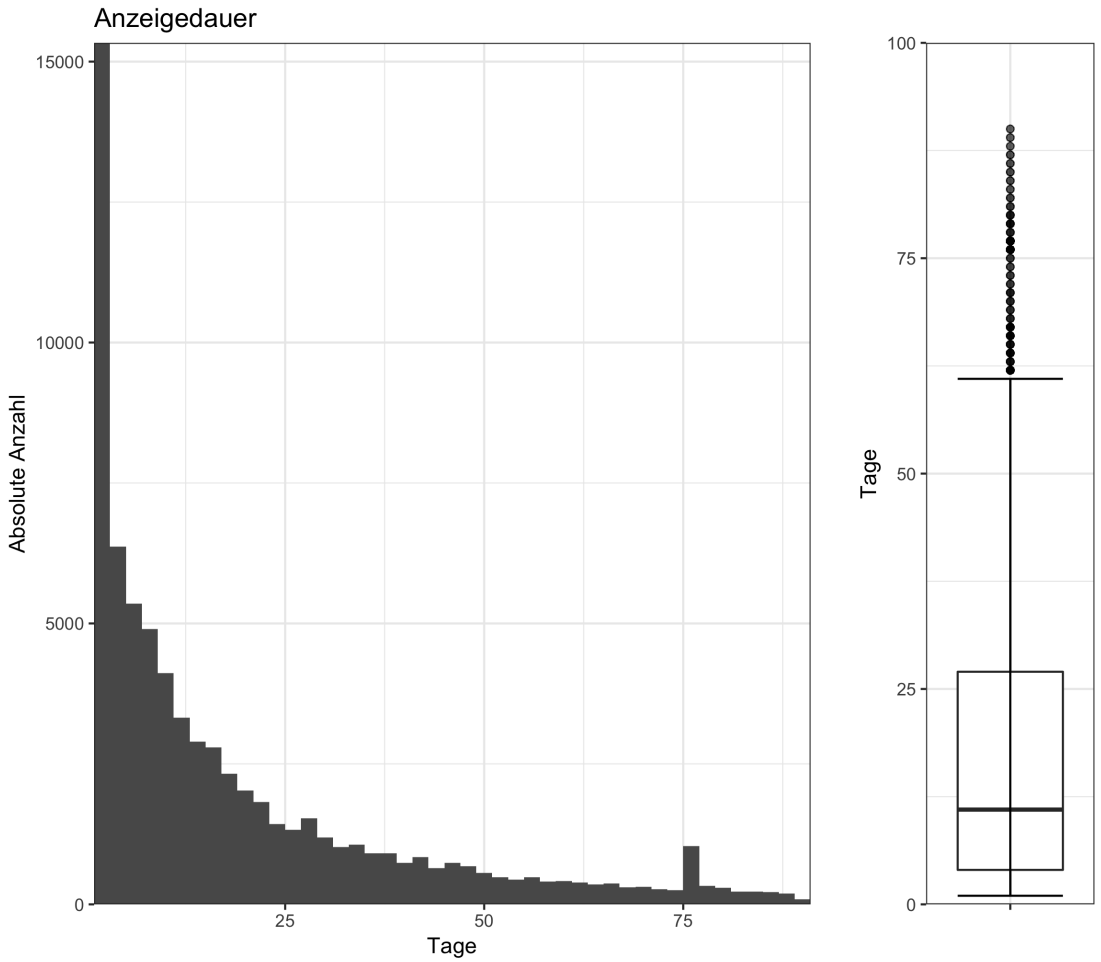

Abbildung 1: Links: Histogramm der Anzeigedauer in Tagen, rechts: Boxplot der Anzeigedauer in Tagen

Das Histogramm in Abbildung 1 ist stark linksseitig steil. Ein Balken entspricht zwei Tagen. Die meisten Wohnungen sind nur wenige Tage online. Bei genauer Analyse ergibt sich, dass die meisten Wohnungen nur zwei Tage inseriert sind. Der Anteil der Anzeigen, die zwischen 75 und 77 Tage online sind, fällt aus dem Raster und ist ca. doppelt so hoch wie die Umliegenden. Bei detaillierter Betrachtung der Daten sieht man, dass eine Dauer von 75 Tagen überdurchschnittlich oft vorkommt. Dies lässt sich nicht durch das Preismodell von „immobilienscout24.de“ erklären, da hier der Ausreißer bei 14, 30 bzw. 90 Tagen liegen müsste. Auch der Boxplot spiegelt die linkssteile Verteilung wider. Durch diese Verteilung entsteht der extreme Unterschied von 8,53 Tagen zwischen Median und Mittelwert.

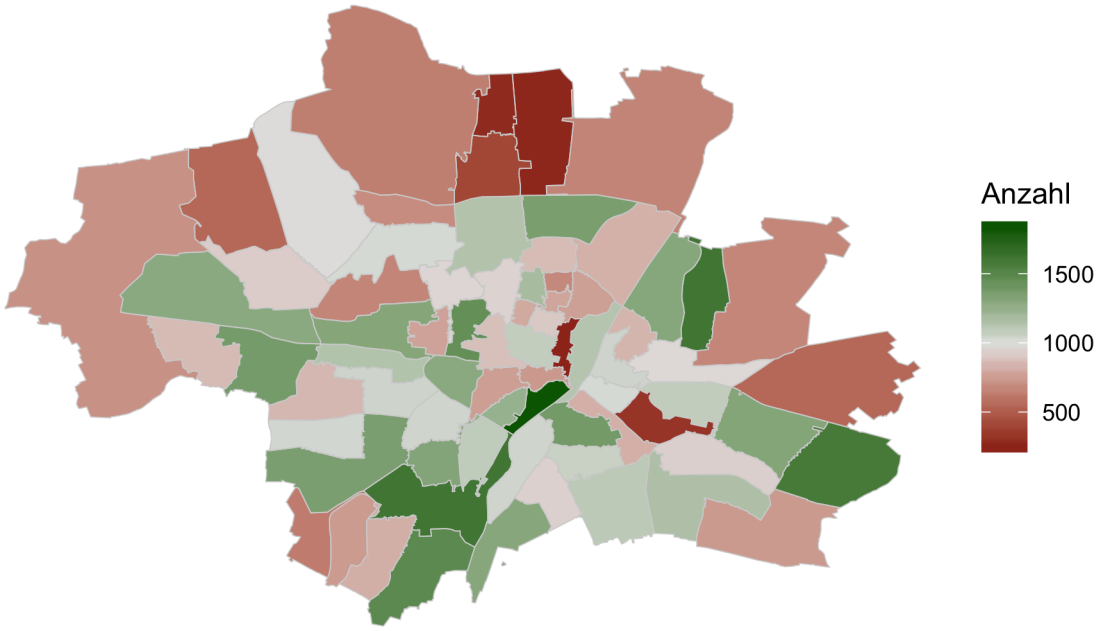

Abbildung 2: Absolute Anzahl der angebotenen Wohnungen verteilt auf die 74 Postleitzahlgebiete Münchens

Ortsdaten Auf „immobilienscout24.de“ ist es keine Pflicht genaue Angaben bzgl. der Adresse zu machen. Jedoch ist die Postleitzahl zu jeder Anzeige gegeben, daher wird diese verwendet, um strukturelle Unterschiede innerhalb von München zu verdeutlichen.

In Verbindung mit der Anzahl der Wohnungen im untersuchen Zeitraum Betrachtet man die Postleitzahlen genauer, so ist in Abbildung 2 sehr deutlich zu

sehen, dass große Unterschiede bei der Anzahl der angebotenen Wohnungen zu finden sind. Grün werden all diejenigen Postleitzahlgebiete, in denen viele Wohnungen offeriert werde, aus roten Gebieten wurde nur wenige Anzeigen aufgezeichnet. Im Mittel gab es ca. 1010 Anzeigen und der Median liegt bei 984 Anzeigen pro Postleitzahlgebiet. Das Gebiet mit den wenigsten angebotenen Wohnungen ist mit 249 Stück (0,35%) über den gesamten Zeitraum „80539“ im Herzen Münchens. Diese geringe Anzahl lässt sich durch die Lage erläutern. Das Gebiet erstreckt sich vom bayrischen Nationaltheater im Süden bis hin zum Siegestor im Norden und wird im Westen durch die Ludwigstraße und im Osten durch den Englischen Garten begrenzt. Ein Großteil der Fläche ist durch öffentliche Bauten, wie z.B. die Residenz, Staatsbibliothek und das Bayrische Nationaltheater, viele Gebäude der Luwig-Maximilians-Universität München und dem Hof-, so wie Finanzgarten nicht für den Mietwohnungsmarkt verfügbar. Die meisten Wohnungen, 1837 oder 2,55%, wurden im Süd-Westlichen Teil der Isarvorstadt mit der Postleitzahl 80469 angeboten. Es gibt keine offensichtliche Erklärung für diesen überdurchschnittlichen Wert.

In Verbindung mit dem Median der Anzeigedauer Durch die extrem linkssteile Verteilung der Anzeigedauer, eignet sich für die Auswertung in Kombination mit den Postleitzahlgebieten am besten der Median und nicht der Mittelwert der Anzeigedauer. Betrachtet man die Verteilung des Medians der Anzeigedauer einzelner Wohnungsanzeigen in den verschiedenen Postleitzahlgebieten, so sieht man in Abbildung 3 sehr gut, dass Wohnungen im Stadtkern deutlich kürzer offeriert werden, als Wohnungen am Stadtrand. Lediglich am „Marienplatz“ und um den„Englischen Garten“ sind Wohnungen länger inseriert. Dies könnte aber ebenfalls, wie bei der Anzahl der Wohnungen, an der besonderen Laage und den damit verbundenen überdurchschnittlichen Mietkosten liegen. Eine Ausnahme stellt hier am Stadtrand das „Hasenbergl“ mit der Postleitzahl „80933“ dar. Dort und in der „Maxvorstadt“(„80799“) liegt der Minimalwert mit 7 Tagen im Median. Mit einer Anzeigedauer im Median von 19 Tagen sind die Wohnungen aus dem Stadtteil „Solln“, „81479“, im Süden Münchens am längsten inseriert.

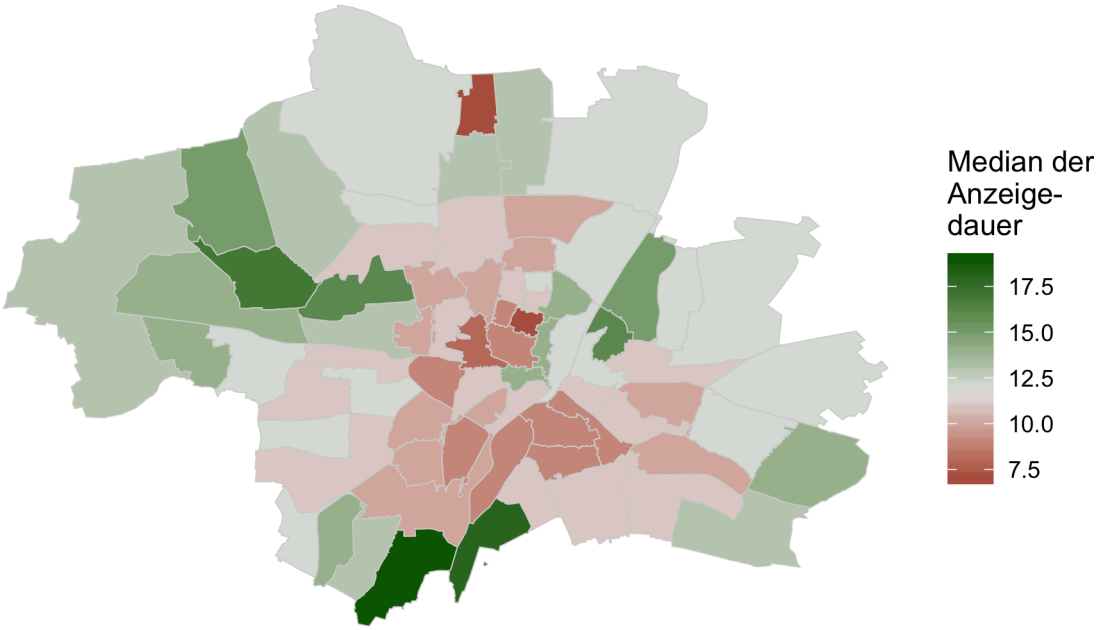

Abbildung 3: Median der Anzeigedauer in Tagen verteilt auf die 74 Postleitzahlgebiete Münchens

Die Korrelation zwischen der Anzeigedauer und der Anzahl beträgt $-0,0328 \approx 0$, somit gibt es fast keine Abhängigkeit zwischen den beiden Variablen.

Nach Datum umstrukturierter Datensatz Wie bereits im Kapitel 2.3.1 erwähnt, wurde der Datensatz so umstrukturiert, dass interessante Variablen anhand des Datum betrachtet werden können.

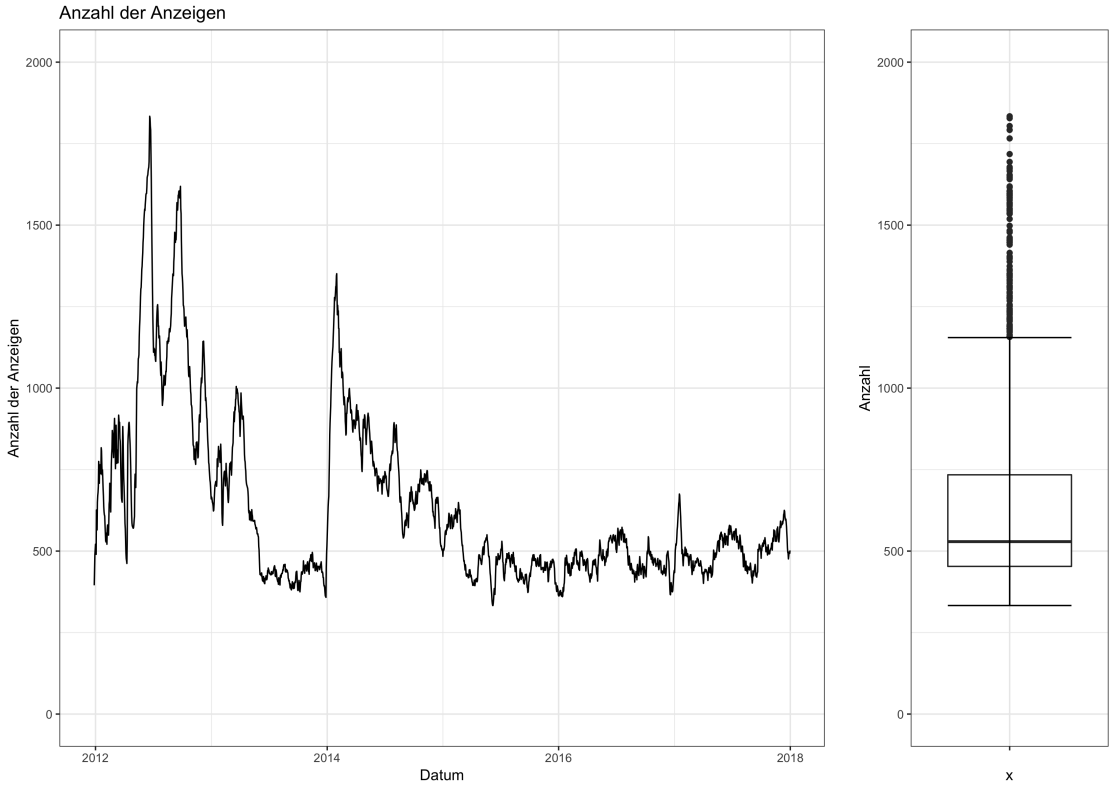

Abbildung 4: Links: Anzahl der Anzeigen nach Anzeigedatum, rechts: Boxplot der Anzahl der Anzeigen

Anzahl der Anzeigen Es ist deutlich links in Abbildung 4 zu sehen, dass die Anzahl der Wohnungsanzeigen zwischen 2012 und 2015 stark schwankt, das Minimum liegt bei 358 und das Maximum bei 1834. Zwischen 2015 und 2018 jedoch schwankt die Anzahl der Anzeigen nicht mehr sehr stark und ist dauerhaft zwischen 333 und ca. 675. Die Werte des Boxplots befinden sich inkl. des Mittelwerts in Tabelle 1.

Min	1.Quantiel	Median	Mittelwert	3.Quantiel	Max
333	453	529	633.6	734	1834

Tabelle 1: Fünf Punkte Zusammenfassung der Anzahl der Anzeigen inklusive Mittelwert.

Auffällig ist der Januar 2014, in dem es zu einem massiven Anstieg der angebotenen Wohnungen kommt, so wie Mitte 2012. Es lässt sich nicht erklären, woher diese Auffälligkeiten in Abbildung 4 kommen. Es scheint auch kein natürliches Wachstum zu sein, da im Januar 2014 die Anzahl der Anzeigen sich fast verdreifacht. Es

gab zwar eine Gesetzesänderung, die Vermieter veranlassen ihre Wohnungen einzustellen, diese trat aber schon 2013 in Kraft. Eine weitere Erklärung durch z.B. ein Neubaugebiete, in denen eine große Zahl von Wohnungen einstanden ist, im Januar 2014 fertiggestellt wurde. Da Neubauwohnungen jedoch nicht im Datensatz enthalten sind, ergibt dies auch keinen Sinn.

3 Regression

„Regression ist die wohl am häufigsten eingesetzte statistische Methodik zur Analyse empirischer Fragestellungen in Wirtschafts-, Sozial- und Lebenswissenschaften.“ (Fahrmeir, Kneib and Lang, 2007, S. 1)
Regressionsmodelle stellen die Abhängigkeit zwischen einer erklärenden Variablen X und einer Zielvariable Y dar und sind somit die Basis vieler statistischer Analysen. Der Zusammenhang stellt sich approximiert wie folgt dar.

$$Y = f(X) + \epsilon$$

mit
X und Y Variablen mit den Ausprägungen x_i und y_i
f: deterministische Regressionsfunktion in Abhängigkeit von X
ϵ: zufälliger Fehler, mit $E(\epsilon_i) = 0$, $i = 1, \ldots, n$

(Fahrmeir, Künstler, Pigeot and Tutz, 2007, S. 475)

Das gesamte Kapitel basiert, sofern nicht explizit erwähnt, auf Fahrmeir, Kneib and Lang (2007).

3.1 Lineare Regression

Dieses Regressionsmodell trifft die sehr starke Annahme, dass der Zusammenhang zwischen den Variablen linear und Y Normalverteilt ist, mit $Var(Y) = \sigma^2$. Somit ergibt sich für $f(X) = \alpha + \beta X$ wobei α der sogenannte Intercept und β der Regressionskoeffizient ist. α und β sind bei der Erstellung des Modells unbekannt und sind somit die zu schätzenden Parameter. Diese Klasse der Modelle heißt Lineare Modelle (LM)(Fahrmeir, Künstler, Pigeot and Tutz, 2007, S. 575 ff.)

3.1.1 Der Koeffizienten

Laut Fahrmeir, Künstler, Pigeot and Tutz (2007, S. 480 f.) wird in der linearen Regression die Kleinste-Quadrat-Methode (KQ) verwendet, um die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ zu bestimmen. Dies bedeutet, dass die Summer der quadratischen Fehler ϵ_i möglichst gering ausfallen soll:

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i) \rightarrow \min_{\alpha,\beta}$$

Somit ergeben sich für β und α die folgenden Schätzer und Eigenschaften:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon_i}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{Y}_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Des Weiteren sind sind die Schätzer $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\sigma}^2$ erwartungstreu.

Naheliegen ist auch, die Parameter mit der Maximum-Likelihodd-Methode zu schätzen. Diese liefert für α und β dieselben Schätzer, die KQ Schätzung entspricht in diesem Falle einem Likelihood-Schätzer. (Fahrmeir, Kneib and Lang, 2007)

3.2 Multiple lineare Regression

Häufig ist es der Fall, dass Y nicht nur von einer erklärenden Variable abhängig ist, sondern von mehreren erklärenden Variablen $X_1, \ldots, X_p$, somit erhält man auch mehrere Regressionskoeffizienten $\beta_1, \ldots, \beta p$. Dies wird in der Funktion $f(X) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, i = 1, \ldots, n$, dargestellt. Daraus ergibt sich wiederum folgendes Grundmodell:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon i, \qquad i = 1, \ldots, n$$

(Fahrmeir, Künstler, Pigeot and Tutz, 2007, S. 494 f.)

Ebenfalls ist eine Matrixschreibweise des Modells möglich:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \vec{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_i \end{pmatrix}$$

Daraus ergibt sich dieses kompakte Grundmodell:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \vec{\epsilon}, \qquad E(\vec{\epsilon}) = 0$$

(Fahrmeir, Künstler, Pigeot and Tutz, 2007, S. 503 ff.)

3.3 Interaktionseffekt

Variablen sind nicht immer unabhängig voneinander. Wird eine Abhängigkeit zischen zwei oder mehreren Variablen vermutet, so muss diese als sogenannter Inter-

aktionseffekt in das Modell aufgenommen werden:

$$f(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

In diesem einfachen Beispiel ist x_1 und x_2 als Haupteffekt in die Gleichung mit aufgenommen, so wie der Interaktionseffekt zwischen den beiden Variablen. (Fahrmeir, Kneib and Lang, 2007, S. 84 f.)

3.4 Generalisierte lineare Regression

Während man bei der klassischen linearen Regression davon ausgeht, dass Y normalverteilt ist, können bei der generalisierten linearen Regression Zielvariablen modelliert werden, die nicht einer Normalverteilung folgen, sondern einer Verteilung die zu den Verteilungen der Exponentialfamilie gehört. Das Modell der generalisierten linearen Regression (GLM), ist somit eine Erweiterung der multiplen linearen Modelle. Diese Modelle nennt man Generalisierte lineare Modelle, oft auch als „GLM“ abgekürzt. (Fahrmeir, Kneib and Lang, 2007, S. 189)

In einem GLM wird die Zielvariable in der Regel nicht direkt geschätzt, sondern ein sogenannter Prädiktor η_i.

$$\eta_i = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

(Fahrmeir, Kneib and Lang, 2007, S. 210)

Für Exponentialfamilien gilt:

1. Die Dichte lässt sich schreiben, als:

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

2. θ: Natürlicher Parameter

3. $f(y|\theta)$ lässt sich normieren

4. $b'(\theta)$ und $b''(\theta)$ existieren

5. Für Erwartungswert und Varianz gilt:

$$\mathbb{E}(y) = \mu = b'(\theta), \qquad Var(y) = \phi b''(\theta)/\omega$$

(Fahrmeir, Kneib and Lang, 2007, S. 218)

β-Schätzer Die Koeffizienten werden mittels ML-Schätzung geschätzt. Daraus ergibt sich, dass $\hat{\beta}_n \overset{a}{\sim} N(\beta, \mathbf{F}^{-1}(\beta))$ ist. (Fahrmeir, Kneib and Lang, 2007, S. 224)

3.4.1 Link

Um den geschätzten Erwartungswert der Zielvariable zu erhalten, muss eine sogenannten Link-Funktion $g(\mu)$ auf den Prädiktor angewandt werden. Dieser Lineare Prädiktor ist abhängig von der Verteilungsannahme der Zielvariable.

Verteilung	$E(\mu) = b'(\theta)$	$b''(\theta)$	$Var(y) = b''(\theta)\phi/\omega$
Normal	$\mathbb{E}(y) = \mu = \theta$	1	σ^2/ω
Bernulli	$\phi = \frac{exp(\theta)}{1+exp(\theta)}$	$\phi(1-\phi)$	$\phi(1-\phi)\omega$
Poisson	$\lambda = exp(\theta)$	λ	λ/ω
Gamma	$\mu = 1 - 1/\theta$	μ^2	$\mu^2\nu^-1/\omega$
Inverse Gauß	$\mu = (-2\theta)^{-1/2}$	μ^3	$\mu^3\omega^2/\omega$

Tabelle 2: Tabelle für Erwartungswert und Varianz gängiger Exponentialfamilien. Quelle: Fahrmeir, Kneib and Lang 2007, S.219

Aus den Erwartungswerten der Tabelle 2 können die Linkfunktionen errechnet werden, wenn die Formel des Erwartungswerts nach θ umgestellt wird. Somit ergibt sich zum Beispiel für die Normalverteilung der sogenannte natürliche Link, für binäre Zielvariablen der Logit-Link und für das Poisson-Modell ein log-linearer Link. (Fahrmeir, Kneib and Lang, 2007, S. 220) Dies bedeutet konkret für den Linearen Prädikator:

Natürlicher Link: $g(\mu) = \mu$
Logit Link: $g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$
Log-linearer Link: $g(\mu) = log(\mu)$

3.4.2 Interpretation des Modells

Eine Interpretation der Koeffizienten ist somit nicht direkt aus dem errechneten Modell möglich, sondern es muss auf die individuellen Modellspezifikationen eingegangen werden. Als Beispiel hier ein fiktives Logit-Link Modell mit folgender Annahme:

$$\eta_i = \beta_0 + \beta_1 x_{i1}$$

mit X_1: Alter in Jahren, Y: Binäre Variable, Beobachtungseinheit braucht eine Brille und $i = 1, \ldots, n$.

Als Ergebnis seinen nun $\hat{\beta}_0 = 0.5$ und $\hat{\beta}_1 = 0.1$. Aus $\hat{\beta}_0$ können keine Interpretationen erfolgen, jedoch $\hat{\beta}_1$. Da die Zielvariable binär ist, aber die Modellformel aber ein stetiges Ergebnis liefert, können nur Wahrscheinlichkeiten angegeben werden. Für einen 20 jährigen ergibt sich durch $exp((0.5 + 0.1 * 20)/(1 + exp(0.5 + 0.1 * 20))) = 0.9241\ldots$ eine Wahrscheinlichkeit von ca. 92% eine Brille zu benötigen. Das Problem an dieser Berechnung ist, dass nur definierte Einzelfälle errechnet werden können. Möchte man nun aber den Koeffizienten des Faktors interpretieren, so ist das nur möglich, vordefinierte Intervalle auf das Odd's zu interpretieren. Dazu wird sich z.B. ein 20 und 25 Jähriger angeschaut und direkt verglichen. Durch kürzen in der Gleichung bleibt $exp(5*\hat{\beta}_1) = exp(5*0.1) = 1.6487\ldots$ übrig. Dies ist das sogenannte Odd's, die multiplikative Veränderung Chance unter Y zu leiden, wenn man 5 Jahre älter ist. Hier gilt zu beachten, dass er Effekt multiplikativ auf die Chance wirkt und $5 * exp(\hat{\beta}_1) \neq \exp(5 * \hat{\beta}_1)$ ist. Die zu interpretierende Erhöhung der Variable X_1 muss demnach individuell errechnet werden und es kann nur die Erhöhung der Chance unter Y zu leiden errechnet werden.

3.5 Modellwahl und Variablenselektion

Häufig stehen, wie im Fall der Immobiliendaten sehr viele Variablen zur Verfügung, die einen Einfluss auf die Zielgröße haben. Häufig werden Herangehensweisen verwendet, die aber nicht optimal sind. Daher wird in diesem Abschnitt auf die möglichen Kriterien eingegangen wie ein optimales Modell gefunden und beurteilt werden kann.

3.5.1 Kriterien

Hier werden einige Kriterien vorgestellt, die im späteren Verlauf der Arbeit wichtig sind.

AIC - Informationskriterium nach Akaike Das Informationskriterium nach Akaike (AIC) ist ein häufig angewendetes Kriterium, um zwei Modelle miteinander zu verglichen.

$$AIC = -2l(\hat{\beta}_M, \sigma^2) + 2|M + 1|$$

mit:
$l(\hat{\beta}_M, \sigma^2)$: Maximaler Wert der Log-Likelihood
$|M + 1|$: Geschätzte Anzahl der Parameter

Fällt das AIC für ein Modell kleiner aus als das eines anderen, so hat es laut AIC eine bessere Vorhersagetreue.

BIC - Informationskriterium nach Bayes Das Bayesianische Informationskriterium sieht dem AIC sehr ähnlich verfolgt aber ein anderen Ansatz:

$$BIC = -2l(\hat{\beta}_M, \sigma^2) + log(n)|M|$$

Im Gegensatz zum AIC bestraft das BIC vor allem komplexe Modelle, somit einstehen einfachere Modelle, als bei der Verwendung des AIC. Wie beim AIC auch, ist ein Modell besser, wenn es ein kleineres BIC annimmt.

CV - Kreuzvalidierung Die CV basiert nicht wie das AIC oder BIC auf der Abweichung der Log-Likelihood, sondern verfolgt den Ansatz, dass die Parameter anhand von vielen Datensätzen geschätzt werden. Dazu werden die vorhandenen Daten in K Teildatensätze geteilt und für jeden Teildatensatz werden die Koeffizienten geschätzt. Die nicht im Teildatensatz enthaltenen Daten werden verwendet, um das Modell zu testen. Die Information der Abweichung zwischen errechneten und echten Werten des Testdatensatzes wird in das Kriterium aufgenommen. Starke Abweichungen führen zu einem höheren CV.

Ein Spezialfall stellt die „leave one out“ KV dar. Dabei wird für die Schätzung der Teilmodelle, jede Beobachtung einmal weggelassen, ansonsten werden alle zur Schätzung des Modells verwendet. Daher entstehen mit N Beobachtungen auch $N = K$ Teildatensätzen und Modelle. Daraus entsteht folgendes Kriterium:

$$CV = \frac{1}{n} \sum \left(y_i - (\hat{y}_{iM})^{-i} \right)^2$$

3.5.2 Variablenselektion

Es gibt verschiedene Methoden, das optimale Modell zu finden. Wobei es nicht unbedingt ein optimales Modell geben muss, jedes Kriterium kann ein eigens Modell als optimal herausstellen. Es gibt verschiedene Algorithmen, die eine Variablenselektion ermöglichen. Hier werden nur drei vorgestellt:

- „Vorwärts-Selektion“: Diese Art der Modellwahl nimmt in jedem Itarationsschritt eine neue Variable, ausgehend von einem leeren Modell, auf. Dabei wird über die Kriterien verglichen, welche neue Variable den größten Effekt auf die Kriterien hat. Das Verfahren bricht ab, wenn keine Verbesserung mehr möglich ist.

- „Rückwärts-Selektion“: Bei dieser Art passiert die Selektion wie bei der der „Vörtwärts-Selektion“, jedoch wird von einem Modell mit allen Einflussvariablen ausgegangenen und es werden die Variablen abgezogen, so dass sich die Kriterien verbessern.

- „Schrittweise-Selektion“: In jedem Iterationsschritt ist es möglich, dass Variablen abezogen oder hinzugefügt werden.

Im Gegensatz zu dem hier nicht näher betrachteten „Leaps and bounds“-Algorythmus liefern die drei genannten Verfahren nicht das bestmögliche Modell, allerdings in den meisten Fällen sehr gute Modelle.

Diagnose Wurde ein geeignetes Modell gefunden, so gilt es dies zu untersuchen. Fahrmeir, Kneib and Lang (2007, S. 168ff.) stellt dabei drei Merkmale als zentral heraus:

- Überprüfen der Modellannahme

- Ungewöhnliche Beobachtungen untersuchen

- Kollinearitätsanalyse

Detaillierte Beschreibungen der einzelnen Methoden zur Untersuchung der Merkmale findet man in Fahrmeir, Kneib and Lang (2007) und werden hier nicht genauer ausgeführt.

3.6 Signifikanz

Eine wichtige Rolle in der Interpretation der Effekte spielt die Signifikanz. Dazu werden die einzelnen Koeffizienten getestet, ob diese einen signifikanten Einfluss auf das Ergebnis haben. Dazu verwendet man einen sogenannten F-Test und stellt die Hypothese auf, dass ein Koeffizient keinen Einfluss hat:

$$H_0 : \beta_i = 0 \qquad vs. \qquad H_1 : \beta_i \neq 0$$

Details zum F-Test finden sich bei Fahrmeir, Kneib and Lang (2007, S. 112 - 115). Im Folgenden wird nur auf das Spezielle Testproblem eingegangen, gezeigt und bewiesen durch Fahrmeir, Kneib and Lang (2007, S. 116f.). Es lässt sich zeigen, dass

$$F = \frac{\hat{\beta}_i^2}{\hat{Var}(\beta_i)} \sim F_{1,n-p}$$

gilt, mit
p: Anzahl der Freiheitsgrade
n: Anzahl der Beobachtungen

Alternativ dazu kann der t-Test durchgeführt werden

$$t = \frac{\hat{\beta}_i}{se\hat{(}\beta_i)}$$

, mit $se\hat{(}\beta_i) = Var\hat{(}\beta_i)^{\frac{1}{2}}$.

Der kritische Wert für den Ablehnbereich von H_0 ergibt sich in beiden Fällen durch das Signifikanzniveau α. H_0 wird abgelehnt, wenn

$$|t| \leq t_{1-\frac{\alpha}{2}}(n-p)$$

oder

$$F > F_{1,n-p}(1-\alpha)$$

gilt.

Das Ergebnis wird durch den p-Wert ausgedrückt. Dieser Hypothesentest überprüft den Fehler 1. Art (Nullhypothese fälschlicherweise nicht verworfen) zum Niveau α. Dieser Test wird von sehr vielen Paketen für Regression in R automatisch für jeden β-Koeffizienten durchgeführt und mit dem Modell ausgegeben. Dadurch kann überprüft werden, inwiefern eine Koeffizient nur zufälligerweise diesen Wert angenommen hat.

4 Überlebensdaten-Analyse

4.1 Einführung in die Überlebenszeitanalyse

Laut Clark et al. (2003) haben Überlebenszeitmodelle einige Eigenschaften, die sie von herkömmlichen Modellen unterscheiden. Dabei ist die Überlebenszeit das Zeitintervall, bis zum Eintreten eines Ereignisses, in den meisten Fällen der Tod oder ein Rückfall. Das Modell werde besonders häufig in der Krebsforschung angewandt und modelliert die Zeit zwischen einen festgelegten Zeitpunkt, z.B. dem Zeitpunkt der Diagnose, bis zum Tod des Patienten oder dem Zeitpunkt des erfolgreichen Abschluss einer Behandlung bis zu einem Rückfall. Des Weiteren führen die Autoren an, dass in den meisten Fällen die Überlebenszeit keiner Normalverteilung folgt und bei Vielen das Ereignis schneller eintrete, wenige aber überleben noch das Ende der Messung. Dies sind die Eigenschaften, auf die ein Überlebenszeitmodell angepasst werden muss.

In dieser Arbeit wird die Überlebenszeit einer Wohnung gemessen. Diese Variable wurde schon unter 2.3.1 eingeführt als Anzeigedauer in Tagen.

4.1.1 Zensierte Daten

In den meisten Überlebensdaten-Analysen stößt man auf das Problem, dass man nicht von allen Bobachtungen das exakte Zeitintervall ermitteln kann. (Kleinbaum and Klein, 2005, S. 5) Dies liegt z.B. daran, dass bei einigen Beobachtungen bis zum Ende der Messung das Ereignis nicht eintritt. Man spricht in diesem Fall von zensierten Daten. Clark et al. (2003) stellen drei verschiedene Szenarien dar, warum Daten zensiert sein können:

1. Das Ereignis ist bis zum Ende der Studie noch nicht eingetreten.
2. Ein Patient verlässt die Studie.
3. Es tritt ein anderes Ereignis ein, das es für den Patienten unmöglich macht, weiter in der Studie teilzunehmen, z.B. tödlicher Autounfall.

Für die Anzeigedauer ist jedoch nur der erste Fall relevant, da Fall Zwei und Drei, geschuldet durch die in Kapitel 2.2 dargestellte Art der Datenerhebung, nicht gemessen werden können. Durch diese Art der Schätzung wird die Überlebenszeit unterschätzt und man spricht auch von richtig und rechts zensierten Daten. Es kann jedoch auch der Fall eintreten, dass der Startzeitpunkt nicht bekannt ist, dieser Typ der Zensur ist jedoch für die Anzeigedauer ebenfalls irrelevant. Diese Eigenschaften der Zensur erfordern es, spezielle Modelle und Formen der grafischen Darstellung zu

entwickeln.(Clark et al., 2003)
Ob eine Beobachtung zensiert werden muss wird dargestellt durch die binäre Variable δ.

$$\delta = \begin{cases} 0, & \text{Zensierte Beobachtung} \\ 1, & \text{Ereigniss ist eingetreten} \end{cases}$$

(Kleinbaum and Klein, 2005, S. 8)

4.1.2 „Survival"- Funktion

Zur Modellierung von Überlebensdaten werden allgemein zwei Wahrscheinlichkeiten verwendet, Überleben „Survival" und Gefahr „Hazard". Die Wahrscheinlichkeit zu überleben $S(t)$, ist die Wahrscheinlichkeit, die eine beobachtete Einheit hat, ab dem Start der Messung, bis zu einem spezifischen Zeitpunkt $t \in T, T \geq 0$ in der Zukunft zu überleben, $S(t) = P(T > t)$. In der Praxis ergibt sich aus dieser Funktion eine linkssteile Treppenfunktion, die nur Null erreichen kann, wenn es keine zensierten Beobachtungen gibt. (Kleinbaum and Klein, 2005, S. 9) Folgende Eigenschaften gelten laut Glomb (2007) für Survival-Funktionen:

1. $S(t) = 1 - F(t); t \geq 0$
2. Die Funktion ist momonton fallend mit $S(0) = 1$ und $lim_{t \to +\infty} = 0$
3. Falls T stetig ist, gilt:
$$S(t) = \int_t^{+\infty} f(u)du$$
4. Falls T diskret ist, ist S eine linkssteile Treppenfunktion:
$$S(t) = \sum_{t_j > t} p(t_j) = 1$$
mit $p(t_j) = P(T = t_j)$

4.1.3 „Hazard"-Funktion

Die Funktion der Gefahr $\lambda(t)$ oder $h(t)$, stellt die Wahrscheinlichkeit dar, dass das Ereignis am Zeitpunkt t, nach Beginn der Messung, eintritt.

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

Eine Hazard-Funktion stellt also die momentane Wahrscheinlichkeit zu einem Zeitpunkt t dar, dass ein Ereignis eintritt. Somit ist die Hazard-Wahrscheinlichkeit eine konditionaler Wahrscheinlichkeit, wobei das Ergebnis von $h(t)$ keine Wahrscheinlichkeit darstellt und das Ergebnis bei gleicher konditionalen Wahrscheinlichkeit von der gewählten Zeiteinheit abhängig ist, wie Kleinbaum and Klein (2005, S. 10 f.) darlegt. Die Hazard-Funktion wird außerdem in der parametrischen Analyse 4.3 von Überlebensdaten verwendet, um die Ausfallverteilung zu bestimmen. Diese liefern eine qualitative Information über den Ausfallmechanismus. Es gibt drei typische Formen, der ein Hazard-Funktion folgen kann:

1. Monoton steigend/fallend
2. Badewannen-förmig
3. Hügel-förmig

(Glomb, 2007)

Laut Kleinbaum and Klein (2005, S. 13) wird die Survival-Funktion häufiger verwendet, da diese sich intuitiver interpretieren lässt. Trotzdem hat die Hazard-Funktion ihre Daseinsberechtigung, da sie benötigt wird, um die momentane Wahrscheinlichkeit zu berechnen, um spezielle Modelle, wie z.B. Weibull, zu berechnen oder als Basis für eine weitere mathematische Modellierung.
Beide Funktionen stehen in einem Verhältnis zueinander, wobei die Überlebenswahrscheinlichkeit einen Fokus darauf liegt, dass kein Ereignis eintritt bis zum Zeitpunkt t und die Gefahr den Fokus auf das Eintreten legt (Clark et al., 2003). Dieses Verhältnis lässt sich folgendermaßen darstellen:

$$S(t) = exp\left[-\int_0^1 h(u)du\right] \qquad \text{oder} \qquad h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

(Kleinbaum and Klein, 2005, S. 14)

4.1.4 Erwartete Restlebensdauer

Eine weitere grundlegende Größe in der Analyse von Überlebensdaten ist die erwartete Restlebensdauer (MRL) diese gibt an mit welcher mittleren Lebensdauer ein Individuum des Alters t noch rechnen kann. Wie Glomb (2007) beweist, gelten folgende Eingenschaften:

$$mrl(t) = \mathbb{E}(T - t|T > t)$$

Ist T stetig, so gilt $mrl(t) = \frac{1}{S(t)} \int_t^{+\infty} S(u)du$
und
ist T diskret so gilt $mtl(t) = \frac{(t_{i+1}-t)S(t_i)+\sum_{j\leq i+1}(t_{j+1}-t_j)S(t_j)}{S(t)}$.
$mtl(t)$ entspricht daher der Fläche unterhalb der Survival-Funktion rechts von t geteilt durch $S(t)$ selbst.

4.2 Kaplan-Meier-Schätzer - Nicht-Parametrisches Modell

Eines der am häufigsten verwendeten und einfachsten Modelle ist der Kaplan-Meier-Schätzer (KM), $\hat{S}_{\mathrm{KM}}(t)$ und wird hier nur kurz als Basis für Survival-Funktionen angesprochen. Dieser schätzt die Wahrscheinlichkeit, dass ein Ereignis bei einer Beobachtungseinheit t_i nicht eintritt. Der KM-Schätzer benötigt keine Verteilungsannahme, daher wird dieser sehr gerne verwendet, wenn diese unbekannt ist.

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{t_{(i)}\leq t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{t_{(i)}\leq t} \left(1 - \frac{d_i}{n_i}\right)$$

mit
$\hat{S}(0) = 1$
d_i = Beobachtungseinheiten bei denen das Ereignis zum Zeitpunkt $t_{(i)}$eingetreten ist
n_i = Beobachtungseinheiten zum Zeitpunkt $t_{(i)}$

(Clark et al., 2003)

Abbildung 5 zeigt den Kaplan-Meier-Schätzer grafisch aufbereitet und ein Beispiel für $\hat{S}(5) = 0,5079$. Die Wahrscheinlichkeit, dass das gemessene Ereignis nach 5 Tagen noch nicht eingetreten ist, ist somit 50,79%. Die einzige Möglichkeit weitere Parameter in das Modell mit aufzunehmen ist, den Datensatz anhand von Parameter zu unterteilen und die Teildatensätze einzeln zu schätzen.

4.3 Geeignete Verteilungen für parametrische Überlebenszeitmodelle

Eine Alternative zu der nicht parametrischen Schätzung, wie den KM-Schätzer, stellt ein parametrisches Modell dar. Hierbei wird davon ausgegangenen, dass das Eintreten des Ereignisses einer Wahrscheinlichkeitsverteilung folgt. Dazu eignet sich im Prinzip jede Verteilung mit nicht-negativen Zufallsvariablen, jedoch hat sich herausgestellt, dass einige Verteilungen besonders geeignet für diesen Fall sind. Ob eine Verteilung sich zur Modellierung der Daten eignet, lässt sich unter anderem über die KM-Kurve grafisch beurteilen.

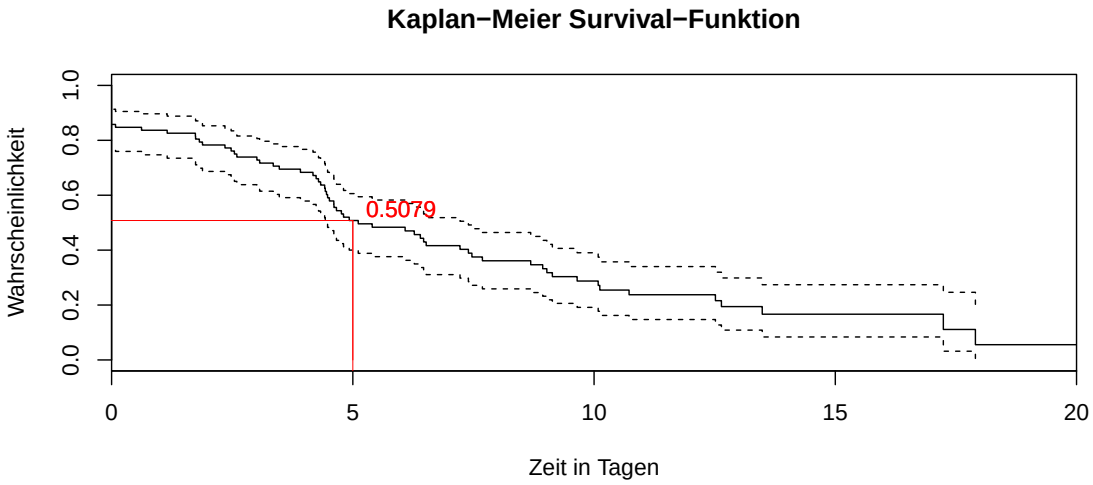

Abbildung 5: Kamplan-Meier-Schätzer von 100 zufällig generierten Daten, die gestrichelte Linie stellt die Standardabweichung des Kaplan-Meier-Schätzers dar. In Rot ein Beispiel für t = 5

4.3.1 Weibullverteilung

Diese Art der Verteilung wird häufig verwendet, um Lebensdauern statistisch zu modellieren.Dieser Abschnitt stützt sich auf Kleinbaum and Klein (2005) und Klösener et al. (2002, S.230 ff.). Der Vorteil der Weibull- gegenüber der Exponentialverteilung, besteht in der Berücksichtigung der Vergangenheit der beobachteten Objekte, sie ist „gedächtnisbehaftet“. Eine Weibull-verteilte Zufallsvariable ist stetig in den positiv reellen Zahlen und wird durch zwei Parameter λ und k definiert, mit $k, \lambda > 0$.

$$f(x) = \lambda k(\lambda x)^{k-1} exp(-(\lambda x)^k)$$

$$F(x) = 1 - exp(-(\lambda x)^k)$$

$$\mathbb{E}(X) = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{k}\right) \qquad Var(X) = \frac{1}{\lambda^2}\left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right)$$

mit Gammafunktion $\Gamma(\cdot)$.
Die zwei Parameter können so gewählt werden, dass die Form der Verteilung eine Normal-, Exponential- oder andere asymmetrische Verteilungen approximiert.

Formparameter Die Form der Verteilung wird maßgeblich durch k bestimmt. Daher spricht man hier auch vom Formparameter, bzw. Shape-Parameter. Für $k = 1$ ergibt die Weibullverteilung eine Exponentialverteilung $Exp(\lambda)$ mit konstanter Ausfallrate, für $k \geq 3.602$ verschwindet die Schiefe annähernd und die Verteilung nähert

sich einer Normalverteilung an.
$k > 1$ bedeutet, dass die Hazard-Wahrscheinlichkeit über die Zeit steigt, bei $k = 1$ bleibt sie gleich und bei $k < 1$ nimmt sie ab.

Skalenparameter Das sogenannte Skalenparameter ist $\frac{1}{\lambda} > 0$. Die charakteristische Lebensdauer T_c entspricht jener Zeitspanne in der bei 63.2% das untersuchte Ereignis eingetreten ist. λ wird häufig durch $\frac{1}{T_c}$ ersetzt. Außerdem gilt damit $T_c * \lambda = 1$.

Survival- und Hazard-Funktion Die Survival-Funktion ergibt aus $S(t) = 1 - F(t) = exp(-(\lambda t)^k)$ und daraus die entsprechende Hazard-Funktion $h(t) = \frac{f(t)}{S(t)} = \lambda k(\lambda t)^{k-1}$, mit Überlebenszeit t. Für $k = 1$ kann die Hazard-Funktion auf $h(t) = \lambda$ reduziert werden.

Schätzer Sowohl der Skalen- als auch Formparameter können nur iterativ geschätzt werden, z.B. über das Newtonverfahren. Auf das detaillierte Schätzverfahren wird hier nicht eingegangen, das Verfahren wird durch Klösener et al. (2002) dargelegt.

Eine weitere nützliche Eigenschaft der Weibullverteilung ist, dass $log(-log(S(t)))$ linear zu $log(t)$ verläuft. Dies erlaubt es die Parameter den $log(-log)$ KM-Schätzer gegen $-log$ der Zeit zu plotten und direkt zu interpretieren.
Betrachtet man das spezielle Weibull-Modell $S(t) = exp(-\lambda t^k)$ so ergibt sich für die $log(-log)$ transformation $ln(\lambda) + k \cdot ln(t)$.So stellt der erste Teil $log(\lambda)$ den Intercept dar.

4.3.2 Logarithmische Normalverteilung

Logarithmische Normalverteilung (logN) beschriebt eine kontinuierliche Wahrscheinlichkeitsverteilung, die stetig in der Menge der positiv reellen Zahlen liegt. Die beschrieben Zufallsvariable X ist logN, wenn die transformierte $Y = log(X)$ einer Normalverteilung folgt. Daraus ergeben sich folgende Eigenschaften für $x > 0$:

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

$$F(x) = \Phi\left(\frac{ln(x) - \mu}{\sigma}\right)$$

$$\mathbb{E}(Y) = exp\left(\mu + \frac{\sigma^2}{2}\right) \qquad Var(Y) = exp(2\mu + \sigma^2)(exp(\sigma^2) - 1)$$

Survival- und Hazard-Funktion Die Funktionen werden, wie in Kapitel 4.1.2, 4.1.3 und praktisch in Kapitel 4.3.1 gezeigt, gebildet.

4.4 Accelerated Failure Time -Modell

Häufig stammen die Daten nicht aus einer homogenen Testumgebung und weitere Faktoren wirken auf die Überlebenszeit. Daher betrachtet man nicht nur den Überlebenszeitvektor $T \geq 0$, sondern auch einen Kovariablenvektor $X = (x_1, \ldots, x_p)'$, mit p Anzahl der erklärenden Variablen. Accelerated Failure Time-Modelle (AFT) stellen eine Möglichkeit dar, diese zusätzlichen Parameter in die Berechnen mit aufzunehmen und sind somit parametrische Survival-Modelle. Da AFT-Modellierung sehr viel Rechenpower benötigt, konnte das AFT-Modell erst in den letzten zwei Dekaden an Bedeutung gewinnen.(Klein et al., 2014, S.58) Dieser Abschnitt stützt sich vor allem auf Klein et al. (2014) und Bradburn et al. (2003).

Im Gegensatz zu dem auch sehr bekannten proportional Hazard-Modell ist es bei AFT-Modellen nötig, eine Verteilungsannahme zu treffen.

$$S(t) = S_0(\varphi t),$$

mit
$S_0(t)$: Basis-Survival-Funktion mit $X = 0$
$\varphi = exp(\sum_{i=1}^{p} \beta_i x_i) = \exp^{X^\top \boldsymbol{\beta}}$: Beschleunigungsfaktor
$\sum_{i=1}^{p} \beta_i x_i = X^\top \boldsymbol{\beta} = \eta$: Linearer Prädiktor
p: Anzahl der Kovariablen
X: Kovariablenvektor, unabhängig von der Zeit
T: Überlebenszeitvektor

Das Konzept des AFT-Modell basiert nach Bradburn et al. (2003) darauf, dass der Effekt der Kovariablen die Überlebenskurve streckt oder staucht, bzw. beschleunigt oder verlangsamt. In anderen Worten bedeutet dies, dass die Kovariablen einen Einfluss auf die Überlebenszeit haben. In Abbildung 6 ist ein Beispiel eins AFT-Modell mit nur einer binären Kovariablen x_1 dargestellt.

$$x_1 = \begin{cases} 0 & \text{Patiente in Placebo Gruppe, mit } \varphi < 0 \\ 1 & \text{Patienten in Gruppe mit neuer Behandlung, mit } \varphi > 0 \end{cases}$$

Die Basisfunktion S_0 ist als durchgängige Linie eingezeichnet, die gestrichelte und

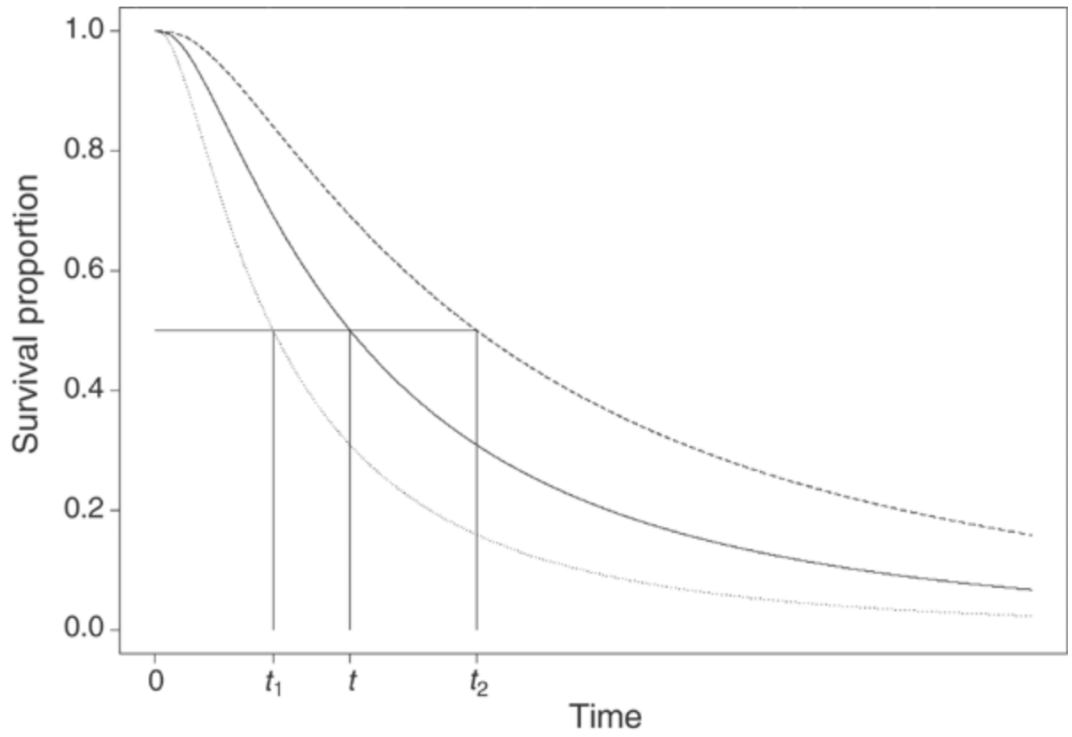

Abbildung 6: Beispiel eines AFT-Modells mit einer einzigen binären Variablen. Quelle: Bradburn et al. 2003

gepunktete Linie sind jeweils die gestauchten (φ<0) , oder gestreckten (φ>0) Funktionen $S_0(\varphi t)$. Patienten mit Placebobehandlung haben dieselbe Wahrscheinlichkeit (50%), dass das Ereignis eintritt zum Zeitpunkt t_1 wie Patienten mit neuer Behandlung zum Zeitpunkt t_2. Die erklärende Variable $x_1 \neq 0$ beschleunigt oder verlangsamt also den erwarteten Zeitpunkt für das Eintreten des Ereignisses gegenüber dem Modell ohne Einflussvariablen. (Bradburn et al., 2003) Der Effekt der Kovariablen ist multiplikativ, daher wird häufig, unter anderem für eine einfachere Interpretation, eine log-Transformation des Modells durchgeführt. Wenn $T > 0$ eine Lebensdauer ist, $Y = ln(T)$, $\boldsymbol{X}$ Matrix aus Kovariablen und Beobachtungen und $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, dann gilt folgender Zusammenhang:

$$Y = log(T) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + bZ$$

Die Verteilung von Z entspricht der Verteilungsannahme des Modells.

Die Schätzung der Parameter b, $\hat{\boldsymbol{\beta}}$ und $\hat{\beta}_0$ erfolgt über das Maximumlikelihood-Verfahren, eine detaillierte Ausführung des Verfahrens findet sich bei Glomb (2007,

S. 69 - 73). Gängige Softwarepakete für R verwenden für die Schätzung das Newton-Raphson-Verfahrens, gezeigt von Lee and Wang (2003, S. 428ff.).

Glomb (2007) zeigt, dass $\boldsymbol{\varphi} = -\boldsymbol{\beta}$ gilt:

$$\begin{aligned} S(t|\boldsymbol{X}) &= P[exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + bZ) > t] \\ &= P[\beta_0 + bZ > texp(-\boldsymbol{\beta}'\boldsymbol{X})] \\ &= S_0[t\ exp(-\boldsymbol{\beta}'\boldsymbol{X})] \end{aligned}$$

Wie bereits in 4.1.3 gezeigt gibt es zu jeder Survival-Funktion auch eine Hazard-Funktion:

$$\lambda(t) = \varphi\lambda_0(\varphi t)$$

Güte des AFT-Modell Es können dieselben Kriterien wie bei der Regressionsanalyse, in Kapitel 3.5.1 dargelegt, verwendet werden, um die Vorhersagegenauigkeit zweier AFTs zu vergleichen.

4.4.1 Interpretation

Die Interpretation der Parameter müssen an das Modell und dessen Verteilungsannahme von Z angepasst werden. Daher wird im Folgenden kurz auf die spezifischen Eigenschaften der Verteilungen aus Kapitel 4.3 eingegangen. Dieser Abschnitt basiert auf den Beweisen und Ausführungen von Glomb (2007, S. 74 - 80).

Weibull-Verteilung Das log-Transformierte Modell ist gegeben durch:

$$Y = ln(T) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + bZ$$

Wobei Z einer Standard-Gumbel-Verteilung, $f(x) = exp(-x)\ exp(-exp(-x))$ und $F(x) = exp(-x)$, mit x Teil der rationalen Zahlen, folgt. Glomb (2007, S. 60) zeigt, dass sich die Weibull und Gumbel-Verteilung extrem ähnlich sind. $\hat{b}$, $\hat{\boldsymbol{\beta}}$ und $\hat{\beta}_0$ werden durch das AFT-Modell geschätzt. Der Beschleunigungsfaktor $\hat{\varphi}$ kann für die i-te Beobachtung mit $\hat{\varphi} = exp(-\hat{\boldsymbol{\beta}}'X)$ berechnet werden. Löst man die Gleichung nach T auf, so erhält man mit dem Fehler Z:

$$\hat{S}_T(t|X) = exp\left(exp\left(\frac{ln(t) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'X}{\hat{b}}\right)\right); t > 0$$

Da gilt, dass $t_p = S^-1(1-p)$ das p-te Quantiel ist kann man für die Beobachtung

i z.B. den Median wie folgt berechnen:

$$\begin{aligned} t_{0,5}(X) &= S_T^{-1}(0,5|X) \\ &= exp(ln(-ln(0,5))\hat{b} + \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'X) \\ &= exp(ln(2)b + \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'X) \end{aligned}$$

Es kann gezeigt werden, dass $\frac{t_p(X_0)}{t_p(X)} = exp(-\hat{\boldsymbol{\beta}}'X) = \varphi$, mit X_0: Modell mit $\hat{\boldsymbol{\beta}}'X = 0$, gilt.

Die Hazard-Funktion kann, wie von Glomb (2007, S.78) gezeigt, für > 0 geschätzt werden:

$$\hat{\lambda}_{T_i}(t|X_i) = \frac{1}{\hat{b}\,t}\left(1 + t\,\hat{b}^{-1} exp\left(\frac{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}X_i}{\hat{b}}\right)\right)^{-1}$$

Logarithmische Normalverteilung Das log-lineare Modell besteht wie bei der Weibull-Verteilung aus:

$$ln(T) = \beta_0 + \hat{\boldsymbol{\beta}}'X + bZ$$

Jedoch geht das AFT-Modell bei einer logarithmischen Normalverteilung davon aus, dass Z Standardnormalverteilt ist ($Z \sim N(0,1)$). Ebenfalls wird durch das AFT-Modell $\hat{b}$, $\hat{\boldsymbol{\beta}}$ und $\hat{\beta}_0$ geschätzt. Der Beschleunigungsfaktor kann für die Beobachtung i ebenfalls direkt aus den Koeffizienten abgelesen werden $\hat{\varphi} = exp(-\hat{\beta}'X_i)$. Die Survival-Funktion kann wie folgt, für $t > 0$ geschätzt werden:

$$\hat{S}_{T_i}(t|X_i) = 1 - \Phi\left(\frac{ln(t) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'X_i}{\hat{b}}\right)$$

Für die Schätzung des $p$ten-Quantiel gilt:

$$\hat{t}_p(X_i) = exp(\Phi^{-1}(p)\hat{b} + \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'X_i)$$

Der Schätzer der Hazard-Funktion:

$$\hat{\lambda}_{T_i}(t|X_i) = \left(1 - \Phi\left(\frac{ln(t) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'X_i}{\hat{b}}\right)\right)^{-1} \Phi'\left(\frac{ln(t) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'X_i}{\hat{b}}\right) t^{-1}\hat{b}^{-1}$$

Interpretation von $\hat{\boldsymbol{\beta}}_i$ Da $\hat{\varphi}_i = exp(-\hat{\boldsymbol{\beta}}'X_i) = exp(-\hat{\beta}_1 X_{i_1} - \hat{\beta}_2 X_{i_2} - \ldots - \hat{\beta}_p X_{i_p})$ gilt, wirken die einzelnen Koeffizienten multiplikativ mit $exp(-\hat{\beta}_i)$ auf $\hat{\varphi}_i$. Der Effekt als prozentuale Änderung von t wahrgenommen werden. Wichtig dabei ist, dass eine β-Koeffizient nur vernünftig interpretiert werden kann, wenn alle anderen β-Koeffizienten gleich bleiben. Häufig wird für die Interpretation als ein Beispiel die

erwartete Lebenszeit von Hunden und Menschen verwendet. Im Volksmund spricht man davon, dass ein Menschenjahr sieben Hundejahre ist, im Modell würde dies wie folgt ausgedrückt werden:

$$S_{Mensch}(t) = S_{Hund}(7t) \Rightarrow \varphi = 7 \Leftrightarrow \beta = -1,9459$$

Ein Mensch hat also nach sieben Lebensjahren dieselbe Überlebens- bzw. Sterbewahrscheinlichkeit, wie ein Hund nach einem Jahr.

5 Modell zur Quantifizierung der Anzeigedauer

Im folgenden Kapitel wird das Modell zur Quantifizierung der Anzeigedauer erstellt und analysiert, um die Thesen dieser Arbeit, vgl. Kapitel 2.1, zu überprüfen.

5.1 KM-Schätzer der Anzeigedauer

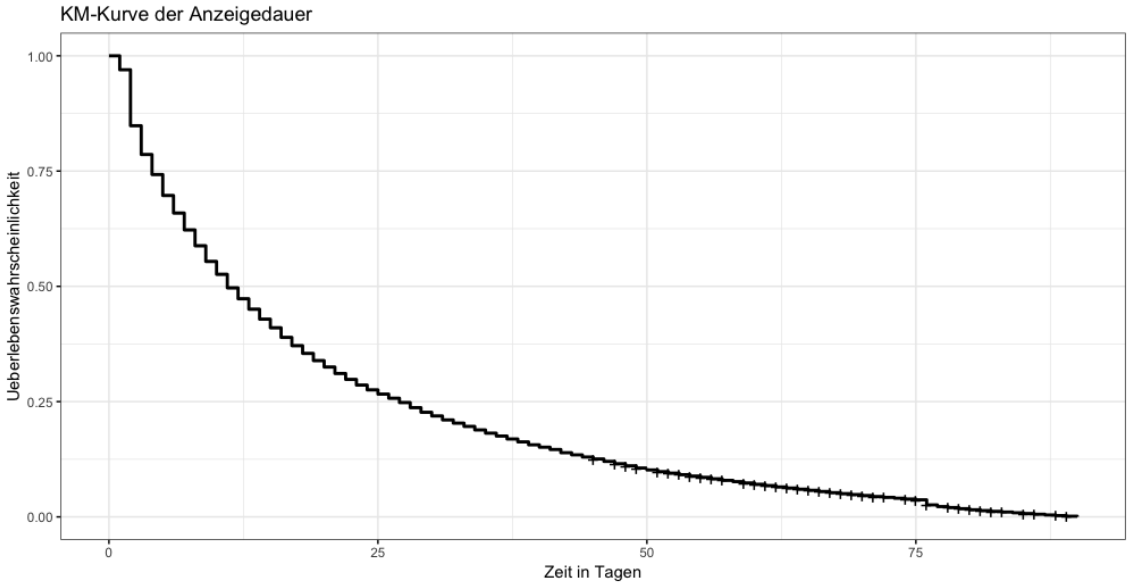

Abbildung 7: KM-Kurve der Anzeigedauer

In Abbildung 7 ist gut zu sehen, was schon in Kapitel 2.3.2 angerissen wurde. Wenige Anzeigen, ca. 3% werden innerhalb der ersten 24 Stunden wieder entfernt, jedoch werden sehr viele, ca. $12,1\%$, in einer Zeitspanne zwischen 24 und 48 Stunden entfernt. Nach 4 Tagen unterschreitet der KM-Schätzer schon die 75% und nach 11 die 50% Überlebenswahrscheinlichkeit.

In Abbildung 8 wurde der KM-Schätzer auf Teildatensätze angewandt. Dabei ergibt sich für jedes Jahr in dem eine Anzeige eingestellt wurde, eine neue Kurve, so wie darunter die absolute Anzahl der Anzeigen, die noch unter Risiko stehen. Es ist deutlich zu sehen, dass mit den Jahren die Kurve immer linkssteiler wird und somit die Anzeigen schneller wieder vom Portal entfernt werden. Die Kurve von 2011 ist mit Vorsicht zu betrachten, da das Jahr 2011 nur zu einem Bruchteil im Datensatz enthalten ist, dies lässt sich auch an der Anzahl unter Risiko erkennen. Während 2011 lediglich 158 Anzeigen überhaupt unter Risiko standen, lag die Zahl ab 2012

im Bereich zwischen 9520 und 14066.
Aus Abbildung 8 wird also deutlich, dass über das untersuchte Zeitintervall die Wohnungen kürzer online waren, je später sie eingestellt wurden. Im Folgenden wird versucht, diesen Effekt genauer zu quantifizieren.

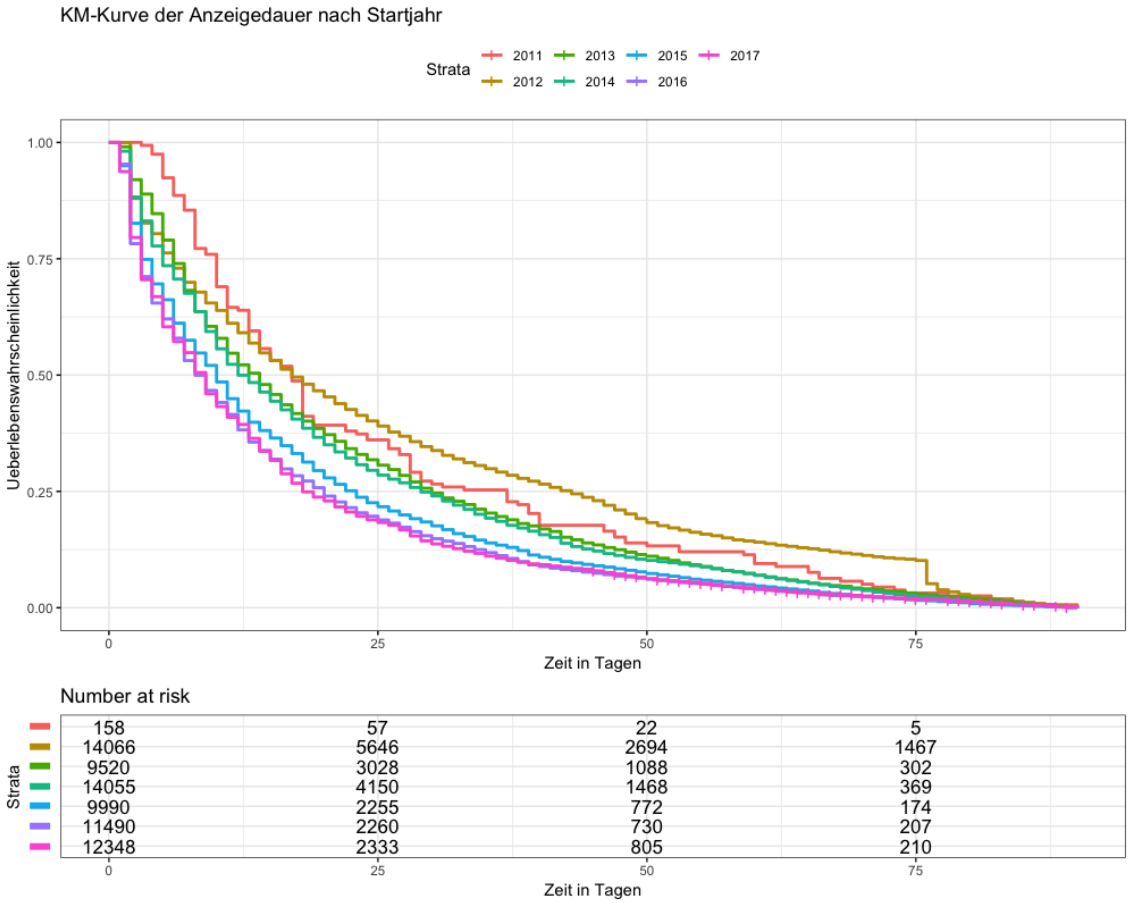

Abbildung 8: Oben: KM-Kurve der Anzeigedauer aufgeteilt auf die Jahre in denen die Anzeige das erste Mal online gestellt wurde. Unten: Absolute Anzahl der Anzeigen unter Risiko

5.2 AFT der Anzeigedauer

In diesem Abschnitt wird das AFT-Modell für die Anzeigedauer aufgebaut, erklärt und analysiert.

5.2.1 Modellannahmen und Struktur

Verteilungsannahme Da man bei einer AFT-Modellierung eine Verteilungsannahme treffen muss, wurde dies anhand der verschieden Kriterien durchgeführt. Alle drei

vorgestellten Informationskriterien lagen bei der Annahme einer Weibull oder logN-Verteilung am niedrigsten. Da die logN-Verteilung leicht niedrigere Werte liefert und später vor allem die Mittelwerte von Interesse sind, wird im folgenden immer eine logN-Verteilung angenommen. Daher müssen für das Modell $\hat{\mu}$ und $\hat{\sigma}$ geschätzt werden.

Log-Transformierte Modell Das schrittweise Verfahren zur Variablenselektion aus Kapitel 3.5.2 anhand des AIC wurde mit allen sinnvoll verwendbaren Variablen durchgeführt. Das Verfahren ergab in beiden Fällen dasselbe Modell mit in Tabelle 3 gelisteten Kovariablen und Einflussgrößen. Es gilt für das Modell:

$$ln(T) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + bZ$$

X_i	$\hat{\boldsymbol{\beta}}$	$exp(-\hat{\beta}_i)$	Signifikanz	Ref. Kategorie
Intercept	$2,4944$	$0,0825$	***	
Startzeit in Jahren	$-0,1445$	$1,1555$	***	
Mittlere Anzahl Anzeigen	$0,0004$	$0,9996$	***	
„2 - 2,5“ Zimmer	$0,0715$	$0,9310$	***	„1 - 1,5“
„3 - 3,5“ Zimmer	$0,0215$	$0,9788$		„1 - 1,5“
„4 - 4,5“ Zimmer	$-0,1223$	$1,1301$	***	„1 - 1,5“
„5 - 6,5“ Zimmer	$-0,3969$	$1,4872$	***	„1 - 1,5“
„> 7“ Zimmer	$-1,0642$	$2,8986$	***	„1 - 1,5“
€/m^2	$0,0869$	$0,9168$	***	
Fläche in m^2	$0,0112$	$0,9889$	***	
€/m^2 * Fläche	$-0,0003$	$1,0003$	***	
Parkplatz vorhanden	$0,0878$	$0,9160$	***	
Balkon oder Terrasse	$-0,0706$	$1,0732$	***	
Einbauküche	$-0,0706$	$1,0731$	***	
Abstellraum	$0,0582$	$0,9435$	***	
Renoviert Keine Angabe	$-0,1736$	$1,1896$	*	Nicht Renoviert
Renoviert	$-0,1360$	$1,1457$	∘	Nicht Renoviert
Bad mit Fenster	$0,0702$	$0,9322$	***	
Bad mit Badewanne	$0,0584$	$0,9433$	***	
Bad mit Dusche	$0,0531$	$0,9483$	***	
Hauswirtschaftsraum	$0,0231$	$0,9771$	*	
Barrierefrei	$0,1093$	$0,8964$	***	
Signifikanz Niveau	$< 0,001$	$< 0,01$	$< 0,05$	$< 0,1$
	***	**	*	∘

Tabelle 3: Kovariablen, deren Einflüsse auf den linearen Prädikator, den prozentualen Einfluss, das Signifikanzniveau und, falls vorhanden, Referenzkategorie des vollen nach AIC optimierten Modells. Ausgenommen: Einfluss der Postleitzahlgebiete und der Etage.

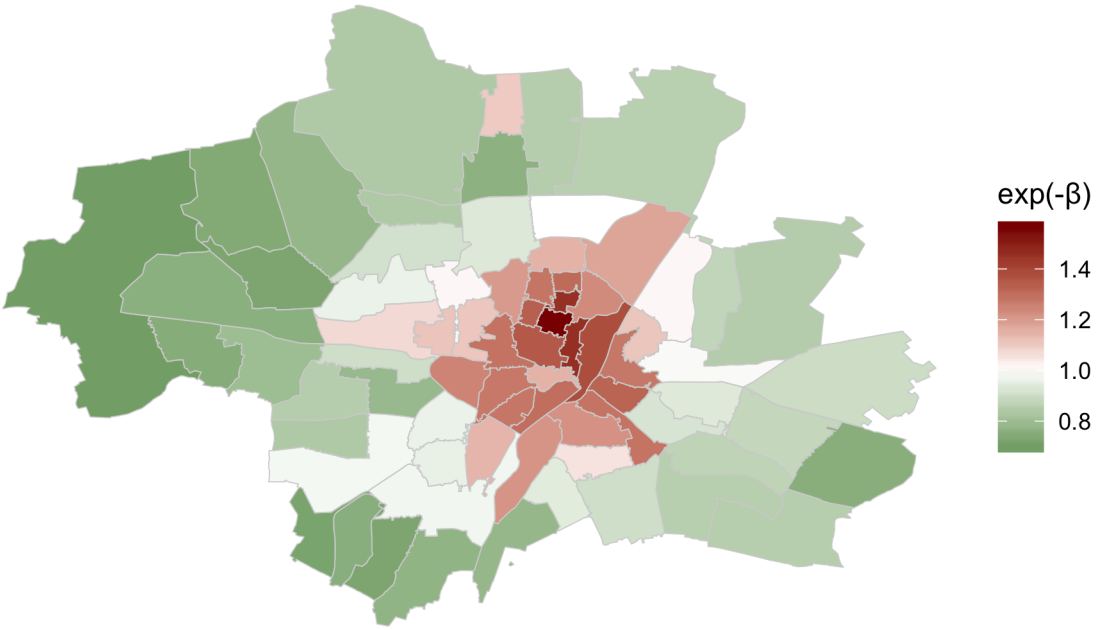

Abbildung 9: Multiplikativer Einfluss der geschätzten Beta-Koeffizienten der Postleitzahlgebiete auf den Verzögerungsfaktor im AFT-Modell. Referenzkategorie: PLZ „80992“.

Der Intercept entspricht dabei $\hat{\beta}_0 = -0,322$ und $b = 1,0477$. Der AIC dieses Modells beträgt 472 351. Eine Überlebenszeitanalyse ohne Einflüsse ergibt einen AIC von 564 534. Es ist also eine deutliche Verbesserung der Vorhersagegenauigkeit eingetreten. Die mittlere Anzahl der Anzeigen, das Startjahr, der Preis pro Quadratmeter und die Fläche in Quadratmetern ist jeweils vom Mittelwert der Variablen aus geschätzt. Dies hat zwar keinen Einfluss auf die $\hat{\beta}$-Koeffizienten, wirkt aber auf den Intercept $\hat{\beta}_0$.

Abbildung 11 zeigt die Survial- und Hazard-Funktion für das Intercept-Modell.

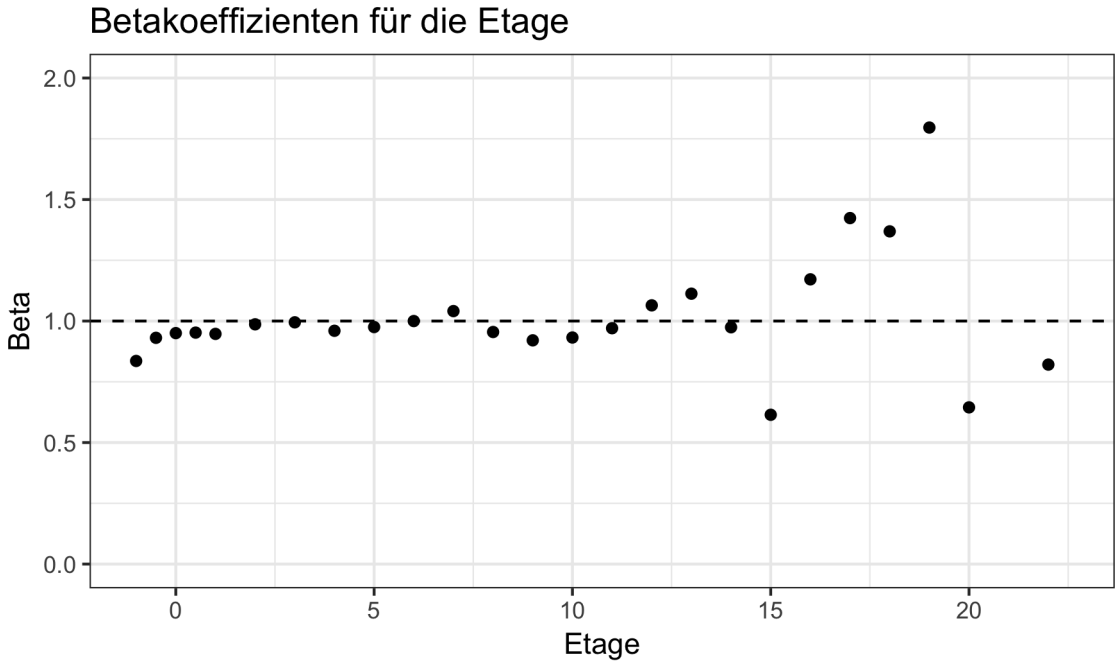

Abbildung 10: Multiplikativer Einfluss der geschätzten Beta-Koeffizienten der Etage auf den Verzögerungsfaktor im AFT-Modell. Referenzkategorie: 6. Stock

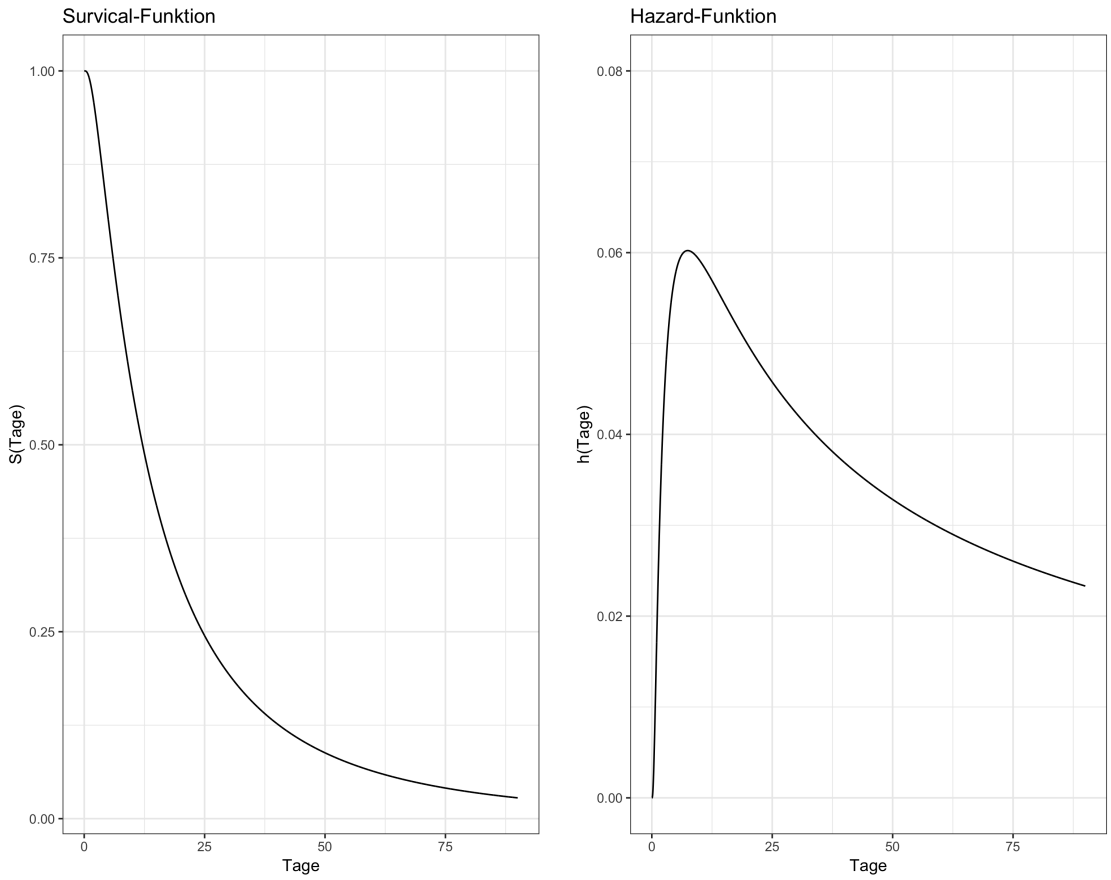

Abbildung 11: Geschätzte Survival- und Hazard-Funktion für das Intercept-Modell

5.2.2 Interpretation

Generell lässt sich der multiplikative Effekt der Variablen auf φ und somit auf die Überlebenszeit aus der Tabelle 3 und Abbildungen 9 und 10 ablesen. Jedoch wird im Folgendem kurz auf die Einflüsse von hohem Interesse eingegangen.

Startdatum Die Auswirkung des Startdatums auf die Anzeigedauer ist Hauptbestandteil der Arbeit. Das AFT zeigt eindeutig, dass das Startdatum der Anzeigen einen negativen Effekt auf die Anzeigedauer hat. Mit einem Signifikanzniveau von $\alpha < 0,1\%$ und einem β-Koeffizienten von $-0,1445$ ist dieser Effekt sehr eindeutig. Somit wirkt auf $\hat{\varphi}$ ein multiplikativer Faktor von $exp(-\hat{\beta}) = 1,1555$ pro Jahr, bei gleichbleibenden sonstigen Kovariablen.

$$\hat{S}_{+1\,Jahr}(t_i) = \hat{S}_0(1,1555\ t_i)$$

Über das gesamte betrachtete Intervall von 6 Jahren, verändert sich also die Anzeigezeit um den multiplikativen Faktor $exp(6 * -(-0,145)) = 1,155^6 = 2,374$, bei gleichbleibenden sonstigen Kovariablen.

$$\hat{S}_{Dez\,2018}(t_i) = \hat{S}_{Jan\,2012}(2,374\ t_i)$$

Man kann also sagen, dass der die Überlebenswahrscheinlichkeit für Wohnungen, bei gleichbleibenden sonstigen Koeffizienten, im Dezember 2018 nach einem Tag dieselbe ist, wie in Januar 2012 nach ca. $2,374$ Tagen. Dieser multiplikative Effekt ist in Abbildung 12 für verschiedene Jahresabständen zwischen zwei Wohnungsanzeigen mit sonstigen gleichen Kovariablen dargestellt.

In Abbildung 13 sind diese Effekte auch deutlich zu sehen. Die erwartete maximale Sterbewahrscheinlichkeit liegt beim Intercept-Modell bei ca. 6% am Tag $7,45$, ein Jahr später schon bei ca. 7% am Tag $6,4$ und sechs Jahre später sogar bei ca. 15% am Tag 3. Entsprechend verhält sich die Survival-Funktion.

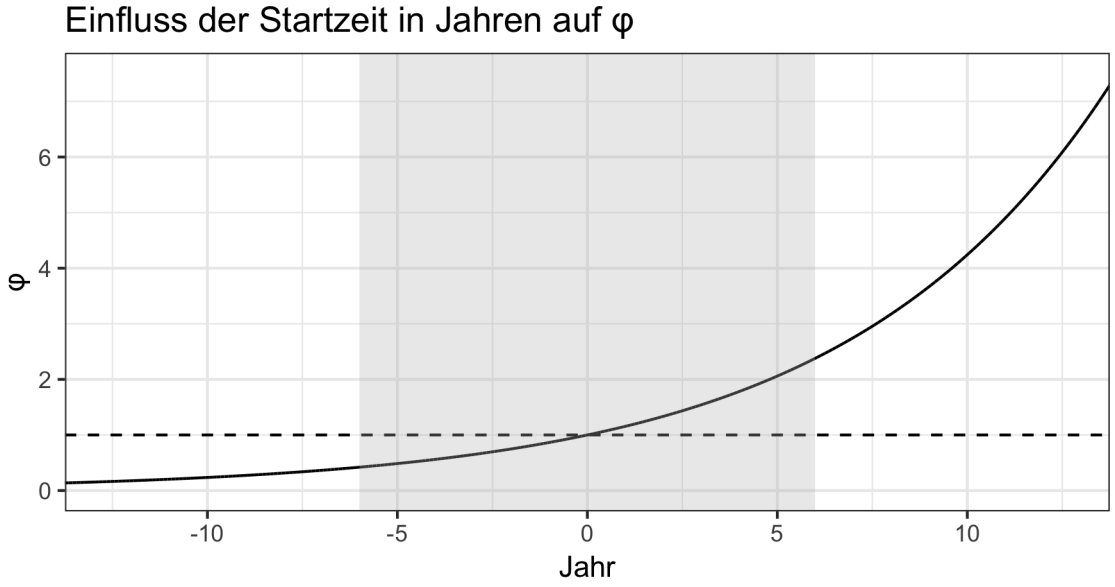

Abbildung 12: Einfluss der Startzeit in Jahren auf den Beschleunigungsfaktor, der graue Kasten markiert den maximal gemessenen Bereich.

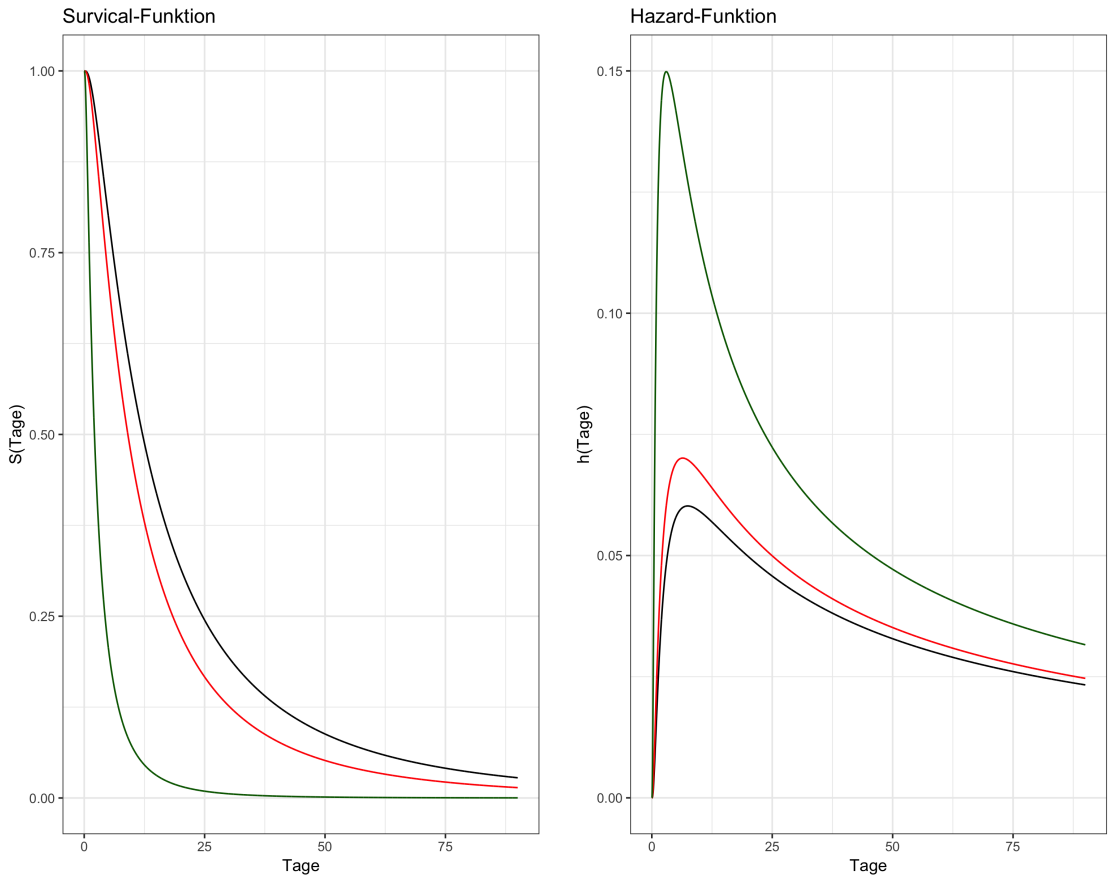

Abbildung 13: Geschätzte Survival- und Hazard-Funktion für das Intercept-Modell in schwarz, in rot entsprechend ein Jahr und in grün 6 Jahre später

Einfluss des Postleitzahlengebiets In Abbildung 9 ist der geschätzte multiplikative Einfluss der Postleitzahlgebiete auf den Beschleunigungsfaktor $\hat{\varphi}$ gezeigt. Als Referenz-Kategorie gilt das Postleitzahlgebiet „80992“ (Teile von „Moosach“ und „Pasing-Obermenzing“) im Norden der Münchener Innenstadt. Deutlich ist zu sehen, dass im Kern der Stadt, der multiplikative Einfluss den Beschleunigungsfaktor vergrößert und am Rand der Stadt, dieser verringert wird. Mit Ausnahme von „80933“ („Feldmoching - Hasenbergel“ und „Milbertshofen - Am Hart“) im Norden von München, mit $exp(-\hat{\beta}_{80933}) = 1,104$. Dieses Postleitzahlgebiet ist jedoch schon bei der deskriptiven Analyse in Abbildung 3 aufgefallen, da verhältnismäßig wenige Wohnungen dort verfügbar sind. Es schient, dass das verminderte Angebot, im Vergleich zu den umgebenden Postleitzahlgebieten, eine erhöhte Nachfrage verursacht, die sich in der Anzeigedauer niederschlägt.

Im Vergleich zur Referenzkategorie sind im Postleitzahlgebiet „80799“ („Maxvorstadt“, „Ludwigvorstadt-Isarvorstadt“ und „Schwabing-Freimann“) Wohnungsan-

zeigen im Schnitt1, 5643 mal kürzer verfügbar, wenn alle anderen Einflussvariablen gleich bleiben. Das Gegenteil dazu stellt „81249“ („Allach-Untermenzing “ „Aubing-Lochhausen-Langwied“ und „Pasing-Obermenzing“) mit einem multiplikativen Einfluss von durchschnittlich 0, 6973 bei gleichen sonstigen Einflussvariablen. Betrachtet man diese Stadtteile im Vergleich zueinander, so sieht man, dass eine Wohnung mit denselben Spezifikationen im Schnitt 2, 2434 mal kürzer in „81249“ angeboten wird, als in „80799“.

$$\hat{S}_{80799}(t_i) = \hat{S}_{81249}(2,2434\ t_i)$$

Dies bedeutet, dass eine Anzeige, bei unveränderten anderen Kovariable, in „81249“ im Schnitt dieselbe Überlebenswahrscheinlichkeit nach 2, 2434 Tagen hat, wie eine Anzeige im Postleitzahlgebiet „80799“ nach nur einem Tag.

Einfluss der Differenz der mittleren Anzahl der Anzeigen Die mittlere Anzahl der Anzeigen während eine Anzeige online ist, hat ebenfalls einen signifikanten Einfluss auf den Beschleunigungsfaktor φ, mit $\alpha < 0,1\%$. In Abbildung 14 kann der multiplikative Effekt auf φ der Differenz der Anzahl der Anzeigen zwischen zwei Beobachtungen, bei gleichbleibenden anderen Kovariablen, abgelesen werden.

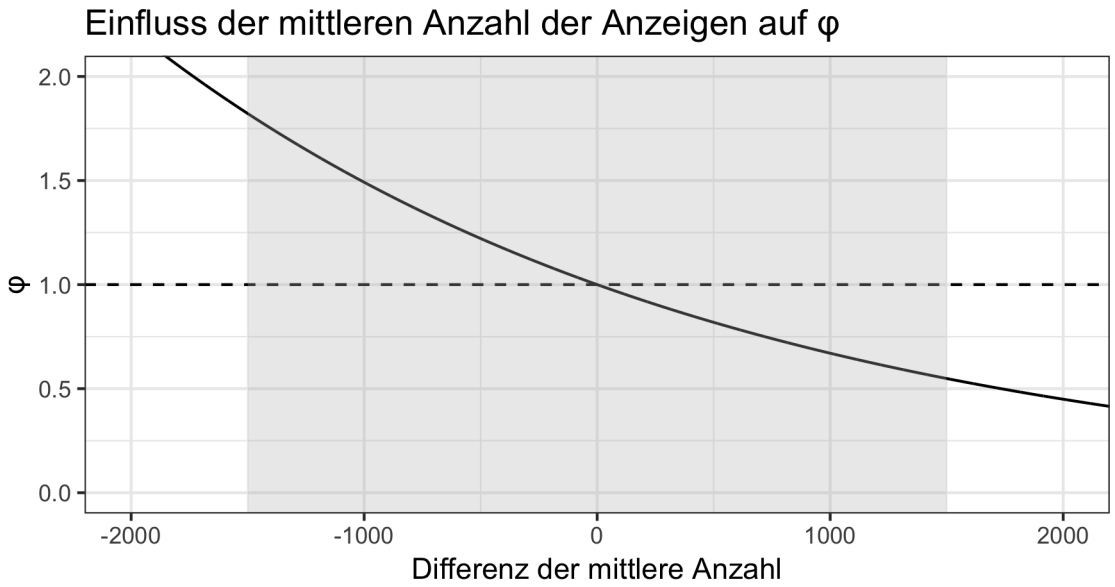

Abbildung 14: Einfluss der Startzeit in Jahren auf den Beschleunigungsfaktor, der graue Kasten markiert den maximal gemessenen Bereich.

Die maximale Differenz zwischen zwei Anzeigen beträgt 1498. Somit ist der multiplikative Faktor auf φ maximal 0, 55 beziehungsweise der Kehrwehr 1, 82 für eine

negative Differenz, in dem Fall, dass alle anderen Einflussvariablen gleich bleiben, für die Differenz von 1498.

6 Abschließende Bemerkungen

6.1 Zusammenfassung

Ziel dieser Arbeit war es zu untersuchen, ob sich der Mietwohnungsmarkt in München verändert hat, indem die Anzeigedauer der Anzeigen untersucht wurde. Nach einer ausführlichen Einführung in die Daten in Kapitel 2, wurde in Kapitel 3 und 4 detailliert auf die verwendeten Accelerated Failure Time Modelle und der darin enthaltenen Regression eingegangen.

Nach dem Methodikteil wird zuerst in Abschnitt 5.1 eine Kaplan-Meier Kurve der Anzeigedauer geschätzt. Auch wird ein KM-Schätzer für die Anzeigedauer unterteilt nach dem Startjahr der Anzeige geschätzt. Dort ist zu sehen, dass die Kurven je nach Startjahr voneinander abweichen. Dieser Effekt wird im Anschluss im Abschnitt 5.2 durch ein AFT geschätzt. Die Modellannahmen und Koeffizienten werden kurz vorgestellt und im Anschluss werden die wichtigsten Koeffizienten interpretiert. Es konnte mit einem signifikanten Effekt (Signifikanzniveau $\alpha < 0,001$) nachgewiesen werden, dass das Startdatum einen Effekt auf die Anzeigedauer hat. Je später eine Anzeige aufgegeben wird, desto schneller wird sie auch wieder entfernt. Auch konnte gezeigt werden, dass der Stadtteil einen Effekt auf die Anzeigedauer hat, so wie die mittlere Anzahl der Anzeigen, die gleichzeitig zur Anzeige online sind.

6.2 Ausblick

Der Effekt auf die Anzeigedauer konnte nachgewiesen und quantifiziert werden. Jedoch könnten hier noch weitere Effekte interagieren. Nicht betrachtet wurde zum Beispiel die Jahreszeit in der die Anzeige geschaltet wurde. So könnte zum Beispiel Semester Anfang und Ende der Hochschulen Münchens mit dem Startdatum interagieren.

Auch wurde im Abschnitt 2.3.2 deutlich, dass viele Wohnungen nur ein oder zweimal vom Web-Scraper erfasst wurden. Daher ist anzunehmen, dass weitere Wohnungen gar nicht erfasst wurden, da sie weniger als 24 Stunden verfügbar waren. Hier könnte eine gezielte Datenerhebung, die in kürzeren Abständen die Webseite analysiert, genauere Ergebnisse liefern.

Außerdem konnte nicht geklärt werden, woher die Auffälligkeiten in der Abbildung 4 stammen. Daher sollte auch hier untersucht werden, warum diese plötzlichen Anstiege in der Anzahl der verfügbaren Wohnungsanzeigen kommen und wie dieser Grund sich auf die Quantifizierung auswirkt.

7 Anhang

7.1 Literatur-, Abbildungs- und Tabellenverzeichnis

Literatur

Bradburn, M. J., Clark, T. G., Love, S. B. and Altman, D. G. (2003), 'Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods', *British journal of cancer* **89**(3), 431–436.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003), 'Survival analysis part i: basic concepts and first analyses', *British journal of cancer* **89**(2), 232–238.

Fahrmeir, L., Kneib, T. and Lang, S. (2007), *Regression: Modelle, Methoden und Anwendungen*, Statistik und ihre Anwendungen, Springer, Berlin and Heidelberg.
URL: *http://nbn-resolving.de/urn:nbn:de:1111-200708029289*

Fahrmeir, L., Künstler, R., Pigeot, I. and Tutz, G. (2007), *Statistik: Der Weg zur Datenanalyse*, Springer-Lehrbuch, 6., überarb. aufl. edn, Springer, Berlin.
URL: *http://dx.doi.org/10.1007/978-3-540-69739-8*

Glomb, P. (2007), Statistische Modelle und Methoden in der Analyse von Lebenszeitdaten, Diplomarbeit, Carl von Ossietzky Universität, Oldenburg.
URL: *https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/Ingenieur/Mueller/Diplomarbeiten/Glomb.pdf*

Immobilien Scout GmbH (n.d.), 'Anzeigenpreise für vermietende eigentümer'.
URL: *https://www.immobilienscout24.de/anbieten/private-anbieter/lp/preise-vm.html*

Immobilienverband IVD Bundesverband e.V. (2018), 'Nutzung von immobilienportalen'.
URL: *https://ivd.net/wp-content/uploads/2018/02/Auswertung-Minutenumfrage-.pdf?utm_source=2018-01+Mitglieder&utm_campaign=a80bff7ab6-EMAIL_CAMPAIGN_2018_02_21&utm_medium=email&utm_term=0_a62d41033e-a80bff7ab6-57078349*

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (2014), *Handbook of Survival Analysis*, Chapman & Hall / CRC Handbooks of Modern

Statistical Methods, Taylor and Francis, Hoboken.
URL: *http://gbv.eblib.com/patron/FullRecord.aspx?p=1563126*

Kleinbaum, D. G. and Klein, M. (2005), *Survival Analysis: A Self-Learning Text*, Statistics for Biology and Health, second edition edn, Springer Science+Business Media Inc, New York, NY.
URL: *http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10229012*

Klösener, K.-H., Elpelt, B. and Hartung, J. (2002), *Statistik: Lehr- und Handbuch der angewandten Statistik*, 13. aufl., 13., unwesentl. veränd. aufl. reprint 2014 edn, De Gruyter Oldenbourg, München.
URL: *http://www.degruyter.com/search?f_0=isbnissn&q_0=9783486810585&searchTitles=true*

Lee, E. T. and Wang, J. W. (2003), *Statistical methods for survival data analysis*, Wiley series in probability and statistics, 3. ed. edn, Wiley-Interscience, Hoboken, NJ.
URL: *http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10304419*

Portal München Betriebs-GmbH & Co. KG (n.d.), 'Stadt kauft 300 wohnungen in sendling'.
URL: *https://www.muenchen.de/aktuell/2018-11/muenchen-kauft-wohnungen-sendling.html*

Sarego GmbH (2017), '5 berühmte zitate für ihre perfekte investmentstrategie - sarego'.
URL: *http://sarego.de/blog/5-beruhmte-zitate-fur-ihre-perfekte-investmentstrategie/*

Schrenk, M. (2012), *Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL*, 2nd ed. edn, No Starch Press, San Francisco.
URL: *http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10574793*

Scout24 AG (n.d.), 'Marken'.
URL: *https://www.scout24.com/Marken/marken-intro.aspx*

Sven Heinen (02.11.2018), 'Marktreport münchen 2017: Noch läuft's - bellevue'.
URL: *https://www.bellevue.de/stories-und-ratgeber/deutschland/*

staedte-specials/immobilien-marktreport-muenchen-2017-noch-laeufts.html

Abbildungsverzeichnis

Tabellenverzeichnis

7.2 Inhalt des elektronischen Anhangs

Die beigefügte CD enthält:

- Bachelorarbeit als PDF
- Ordner **Analyse** mit folgendem Inhalt:
  - „ims.RData“- Original zur Verfügung gestellter Datensatz
  - „ims2.RData“ - Bearbeiteter Datensatz „ims.RData“
  - „byDate.RData“- Umstrukturierter Datensatz nach Datum
  - „bd.RData“- Kopie von „byDate.RData“
  - „StepAIC.txt“ - Enthält den R-Output für das schrittweise Verfahren zur Variablenselektion
  - Verschiedene Daten zum erstellen der Postleitzahlplots
- Ordner **Plots** enthält in Unterordnern Grafiken im „.png“-Vormat

7.3 Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum	Tobias Kaller