Seminar „Philosophische Grundlagen der Statistik“
Thema: Statistische Modellierung

Seminararbeit
am Institut für Statistik
Ludwig-Maximilians-Universität

München, im Januar 2011

Fachliche Betreuung: Prof. Dr. Thomas Augustin
Autor: Anne Göpfert

Abstract

„The word 'model' is highly ambiguous, and there is no uniform terminology used by either scientists or philosophers.“
Internet Encyclopedia of Philosophy [4]

Diese Seminararbeit hat als Thema „statistische Modellierung“. Wie aus dem Begriff unschwer zu erahnen ist, ist dies ein sehr weitläufiges und „schwammiges“ Thema, das auf vielfältige Art und Weise interpretiert werden kann. Aus diesem Grund ist die vorliegende Arbeit auch sehr breit gefächert und gibt einen Überblick über verschiedene Aspekte der statistischen Modellierung. Der Fokus liegt dabei stärker auf einem breiten Überblick als auf einer eingehenden Eröterung von spezifischen Themen oder Theorien.
Zunächst werden mit der bayesianischen, der frequentistischen und der fisherschen Sicht drei verschiedene Denkweisen der Statistik vorgestellt und ihre Bedeutung für die statistische Modellierung diskutiert. Konkret mit dem Thema Modellierung beschäftigt sich das nächste Kapitel. Hier wird zunächst ganz allgemein geklärt, was unter dem Begriff „Modell“ zu verstehen ist, aus welchen „Bausteinen“ ein (statistisches) Modell besteht, sowie welche Methoden für Modellwahl und -diagnose existieren. Anschließend wird eine kurze Typisierung von verschiedenen Modellen gegeben. Im Kapitel „beispielhafte Anwendungen“ wird die Verwendung von statistischen Modellen in drei unterschiedlichen wissenschaftlichen Disziplinen erötert. Die Modellierung von Verweildauern auf non-parametrische und auf parametrische Weise wird anhand der Verwendungsmöglichkeiten in der Medizin und in der Ökonometrie, die Verwendung von Prognosemodellen anhand eines Klimaszenarios gezeigt. Im letzten Kapitel wird ein Fazit gezogen, sowie ein Ausblick auf weiterführende Themen gegeben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ANOVA Analysis Of Variance

AIC Informationskriterium nach Akaike

ARMA Autoregressiv Moving Average

BIC Bayesianische Informationskriterium

MIT Massachusetts Institute of Technology

1 Überblick: Verschiedene Denkweisen der Statistik

Dieses Kapitel beschäftigt sich noch nicht konkret mit dem Thema „statistische Modellierung“, sondern geht auf unterschiedliche Denkweisen in der Statistik ein und stellt Gemeinsamkeiten und Unterschiede der verschiedenen Inferenzschulen dar. Wie sich jedoch zeigen wird, ist die Kenntnis dieser Inferenzschulen für eine überlegte Vorgehensweise in der Modellierung eine wichtige Voraussetzung und wird deshalb an dieser Stelle noch einmal kurz dargestellt.
In der Statistik stehen sich mit den Frequentisten und den Bayesianern zwei unterschiedliche Denkweisen gegenüber. Efron [15, 16] und Dempster [14] nennen noch die Anhänger Fishers als eigene kleinere Gruppe. In den folgenden Abschnitten werden die Philosophien dieser drei unterschiedlichen Gruppen jeweils kurz dargestellt und anschließend ihre Bedeutung für die statistische Modellierung erötert.

1.1 Frequentistische Inferenzschule

In diesem Abschnitt wird auf die frequentistische Inferenzschule eingegangen. Wichtige Begriffe sind dabei „Erwartungstreue“ und „Konsistenz“.
Ein Schätzer wird als *erwartungstreu* bezeichnet, falls gilt [22]:

$$E_\vartheta(T_n(X)) = \vartheta \tag{1.1}$$

Man kann sagen, dass „im Mittel“ der richtige Wert getroffen wird. Trifft dies nicht zu, so bezeichnet man den Schätzer auch als verzerrt [22]. Eine etwas schwächere Forderung als Erwartungstreue ist die *asymptotische Erwartungstreue*. Von asymptotischer Erwartungstreue spricht man, wenn ein Schätzer mit wachsendem Stichprobenumfang $n \to \infty$ gegen den wahren Parameter ϑ konvergiert, für einen festen Stichprobenumfang n muss der Schätzer hier nicht erwartungstreu sein [22]. Eine weitere Eigenschaft, die häufig in Zusammenhang mit dem frequentistischen Wahrscheinlichkeitsbegriff genannt wird, ist die *Konsistenz* eines Schätzers. Man bezeichnet einen Schätzer als schwach konsistent, falls die Zufallsvariable $T(X)$ nach unendlich vielen Wiederholungen eines Zufallsexperiments mit Wahrscheinlichkeit gegen ϑ konvergiert und als stark konsistent, falls er fast sicher gegen ϑ konvergiert. Damit steht der Begriff der Konsistenz im engen Bezug zur asymtotischen Erwartungstreue [22].
Zentral beim frequentistischen Wahrscheinlichkeitsbegriff sind die (zumindest theoretisch) unendlich vielen Wiederholungen ein und desselben Zufallsexperiments. Dies rückt die Bedeutung von Experimenten in den Vordergrund. Frequentisten fokusieren sich damit, im

Vergleich zu Bayesianern, stärker auf die Planung von Experimenten und der Theorie von Modellen [10, 27]. Lehmann [27] weist allerdings auch darauf hin, dass ein gutes erklärendes Modell verlangt, dass der Anwender sich mit der dahinter liegenden Wissenschaft auseinander gesetzt hat. Hier sieht Dempster [14] ein Problem bei den frequentistischen Methoden, da durch alleinige Konzentration auf die mathematische Theorie, praktische Implikationen (wie z.B. bestehendes Vorwissen) ignoriert werden können.

1.2 Bayesianische Inferenzschule

Die bayesianische Sichtweise geht, im Gegensatz zu der frequentistischen Sichtweise, davon aus, dass bereits bestehendes Vorwissen auch schon in die Schätzung von Wahrscheinlichkeiten eingebracht werden kann und muss. So wird z.B. die Wahrscheinlichkeit, dass die Linkspartei einen hohen Prozentsatz bei einer Landtagswahl erhält, je nach Bundesland unterschiedlich sein. Technisch wird der unbekannte Parameter ϑ nicht mehr als fest (wie bei den Frequentisten), sondern als Zufallsvariable mit Mittelwert m und Standardabweichung s angesehen [15]. Die Verteilung für diese „Zufallsvariable“ bezeichnet man als *Priori-Verteilung* $f(\vartheta)$. Nach Beobachtung der Daten $X = x$ erhält man die *Posteriori-Verteilung* $f(\vartheta|x)$, die auf die Beobachtungen bedingt ist [22]. Dabei können unterschiedliche Priori-Verteilungen gewählt werden. Eine Möglichkeit ist, für den Priori-Parameter eine Gleichverteilung zu wählen. Umstritten ist die Zielsetzung mit der eine solche Gleichverteilung gewählt wird. So wird z.B. dieser Ansatz häufig verwendet um, „wenig Vorwissen“ in die Priori-Verteilung eingehen zu lassen [22]. Diskussionen, auch in diesem Seminar zu den Philosophischen Grundlagen der Statistik, weisen aber auch darauf hin, dass eine Gleichverteilung durchaus ein bestimmtes Vorwissen beinhaltet, das auf diese Weise in die Modellierung eingebracht wird. Eine andere, pragmatischere Vorgehensweise ist die Wahl einer konjugierten Priori-Verteilung. In diesem Fällen gehört die Posteriori-Verteilung zur gleichen Verteilungsklasse wie die Priori-Verteilung (z.B. ist die Beta-Verteilung konjugiert zur Binomal-Verteilung [22]). Abhandlungen über verschiedene Wahlmöglichkeiten für die Priori-Verteilung findet man bei Held [22] oder Bayarri und Berger [10]. Die konkrete Wahl für eine Priori-Verteilung wird häufig von individuellen Einstellungen und Ansichten beeinflusst. Bayarri und Berger [10] weisen allerdings darauf hin, dass die Anwendung von bayesianischen Methoden nicht immer etwas mit der „inneren Einstellung“ zu tun haben muss. Auch die fortschreitenden Entwicklung computerintensiver Methoden (z.B. Gibbs Sampling oder Markov Chain Monte Carlo) führt zu einem verstärkten Interesse für Bayes. Die Verwendung von bayesianischen Methoden und damit die Herausforderung die richtige Priori-Verteilung zu wählen stellt den Anwender vor eine Reihe von Herausforderung. Efron [16] nennt hier konkret:

- Schwierigere Anwendbarkeit in der Praxis. Man muss sich im konkreten Fall jeweils Gedanken machen, welche Priori-Verteilung gewählt wird
- Wissenschaftliche Objektivität lässt sich mit Bayes schwerer erreichen
- Vor allem in Situationen, in denen mehr als ein Parameter geschätzt wird, ist es sehr

schwer, für jeden Parameter die richtige Priori-Verteilung zu finden

- Bayes beschäftigt sich hauptsächlich mit Inferenz. Dies ist zwar die „Königsdisziplin“ in der Statistik, aber nicht notwendigerweise die wichtigste. Bei Neymann und Fisher erhalten vorgelagerte Aspekte der statistischen Modellierung (wie z.B. Randomisierung und Planung von Experimenten) ein stärkeres Gewicht.

1.3 Sichtweise von Sir Ronald Fisher als 3. Denkrichtung

In diesem Abschnitt wird die Fiduzialinferenz und die Sichtweise von Sir Ronald Fisher vorgestellt. Diese Sichtweise wird von einigen Autoren (z.B. Efron [15, 16] oder Dempster [14]) als 3. Denkrichtung in der Statistik angesehen. Fisher [19] definiert Wahrscheinlichkeit als einen gut spezifizierten Status von „logischer Unsicherheit“. Er listet drei Anforderungen auf, die dafür benötigt werden:

1. Es gibt ein messbares Referenzset (z.B. Geschlecht eines Babys bei einer Geburt)
2. Das interessierte Subjekt gehört zu diesem Referenzset
3. Nicht relevante Sub-sets können identifiziert werden (z.B. könnte das Geschlechterverhältnis bei erstgeborenen Babies anders sein)

Dempster [14] geht in seinem Artikel „Logical Statistics I. Modells and Modelling“ auf die Unterschiede zwischen den Denkweisen des Frequentisten Neyman und Sir Ronald Fisher ein. Formal gesehen sind sich die Vorgehensweisen, die z.B. beim Schätzen eines Parameters angewendet werden, sehr ähnlich. Der Unterschied zwischen beiden liegt in der „Vorstellung“, wie eine Vorgehensweise gewählt wird. Fisher bevorzugt die oben erläuterte Form von „logischer“ Inferenz, ohne die zwangsläufige Berücksichtigung des „long run“ (also den theoretisch unendlich vielen Wiederholungen). Dabei schließt er die klassischen Eigenschaften frequentistischer Schätzer nicht zwangsläufig aus [15]. Nach Dempster [14] besteht der Hauptunterschied zwischen Fisher und Neyman, in der Tatsache, dass Fisher, im Gegensatz zu Neymann, auch subjektive Wahrscheinlichkeiten zulässt. In der Fiduzialinferenz von Fisher hängt die Beobachtung, und damit das Wissen über den unbekannten Parameter ϑ, von den beobachteten Daten x ab. ϑ bleibt weiterhin fest, die Schätzung des Parameters wird jedoch von der jeweiligen Stichprobe beeinflusst. Die Rollen der Zufallsvariable x und des Parameters ϑ sind damit vertauscht [12, 14]. Inwieweit kann Fisher nun als Bayesianer angesehen werden? Fisher [19] ist der Auffassung, dass eine bayesianische Vorgehensweise immer dann angemessen ist, wenn Wissen über die Priori-Verteilung eines Parameters gegeben ist. Dennoch ist er der Auffassung, dass es wichtig ist, alternative Formen der Inferenz zu entwickeln [14]. Nach Dempser [14] kann Fisher deshalb als Frequentist oder Bayesianer in „a limited way“ angesehen werden.

1.4 Bedeutung für die statistische Modellierung

In diesem Kapitel wurden die verschiedenen Denkrichtungen in der Statistik dargestellt. Im Folgenden werden sie noch einmal kurz zusammengefasst und Bedeutung für die statistische

Modellierung diskutiert. Abbildung 1.1 fasst die Gemeinsamkeiten und Unterschiede der frequentistischen, bayesianischen und fisherschen Sicht zusammen.

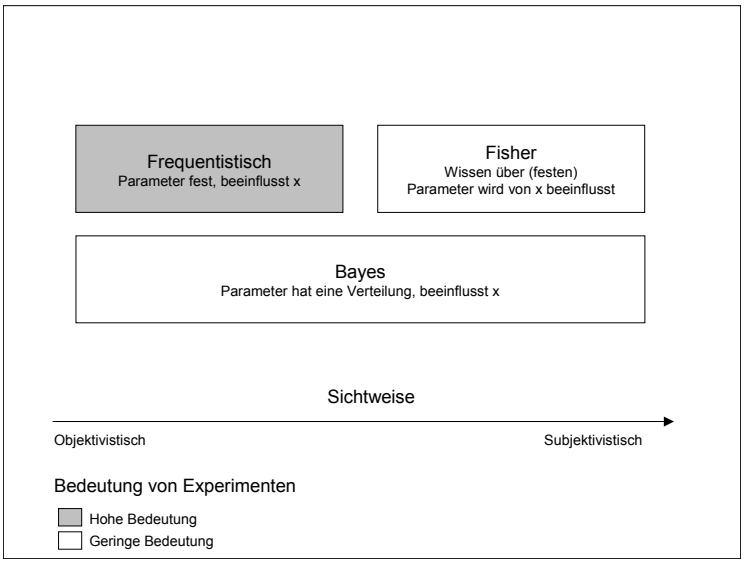

Abbildung 1.1: Vergleich der drei verschiedenen Denkweisen in der Statistik

Insgesamt lässt sich keine eindeutige Empfehlungen für die Bevorzugung einer „Denkrichtung“ wie frequentistische, bayesianische oder fishersche Sicht geben. Häufig werden verschiedene Methoden nebeneinander verwendet. So weisen Bayarri und Berger [10] darauf hin, dass Baysianer z.B. für die Modellwahl frequentistische Methoden verwenden (siehe auch Abschnitt 2.3.2).
In der Einleitung zu diesem Kapitel wurde schon einmal erwähnt, dass die Kenntnis dieser Denkweisen bzw. Inferenzschulen wichtig für die nachfolgenden Kapitel zur Modellierung ist. Auf den ersten Blick erscheint dies zunächst nicht so: Im Kapitel 2 „statistische Modellierung“ werden Themen wie „Bausteine eines statistischen Modells“ oder Gütekriterien zur richtigen Modellwahl erötert. Das Kapitel 3 „Beispielhafte Anwendungen“ stellt Modellierungsbeispiele aus unterschiedlichen wissenschaftlichen Disziplinen dar. Wie hängen nun also diese Themen mit der „Entscheidung“ für eine Inferenzschule zusammen? Die Wahl eines bestimmten Modells wird davon bestimmt was man modellieren möchte (ausführliche Erläuterung siehe Abschnitt 2.1) und mit welcher Vorgehensweise dies geschieht. Diese Vorgehensweise wird von unterschiedlichen Aspekten beeinflusst. Der Abschnitt 3.1 „Unterschiedliche Modellierungen von Verweildauern“ stellt beispielweise zwei Arten der Modellierung von Verweildauern in zwei wissenschaftlichen Disziplinen dar. Nicht zuletzt wird die Vorgehensweise aber auch der Inferenzschule, der ein Anwender anhängt, maßgeblich beeinflusst (z.B. Annahme einer bestimmten Priori-Verteilung, siehe Abschnitt 1.2). Damit hat die Kenntnis der verschiedenen Inferenzschulen für die statistische Modellierung eine große Bedeutung.

2 Statistische Modellierung

In diesem Kapitel wird zunächst geklärt, was unter dem Begriff „Modell“ zu verstehen ist und welche Ziele mit einem Modell verfolgt werden. Anschließend werden kurz die „Bausteine“ und deren Zusammenhänge eines (statistischen) Modells dargestellt. Danach werden, teilweise anhand von speziellen Methoden (z.B. lineare Regression), die Themen Modellwahl und Validierung erötert und am Ende des Kapitels wird eine „Typisierung“ von statistischen Modellen vorgenommen.

2.1 Was ist ein „Modell“?

Modelle sind keine neue Erfindung der Statistik oder anderer Disziplinen, es gibt sie schon seit Urzeiten. So weist z.B. Müller [31] darauf hin, dass das Modelle so alt sind wie die Menschheit selbst. Er sieht in der Verwendung von Werkzeugen in der Altsteinzeit bzw. Höhlenmalerei und in Mythen etwa zur Entstehung des Kosmos erste Ansätze von Modelldenken. Der Modellbegriff muss sehr weitläufig interpretiert werden. Puppen und Modelleisenbahnen fallen dabei ebenso darunter wie Stadtpläne und Regressionsgleichungen bei statistischen Modellen. Abbildung 2.1 zeigt beispielhaft ganz unterschiedliche Modellierungen aus verschiedenen Bereichen.

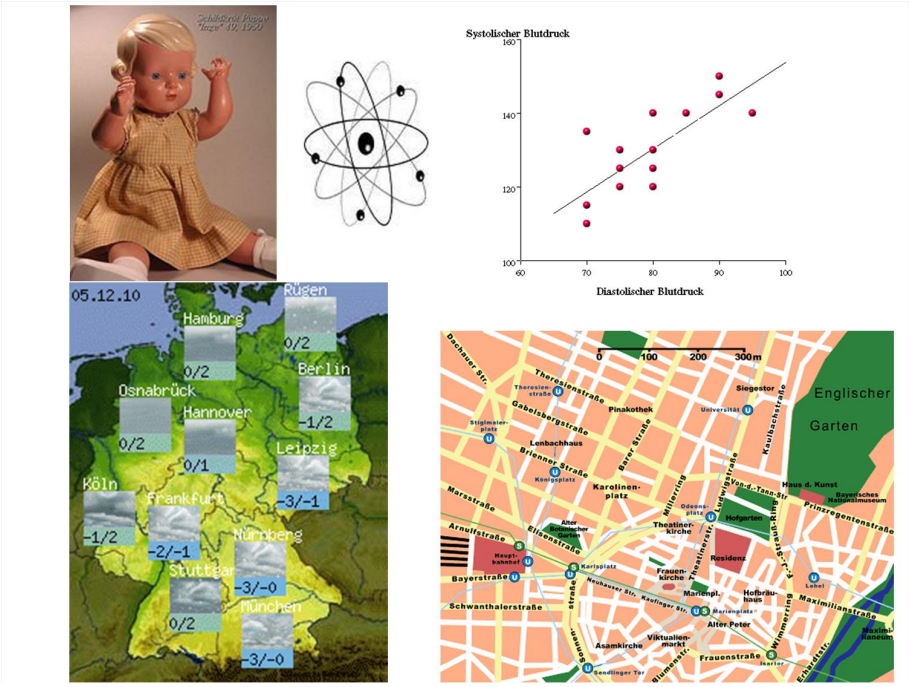

Abbildung 2.1: Beispiele für unterschiedliche Modelle aus verschiedenen Bereichen

Der Begriff „Modell“ kann also keinesfalls als einheitlich und allgemeingültig angesehen werden. Dies zeigt auch folgendes Zitat aus der „Internet Encyclopedia of Philosophy“ [4]: „The word 'model' is highly ambiguous, and there is no uniform terminology used by either scientists or philosophers.“
Alle Modelle haben jedoch eine Sache gemeinsam. Sie repräsentieren eine bestimmte Idee bzw. eine bestimmte Sache und dienen einem bestimmten Zweck. Wie ein spezielles Modell aussieht, hängt dabei immer von seiner Absicht und seinem Anwender ab. So wird sich eine Puppe für einen Sammler von einer Puppe, die zum Spiel und zur Erziehung eines kleinen Kindes gedacht ist, unterscheiden. Modelle lassen immer bestimmte Aspekte weg und stellen andere u.U. falsch dar [24]. Dies ist jedoch nicht problematisch, solange sie ihre „Aufgabe“, eine bestimmte Sache zweckgemäß darzustellen, noch richtig „erfüllen“. Auf dem Pragmatismus bei der Modellbildung verweist auch Stachowiak [30]. Er fordert den Modellbegriff „dreifach pragmatisch“ zu relativieren: Modelle sind „für jemanden“, sie sind „innerhalb eines bestimmten Zeitintervalls“ und für „einen bestimmten Zweck“. Es müssen also immer folgende Fragen beanwortet werden [30]:

- *Für wen* ist das Modell?
- *Wann* brauchen wir dieses Modell?
- *Wozu* brauchen wir dieses Modell?

Eine gute Definition für ein Modell findet man bei Kaplan (S. 6) [24]: „A model is a representation for a particular purpose.“. Modelle bilden also die Realität immer nur in einem bestimmten Maße ab. Wie genau diese Abbildung aussieht, hängt dabei immer von der spezifischen Zielsetzung ab. Die Eigenschaften eines wissenschaftlichen Modells werden bei Pittioni [32] erwähnt:

- Modelle sind Systeme, d.h. ein Beziehungsgefüge
- Modelle beziehen sich auf ein Objekt oder einen Prozess des Originals. Die Art und Weise des Bezugs wird dabei vom Anwender (Modellsubjekt) vorgegeben
- Modelle sind dem Original ähnlich. Der Grad der Ähnlichkeit hängt dabei von den Absichten des Anwenders ab
- Modelle haben praktische Vorteile gegenüber dem Original. Diese Vorteile können z.B. darin bestehen, dass das Modell billiger oder für den Anwender verständlicher ist

Auch dies stimmt überein mit den vorhergehenden Definitionen der Abstraktion und der Zweckgebundenheit eines Modells.
Welche Ziele sind nun mit der Erstellung von Modellen verbunden? Dempster [14] nennt drei Ziele bei der Aufstellung von Modellen.

1. Annäherung bzw. Approximation
2. Erklärung

3. Vorhersage

Je nach Sicht und Einstellung des Anwenders (z.B. Baysianer vs. Frequentisten) können unterschiedliche Ziele in den Vordergrund gestellt werden.
Nach dieser allgemeinen Einführung zum Modellbegriff liegt der Schwerpunkt im Folgenden wieder auf der statistischen Modellierung.

2.2 „Bausteine“ eines (statistischen) Modells

In diesem Abschnitt werden die „Baustandteile“ eines Modells und ihre „Beziehungen“ untereinander dargestellt. Kaplan trifft in seinem Buch „Statistical Modelling“ folgende Aussage [24]: Bestandteile eines Modells sind *Daten*. Diese Daten bestehen aus Objekten (z.B. Personen, die in einer Studie untersucht werden), den *Fällen*. Jeder Fall hat mehrere Attribute mit unterschiedlichen quantitativen oder qualitative Ausprägungen, die *Variablen*. Die Auswahl der Daten und der Variablen muss dem jeweiligen Untersuchungsziel und der Verfügbarkeit der Daten angepasst werden. Dempster [14] fasst dagegen den Begriff „Daten“ weitläufiger. Z.B. sieht er jeden Input in einem Computer, wie etwa ein Programm, als Daten an. Seiner Ansicht nach sind aber nur empirische Daten, Daten im Sinne von statistischen Modellen.
Die Variation in den Variablen ist oft Schwerpunkt der Modellierung, sie kann in zwei Teile, erklärte und unerklärte Variation, aufgeteilt werden. „Typische“ Fälle sind nahe an den Gesamtmittelwert angesiedelt, die Abweichungen der individuellen Fälle vom Gesamtmittelwert sind die Residuen. *Ausreißer* sind dagegen „untypische“ Fälle, die sich in ihrer Ausprägung deutlich von anderen Daten unterscheiden. Barnett [9] nennt vier verschiedene Möglichkeiten mit Ausreißern umzugehen:

1. „Accommodation“: durch die Verwendung von robusten Methoden (z.B. Abweichung vom Median)

2. „Incorporation“: wenn sich der Ausreißer durch die Verteilungsannahme erklären lässt

3. „Identification“: Aus dem Ausreißern selbst kann auch schon eine wichtigte Information gezogen werden (z.B. können ungewöhnlich hohe Einkaufsmengen damit erklärt werden, dass in diesen Fällen keine wöchentlichen Einkäufe, sondern Einkäufe über einen längeren Zeitraum erfolgen). Diese Erkenntnis kann dazu führen, dass ein neues Modell aufgesetzt wird (z.B. Modellierung einer Mischverteilung über die verschiedenen Einkaufstypen)

4. „Rejection“: wenn es keine anderen Alternativen gibt mit dem Ausreißer umzugehen (z.B. offensichtliche Fehleingaben)

Für die *grafische Darstellung* der Variation im Gesamtdatensatz gibt es unterschiedliche Methoden wie z.B. Boxplots oder Histogramme. Aus den Plots lassen sich erste Rückschlüsse auf die Verteilung ziehen. Eine ausführliche Abhandlung zur Variation bei Modellen ist bei Kaplan [24] zu finden.
Wichtig für die statistische Modellierung ist jedoch nicht nur die Kenntnis der Variation

zwischen den Daten, sondern auch die Beziehung zwischen den einzelnen Variablen. Diese wird in einer mathematischen Funktion dargestellt. Die *Response-Variable* oder abhängige Variable ist die Variable, über die man etwas herausfinden will, ihre Variation wird über die restlichen Variablen (*Kovariablen*), die als Input in das Modell kommen, erklärt [24]. Die *Residuen* zeigen für jeden einzelnen Fall wie weit er von seinem theoretischen Modellwert abweicht. Teilt man die Varianz in einem Modell in erklärte und unerklärte Varianz, dann repräsentierten die Residuen den unerklärten oder zufälligen Teil der Varianz. Auch für Dempster [14] bestehen Modelle aus den Bestandteilen „data“ und „laws“ (d.h. mathmatische Funktionen). Wie die „laws“ genau aussehen hängt auch davon ab, ob das Modell deterministisch ist oder nicht. Die Kenntnis über die formalen Bestandteile eines Modells bedeutet noch nicht, dass auch immer die richtigen Variablen und mathematischen Funktionen für das Modell ausgewählt werden. Die Wahl eines „richtigen“ und angemessenen Modells ist jedoch von entscheidender Bedeutung.
In den nächsten Abschnitten geht es daher um die Themen Modellwahl und -diagnose und Validierung.

2.3 Methoden zur Modellwahl

Fahrmeir et al. [18] schlagen folgende Vorgehensweise zur Auswahl eines Modells vor:

1. Substanzwissenschaftliche Überlegungen bzw. Vorwissen: Damit wird eine Vorauswahl potentieller Modelle getroffen. Diese Modelle können sich sowohl in der Anzahl der Variablen als auch in der Art der Modellierung unterscheiden. Insgesamt sollte die Anzahl der Modelle aber relativ klein sein. Auch jeweilige Inferenzschule wird die grundlegende Wahl eines Modells beeinflussen (vergleiche Abschnitt 1.4)

2. Beurteilung anhand von gewählten Modellwahlkriterien

Hastie et al. [21] splitten den letzten Punkt noch in zwei Unterschritte bzw. Ziele auf:

- Modellselektion: Die Performance von verschiedenen Modell wird geschätzt, mit dem Ziel das Beste auszuwählen

- Modellbeurteilung: Schätzung des Vorhersagefehlers auf einem neuen Datensatz

Auf Methoden zum Erreichen dieser Ziele wird im Folgenden ausführlicher eingegangen. Zunächst werden verschiedene Möglichkeiten für das Testen von Hypothesen bei Modellen gezeigt. Anschließend erfolgen zwei Abschnitte zur Beurteilung der Modellgüte und zum Überprüfen der gewählten Annahmen (Modelldiagnose). Im Abschnitt 2.3.4 wird die Kreuzvalidierung vorgestellt und erötert.

2.3.1 Hypothesentests

In diesem Abschnitt werden drei verschiedene Möglichkeiten zum Testen von linearen Hypothesen bei Regressionsmodellen vorgestellt [18]. Diese Hypothesen sind üblicherweise folgendermaßen definiert:

$$H_0 : \mathbf{C\Theta} = \mathbf{d} \tag{2.1}$$

gegen

$$H_1 : \mathbf{C\Theta} \neq \mathbf{d} \tag{2.2}$$

$\mathbf{\Theta}$ ist dann beispielsweise der Parametervektor β für Regressionsmodelle. Die Signifikanz des Einflusses von Kovariabeln kann über einen r-dimensionalen Teilvektor $\mathbf{\Theta_r}$ getestet werden. Mit diesen Voraussetzungen können nun drei verschiedene Tests durchgeführt werden (für eine ausführlichere Darstellung siehe Fahrmeir et al. [18]):

- Der *Likelihood-Quotienten-Test* vergleicht den Maximum-Likelihood-Schätzer unter der Alternativhypothese mit dem Maximum-Likelihood-Schätzer unter der Nullhypothese:

$$LQ = \frac{L(\hat{\mathbf{\Theta}})}{L(\tilde{\mathbf{\Theta}})} \tag{2.3}$$

  Dabei ist $\hat{\mathbf{\Theta}}$ der Schätzer unter der Alternativhypothese H_1 und $\tilde{\mathbf{\Theta}}$ der Schätzer unter der Nullhypothese H_0. Größere Werte für den Likelihood-Quotienten sprechen dafür die Nullhypothese abzulehnen. Der Likelihood-Quotienten-Test hat als voraussetzung, dass genestete Modelle vorliegen.

- Bei der *Wald-Statistik* wird die Differenz zwischen $\mathbf{C\hat{\Theta}}$ und $\mathbf{d}$ gemessen und mit der inversen Kovarianzmatrix gewichtet. Auch hier weisen größere Werte auf eine Ablehnung der Nullhypothese hin.

- Bei der *Score-Statistik* wird die gewichtete Distanz zwischem dem Wert $\mathbf{0} = s(\hat{\mathbf{\Theta}})$ der Score-Funktion und dem Wert $s(\tilde{\mathbf{\Theta}})$ ausgewertet am restringierten Schätzer $\tilde{\mathbf{\Theta}}$.

Die drei Teststatistiken sind unter H_0 approximativ χ^2-verteilt mit r Freiheitsgraden. Für die Auswahl eines geeigneten Tests sollte auf bereits geschätzte Modelle geachtet werden. Wald-Tests bieten sich an, wenn zu einem bereits geschätzten Modell ein Teilmodell getestet werden soll, Score-Tests sind dagegen besser geeignet, wenn ein geschätztes Modell gegen ein Obermodell getestet wird [18].

2.3.2 Gütekriterien

In diesem Abschnitt werden verschiedene Gütekriterien zur Bewertungen und zur Auswahl von Modellen vorgestellt. Dabei ist zu beachten, dass viele Gütekriterien oft nur unter bestimmten Voraussetzungen gelten und nicht für jede Methode angewendet werden können. Ein Kriterium, dass in *linearen Regressionsmodellen* herangezogen wird, ist das sogenannte Bestimmtheitsmaß R^2 (siehe z.B. Fahrmeir et al. [18]). Dieses wird mit folgender Formel berechnet:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.4}$$

Es gilt $0 \leq R^2 \leq 1$. Je näher das Bestimmtheitsmaß bei 1 liegt, umso besser ist die Datenanpassung des Modells, Werte nahe 0 weisen auf eine schlechte Datenanpassung hin. Ein Nachteil im Zusammenhang mit dem Bestimmtheitsmaß ist, dass dieses automatisch größer wird, wenn man eine neue Kovariable in das Modell aufnimmt. Der Vergleich von verschiedenen R^2 ist deshalb nur für Modelle mit der gleichen Anzahl an Parametern oder durch das korrigierte Bestimmtheitsmaß möglich. Das korrigierte Bestimmtheitsmaß enthält eine Korrektur für die Anzahl der Parameter. Damit wird es mit Aufnahme einer zusätzlichen Kovariable nicht zwangsläufig größer. Fahrmeir et al. [18] raten jedoch von der Verwendung des korrigierten Bestimmtheitsmaßes ab, da die Bestrafung für zusätzliche Kovariablen relativ gering ist.
Ein Gütekriterium, dass im Rahmen der *Maximum-Likelihood-Inferenz* verwendet wird, ist das Informationskriterium nach Akaike (AIC) (siehe z.B. Fahrmeir et al. [18]). Dieses ist definiert als:

$$AIC = -2l(\hat{\boldsymbol{\Theta}}) + 2p \tag{2.5}$$

Durch den Term $2p$ (entspricht Anzahl der Parameter) wird ein überparametrisiertes Modell bestraft [18]. Je kleiner der Wert des AIC, umso besser ist das Modell. Eine Alternative zum AIC ist das Bayesianische Informationskriterium (BIC) (auch Schwartz-Kriterium genannt)[18]:

$$BIC = -2l(\hat{\boldsymbol{\Theta}}) + log(n)p \tag{2.6}$$

Beim BIC werden komplexere Modelle stärker bestraft, es werden also sparsamere Modelle gewählt als beim AIC [18]. Die Kriterien AIC und BIC werden z.B. bei der Auswahl von Variablen mit Vorwärts- und Rückwärtsselektion (engl.: Forward- und Backward-Selection) verwendet (siehe z.B. bei Fahrmeir et al. [18]). Während bei der Vorwärtselektion in jedem Schritt diejenige Kovariable aufgenommen wird, welche die größte Reduktion des Modellwahlkriteriums liefert, startet man bei der Rückwärtselektion mit einem vollen Modell, das alle Kovariablen enthält. In jeder Interation wird nun die Kovariable aus dem Modell entfernt, welche die mit der größten Reduktion des gewählten Modellwahlkriteriums einhergeht. Eine Kombination aus beiden Methoden stellt schließlich die schrittweise Selektion (engl.: Stepwise-Selection) dar. Hier können in jedem Schritt Variablen aufgenommen oder entfernt werden.
Eine Alternative zu den likelihood-basierten Methoden AIC und BIC ist die baysianische Modellwahl. Bei dieser Methode werden die Posteriori-Modellwahrscheinlichkeiten betrachtet und anschließend das Modell mit der höchsten Posteriori-Wahrscheinlichkeit ausgewählt [22].

2.3.3 Modelldiagnose

Bei der Modelldiagnose geht es primär darum, zu überprüfen, ob die gewählten Annahmen auch richtig sind. Da bei der Anpassung eines Modells am Computer i.d.R. fast jedes Modell - unabhängig davon, ob es geeignet ist oder nicht - gefittet wird [7], ist dieser Check

sehr wichtig. Häufig geschieht dies durch das Betrachten der Residuen. Im Folgenden wird die Modelldiagnose anhand der Beispiele des klassischen linearen und des generalisierten linearen Modells erläutert.
Beim *klassischen linearen Modell* werden folgende Annahmen getroffen [18]:

1. Der Erwartungswert der Residuen ist gleich $\mathbf{0}$ ($E(\epsilon) = \mathbf{0}$)
2. Die Residuen sind von einander unabhängig ($Cov(\epsilon) = \sigma^2\mathbf{I}$)
3. Die Designmatrix $\mathbf{X}$ besitzt vollen Spaltenrang
4. Die Residuen sind normalverteilt ($\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$)

Gerade Annahme 2 ist bei Modellen mit Zeitreihen jedoch häufig nicht erfüllt. Die einzelnen Werte hängen (zeitverzögert) voneinander ab. Man spricht in diesem Fall von Autokorrelation (siehe z.B. Fahrmeir et al. [18]).
Abbildung (beispielhafte Darstellung nach Fahrmeir et al. [18])[1] zeigt die Überprüfung von Modellen auf Homo- und Heteroskedastizität bzw. Varianzengleichheit. Wenn Homoskedastizität vorliegt, bedeutet dies, dass die Fehler nicht systematisch größer oder kleiner werden, liegt dagegen Heteroskedastizität vor, so lässt sich eine Systematik in den Residuen erkennen. Ist diese Annahme jedoch verletzt, so führt dies dazu, dass Hypothesentest und Konfidenzintervalle der Regressionsparameter beeinflusst werden [18].

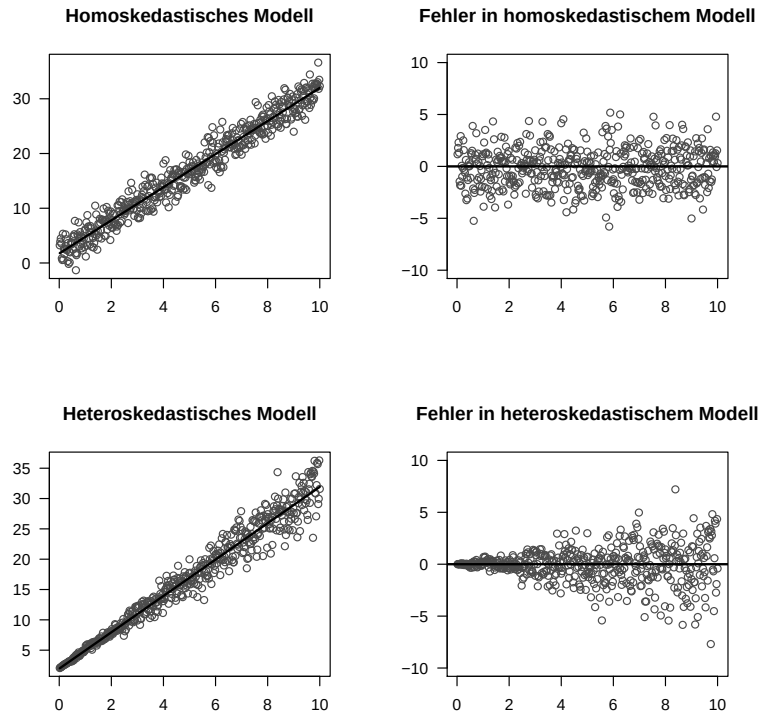

Abbildung 2.2: Überprüfung Homoskedastizität

[1] R-Code aus Tutorium zur Vorlesung „Schätzen und Testen I“ im Wintersemester 2009/10 [37]

Einer weiterer Punkt, der in der Regel bei *generalisierten linearen Modellen* überprüft wird, ist das Vorhandensein von Über- bzw. Unterdispersion (siehe z.B. Fahrmeir et al. [18]). Generalisierte lineare Modellen umfassen Zielvariablen, die nicht zwangsweise normalverteilt sind. Dies können etwa binäre Zielgrößen (Kunde ist kreditwürdig: ja oder nein) oder poissonverteilte Zielgrößen (Anteil von Schadensfällen in einer Versicherung) sein. Bei diesen Zielgrößen hängen Erwartungswert und Varianz teilweise voneinander ab. So gilt für die Poissonverteilung, dass Erwartungswert und Varianz gleich sind. Ist nun die Varianz größer (kleiner) als erwartet, so liegt Überdispersion (Unterdispersion) vor. Eine Möglichkeit damit umzugehen bietet z.B. die Methode der Quasi-Likelihood-Schätzung, die Erwartungswert und Varianz separat spezifiziert.

2.3.4 Modellkomplexität und Kreuzvalidierung

Will man sich zwischen verschiedenen konkurrierten statistischen Modellen entscheiden, so muss man immer einen Kompromiss zwischen möglichst guter Datenanpassung und zu großer Modellkomplexität, die auf eine hohe Anzahl von Parametern zurückzuführen ist, treffen [18]. Einige Gütekriterien, die die Anzahl der Parameter berücksichtigen, wurden schon in Abschnitt 2.3.2 errötert. In diesem Anschnitt wird nun konkret auf das Wechselspiel und die Zusammenhänge zwischen Modellkomplexität, Modell- und Prognosegüte eingegangen. Zusätzlich wird ein weiteres Modellwahlkriterium, die *Kreuzvalidierung* näher vorgestellt (siehe z.B. bei Fahrmeir et al. [18] oder Hastie et al. [21]). Mit der Validierung eines Modells wird geprüft, ob es (z.B. für die Prognose von Krankheitsverhäufen) zufriedenstellende Ergebnisse liefert [8]. Am Beispiel einer medizinischen Studie unterscheiden Altmann und Royston [8] drei Arten von Validierung: interne, zeitliche und externe Validierung. Bei den letzten beiden Arten erfolgt die Validierung anhand von weiteren Datensätzen, bei der internen Validierung wird für die Modellierung und die Validierung derselbe Datensatz verwendet. Diese Methode bezeichnet man als Kreuzvalidierung. Dafür werden die Daten in einen *Trainingsdatensatz* und in einen *Testdatensatz* aufgeteilt. Die Aufteilung der Datensätze kann dabei zufällig oder nicht zufällig geschehen [8]. Das Modell wird nun zunächst auf dem Trainingsdatensatz gefittet. Anschließend wird mit dem Testdatensatz der Vorhersagefehler ermittelt. Alternativen zu der einfachen Kreuzvalidierung sind die k-fache Kreuzvalidierung und die n-fache Kreuzvalidierung auch als „leaving-one-out cross validation“ bezeichnet. Bei der erstgenannten Option wird der Datensatz in k Teildatensätze aufgeteilt. k-1 Datensätze werden nun als Trainingsdatensatz und der k-te Datensatz als Testdatensatz verwendet. Die ganze Prozedur wird k-mal wiederholt bis jeder Teildatensatz einmal als Testdatensatz verwendet wurde. Als Ergebnis erhält man k Fehlerraten, die zu einer gesamten Fehlerrate gemittelt werden. Hastie et al. [21] empfehlen als optimale Wahl für k die Unterteilung in 5 oder 10 Teildatensätze. Bei der n-fachen Kreuzvalidierung besteht der Trainingsdatensatz aus jeweils n-1 Fällen und der Prognosefehler wird anhand des n-ten Falles ermittelt. Zudem existiert noch die geschachtelte Kreuzvalidierung. Diese besteht aus zwei Stufen. In der ersten Stufe werden die Parameter ausgewählt (z.B. Grad des Polynoms bei einem Regressionsmodell). In der zweiten Stufe werden diese gewählten Parameter optimiert. Bei der Angabe des Fehlers muss darauf geachtet werden, dass man

immer den geschachtelten Kreuzvalidierungsfehler angibt und nicht nur den kreuzvalidierten Fehler [21]. Ein Vorteil der Kreuzvalidierung liegt in ihrer Plausibilität und Einfachheit. Als Nachteil kann, gerade bei der n-fachen Kreuzvalidierung, die erhöhte Rechenkapazität angesehen werden [21].
Je komplexer ein Modell ist, umso besser erscheint zunächst die Modellgüte. Hier muss allerdings beachtet werden, dass bei einer erhöhten Anzahl an Parametern auch viel Rauschen aus den Trainingsdaten erklärt wird. Wendet man nun einen unabhängigen Validierungs- oder Trainingsdatensatz auf dieses Modell an, so kann es sogar sein, dass ab einer bestimmten Anzahl an Parametern die Fehler wieder steigen. Abbildung 2.3 [2] stellt dies beispielhaft dar. Die blaue Linie zeigt dabei die Rate des Testfehlers, die rote Linie die Rate des Trainingsfehlers. Gleichzeitig ist in dieser Abbildung noch das Wechselspiel zwischen Bias und Varianz ersichtlich. Je komplexer das Modell wird, umso stärker passt sich der Trainingsdatensatz an die vorliegenden Strukturen an und der Bias sinkt, bei gleichzeitigen Anstieg in der Varianz der Daten [21].

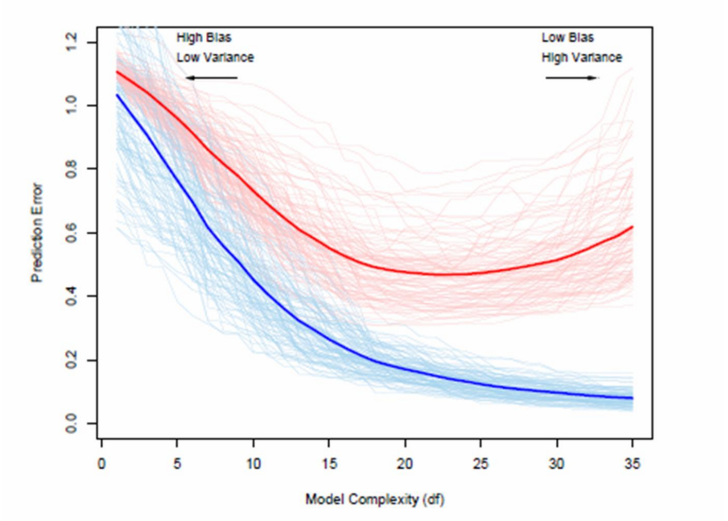

Abbildung 2.3: Test- und Trainingsfehler

Eine zu starke Anpassung an die Trainingsdaten wird auch als *Overfitting* bezeichnet. Bei Regressionsmodellen bedeutet dies z.B., dass die Anzahl der Regressionskoeffizienten steigt. Mit den zusätzlichen Parametern wird jedoch hauptsächlich Rauschen aus den Trainingsdaten erklärt, die Übertragbarkeit auf andere Daten sinkt. Ein Extremfall ist hier das saturierte Modell. Dieses besitzt soviele Parameter wie Freiheitsgrade. Damit ist das Modell allerdings deterministisch und es wird keine Statistik mehr betrieben.

2.4 Typen von Modellen

Wie bereits mehrfach erläutert, gibt es viele Arten bzw. Typen von (statistischen) Modellen. Es erscheint daher logisch und sinnvoll, eine Einteilung verschiedener Modelle nach

[2] Abbildung aus „The Elements of Statistical Learning: Data Mining, Inference and Prediction.“ [21]

bestimmten Gesichtspunkten vorzunehmen. In der Literatur gibt es, ja nach Autor, unterschiedliche Einteilungen. Für die folgenden Abschnitte wird die Systematik nach Dempster [14] verwendet. Dieser unterscheidet Modelle nach ihrer Art der Relation, also den in Abschnitt 2.2 erwähnten „Beziehungen“, in empirische, stochastische und prädiktive Modelle.
Empirische Modelle sind ein sind ein „Teilgebiet“ der explorativen Datenanalyse [14], sie basieren weniger auf subjekt-spezifischen Überlegungen, als darauf was die Parameter auf „lange Sicht“ (mit vielen hypothetischen Wiederholungen) sein würden. Die Darstellung empirischer Modelle erfolgt oft graphisch (z.B. durch unterschiedliche Farben und durch 2D- bzw. 3D-Darstellungen [14]). Für die empirische oder auch explorative Datenanalyse eignen sich nicht-parametrische Verfahren sehr gut [23]. Basierend auf diesen Überlegungen können Hypothesen über mögliche Verteilungen aufgestellt werden. So weist z.B. Dempster [14] darauf hin, dass an die empirischen Daten einer Kaplan-Meier Kurve auch eine parametrische Verteilungsfunktion wie z.B. eine Weibull-Verteilung angepasst werden kann. Auf die Unterschiede der Anpassung von Modellen mit Kaplan-Meier Kurven und mit Weibull-Verteilung in der Praxis wird in Kapital 3 noch einmal ausführlicher eingegangen. Häufiger sind Modelle, die unbekannte Parameter in der Gesamtheit schätzen [11]. Der Zusammenhang zwischen den einzelnen Variablen beruht nun auf stochastischen Beziehungen. Die Daten für ein *stochastisches Modell* beruhen auf zufälligen Stichproben. Wie im Urnenmodell auch hat jedes Element in der Stichprobe eine (berechenbare) Wahrscheinlichkeit, in die Stichprobe zu kommen. Man hat nun eine Zufallsvariable. Über ein bestimmtes Maß der Wahrscheinlichkeit werden alle möglichen Realisierungen angegeben. Wie das Maß der Wahrscheinlichkeit und die möglichen Realisierungen aussehen, hängt von der zugrunde liegenden Verteilung ab [14]. Der Zusammenhang zwischen empirischen und stochastischen Modellen ist durch formale mathematische Strukturen gegeben [14]. Der Unterschied ist, dass bei empirischen Modellen, etwa im Fall der Normalverteilung, Mittelwert und Varianz als gefittete Werte angesehen werden, während sie bei stochastischen Modellen als Schätzungen über die entsprechenden Werte in der Gesamtpopulation angesehen werden können.
Bei einem *prädiktiven Modell* wird davon ausgegangen, dass es zwischen der zu prognostizierenden, zukünftigen Variable y und den erklärenden beobachtbaren Variablen einen Zusammenhang gibt [23]. Dempster [14] sieht, wie bei empirischen Modellen auch, stochastische Modelle wieder als eine Untergruppe der prädiktiven Modelle. Das entscheidende Kriterium ist hier wieder, dass die Aussagen über die prognostizierende Variablen wieder mit einer Wahrscheinlichkeit getroffen werden.
In Abschnitt 2.1 werden als Ziele von Modellen Annäherung bzw. Approximation, Erklärung und Vorhersage genannt. Je nach (Haupt-)ziel wird nun wahrscheinlich ein Modelltyp bevorzugt werden. So wird ein Anwender, der am Beginn einer Analyse ist, wahrscheinlich erst einmal mit einem empirischen Modell beginnen. Je nach weiterer Zielsetzung wird er danach eventuell eher ein stochastisches oder eher ein prädiktives Modell bevorzugen.
Insgesamt ist jedoch festzustellen, dass die Einordnung eines Modells in eine bestimmte Klasse sehr schwierig ist und eine Zuordnung oft nicht zu 100% erfolgen kann (siehe auch Lehmann [27]).

3 Beispielhafte Anwendungen

In diesem Kapitel werden drei verschiedene beispielhafte Modellierungen aus der Medizin, der Ökonometrie und der Klimatologie vorgestellt und ihre Gemeinsamkeiten und Unterschiede herausgearbeitet. Dabei werden zunächst nicht-parametrische und parametrische Modellierungen von Verweildauern in der Medizin und in der Ökonometrie gegenübergestellt. Anschließend wird anhand eines Klimaszenarios die Verwendung eines prädiktiven Modells gezeigt.

3.1 Unterschiedliche Modellierungen von Verweildauern

In vielen wissenschaftlichen Disziplinen werden Überlebenszeiten bzw. Verweildauern untersucht. Beispiele dafür sind:

- Medizin: Überlebenszeit eines Patienten mit einem bestimmten Tumor bei unterschiedlichen Behandlungsmethoden
- Ökonometrie: Verweildauer in der Arbeitslosigkeit, abhängig von Alter, Geschlecht und Qualifikation
- Marketing: Dauer bis zum Produktwechsel bei unterschiedlichen Werbemaßnahmen

In allen Fällen ist das Ziel immer dasselbe: Es wird das „Verbleiben“ in einem Zustand bis zum Eintritt eines bestimmten Ereignisses modelliert. Die Methoden und die Begriffe unterscheiden sich in den einzelnen Wissenschaften jedoch voneinander. Während in der Medizin häufig non-parameterische Methoden und der Begriff Überlebenszeitanalyse (engl.: survival analysis) verwendet wird, spricht man in der Ökonometrie und in den Sozialwissenschaften eher von Verweildaueranalyse (engl.: failure time analysis) und bevorzugt parametrische Modellierungen [39, 33, 28]. In den folgenden zwei Abschnitten wird jeweils eine non-parametrische und eine parametrische Methode zur Modellierung von Überlebenszeiten bzw. Verweildauern dargestellt und anschließend verglichen.
Für der Verständnis der nächsten beiden Abschnitte werden zunächst einige Begriffe aus der Lebensdaueranalyse kurz vorgestellt:

- Die *Lebensdauer* T ist eine nichtnegative Zufallsvariable. Sie gibt an wie lange das „Verbleiben“ in einem bestimmten Zustand dauert
- Veränderungen in einem Zustand (z.B. Patient stirbt, Individuum fängt wieder an zu arbeiten) werden als *Ereignisse* bezeichnet
- Die *Hazardrate* λ gibt an, wie groß das „Risiko“ ist aus dem Zustand auszuscheiden, gegeben, man hat bis zum aktuellen Zeitpunkt t „überlebt“

- Die *Survivalfunktion* $S(t) = 1 - F(t)$ gibt an, mit welcher Wahrscheinlichkeit ein Zeitpunkt t überlebt wird
- *Zensierung* bezeichnet die Tatsache, dass ein Individuum aus der Studie ausscheidet, ohne das ein Ereignis stattgefunden hat. Dies ist der Fall, wenn kein Ereignis bis zum Ende des Beobachtungszeitraum stattgefunden hat oder wenn ein Individuum (z.B. durch Umzug) nicht mehr unter Beobachtung steht
- *Unter Risiko* ist ein Individuum, solange es unter Beobachtung steht. Dies ist solange der Fall, bis eine Zensierung oder ein Ereignis stattfindet

Eine vertiefte Darstellung zur Lebensdaueranalyse findet sich beispielsweise bei Klein und Moeschberger [26].

3.1.1 Medizin: Non-parametrische Modellierung

Toutenburg [35] weist darauf hin, dass die Verwendung parametrischer Verfahren voraussetzt, dass

1. die Verteilungsform in der Grundgesamtheit bekannt ist
2. die Verteilung mit einer mathematischen Funktion zu beschreiben ist

Dies ist bei vielen medizinischen Untersuchungen nicht gegeben, so dass hier häufig non-parametrische Modellierungen gewählt werden.
Eine non-parametrische Methode zur Schätzung von Überlebenszeiten ist die Kaplan-Meier Methode. Die Grundidee bei der Kaplan-Meier Methode ist, dass die Ereignisse durch die Beobachtungsintervalle nicht fest vorgeben sind, sondern durch die Ereignisse definiert werden. Ein neues Zeitintervall wird durch ein Ereignis (z.B. Tod eines Patienten) definiert. Man berechnet nun die bedingte Wahrscheinlichkeit, dass ein Patient ein bestimmtes Zeitintervall überlebt, falls er zu Beginn des Zeitintervalls noch gelebt hat [39]. Die Gesamtüberlebenszeit lässt sich als Produkt der bedingten Überlebenswahrscheinlichkeiten darstellen, der erhaltene Schätzer wird deshalb auch als „Product-Limit-Estimator“ bezeichnet (siehe z.B. bei Ziegler, Kaplan, Klein und Moeschberger oder Fahrmeir [39, 25, 26, 17]):

$$\hat{S(t)} = \prod_{t_{(k)} \leq t} \frac{n_k - d_k}{n_k} \tag{3.1}$$

Dabei ist d_k die Anzahl der Ereignisse, und n_k die Anzahl der Personen unter Risiko (jeweils zum Zeitpunkt $t_{(k)}$).
Gruppenunterschiede können mit Hilfe des Log-Rank-Tests untersucht werden [39]. In Anlehnung an [34] werden in der folgenden Abbildung 3.1 beispielhaft zwei unterschiedliche Kaplan-Meier Kurve für Patienten, die an akuter lymphatischer Leukämie leiden, dargestellt. Bei Patienten mit der grünen Kaplan-Meier Kurve wurde dabei die Behandlung mit der Chemotherapie fortgeführt, bei Patienten mit der roten Kaplan-Meier Kurve wurde sie nicht weitergeführt. Wie erkennbar, ist die Überlebenswahrscheinlichkeit bei Fortführung der Chemotherapie größer.

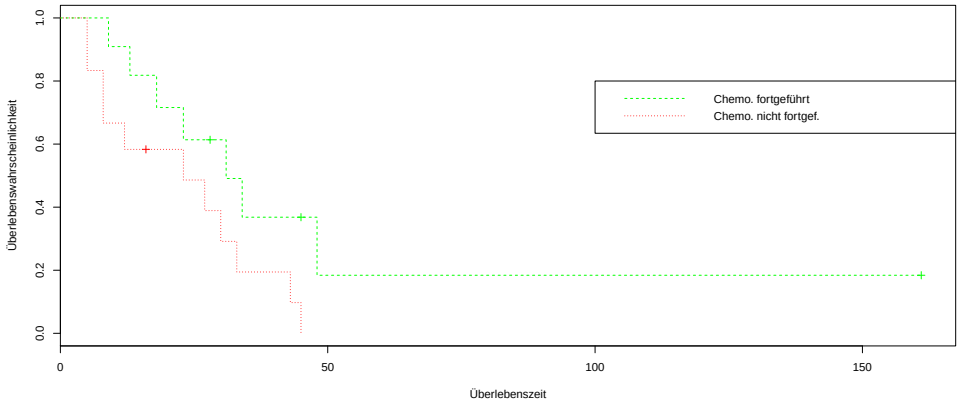

Abbildung 3.1: Beispielhafte Darstellung einer Kaplan-Meier Kurve

Wie in Abschnitt 2.4 schon erläutert, können nun aus dem Verlauf der Kaplan-Meier Kurve Rückschlüsse auf die zugrunde liegende Verteilung gezogen werden. Häufig werden jedoch parametrische Modelle direkt, ohne vorheriges Anpassen, geschätzt. Der nächste Abschnitt zeigt eine Möglichkeit der parametrischen Modellierung von Verweildauern.

3.1.2 Ökonometrie: Parametrische Modellierung von Arbeitslosigkeit

Auch wenn durch den konjunkturellen Aufschwung 2010 die Zahl der Arbeitslosen in Deutschland gesunken ist [6], bleibt die Verweildauer in der Arbeitslosigkeit weiterhin ein Schwerpunktthema in der Ökononometrie [3, 38]. Es ist daher von Interesse, welche Faktoren die Verweildauer in der Arbeitslosigkeit beeinflussen. Die Problemstellung ist hier, wie in Abschnitt 3.1 schon erläutert, analog zur Überlebenszeitanalyse in der Medizin. Der Unterschied besteht zunächst anscheinend nur inhaltlich. Es wird nicht das Überleben bei einer Krankheit sondern die Verweildauer in einem bestimmten Zustand wie z.B. Arbeitslosigkeit modelliert. Anstelle von unterschiedlichen Medikamenten gehen nun Faktoren wie Geschlecht, Mobilität und Bildung als Kovariablen in den Analyse ein. Im Gegensatz zur Medizin werden in der Ökonometrie und der Soziologie statistische Modelle allerdings eher parametrisch modelliert, nichtparametrische Modelle werden vermieden [39, 33, 28]. So modellierte z.B. die Bundesagentur für Arbeit eine Studie zur Verweildauer in der Bezug vom Sozialleistungen durch einen stationären Prozess, der durch eine feste Verteilung bestimmt war [6].
Greene [20] führt folgende mögliche Verteilungsannahmen für die Verweildaueranalyse auf:

- Exponential-Verteilung
- Weibull-Verteilung
- Lognormal-Verteilung
- Loglogistische Verteilung

Diese unterscheiden sich u.a. auch durch die Hazard-Rate: Während sie für die Exponential-Verteilung über die gesamte Dauer der Arbeitslosigkeit konstant ist, ist sie für die Weibull-Verteilung monoton steigend oder fallend (abhängig vom Wert für den zweiten Parameter). Die Lognormal-Verteilung und die Loglogistische Verteilung haben zuerst eine steigende und dann eine fallende Hazard-Rate [20]. Die Wahrscheinlichkeit für einen Arbeitslosen wieder Arbeit zu finden (also das „Risiko“) wird sich erwartungsgemäß über die Zeit der Arbeitslosigkeit ändern. So ist es durchaus denkbar, dass es mit steigender Arbeitslosigkeit immer schwieriger wird, diesen Zustand wieder zu verlassen. Die Hazard-Rate ist in diesem Fall also nicht konstant. Die Modellierung der Verweildauer in der Arbeitslosigkeit kann mit den letztgenannten Verteilungen besser an die Realität angepasst werden. Sehr populär ist die Verwendung der Weibull-Verteilung [20]. Bislang modellieren die oben genannten Modelle die Verweildauer in der Arbeitslosigkeit noch ohne Kovariablen. Diese können allerdings durch folgende Definition der Hazard-Rate berücksichtigt werden (siehe z.B. bei Greene [20]):

$$\lambda_i = \exp -x_i'\beta \tag{3.2}$$

Dabei ist λ_i ein individueller „Beschleunigungs-“ bzw. „Verzögerungsfaktor“, der das individuelle „Risiko“ wieder Arbeit zu finden erhöht oder vermindert. Solche Modelle werden deshalb auch als „accelerated failure time models“ bezeichnet, weil die Kovariablen die Anzahl der gemessenen Einheiten auf der Zeitachse beeinflussen [20].

3.1.3 Vergleich

In den Abschnitten 3.1.1 und 3.1.2 wurden zwei unterschiedliche Methoden vorgestellt, die Überlebenszeiten bzw. Verweildauern in zwei ganz unterschiedlichen wissenschaftlichen Diziplinen modellieren. Dabei wurde gezeigt, dass in der Medizin häufig non-parametrische Methoden wie Kaplan-Meier und in der Ökonometrie eher parametrische Methoden wie z.B. ein Modell mit einer Weibull-Verteilung verwendet werden. Dabei ist allerdings zu beachten, dass dies natürlich keinesfalls allgemein gültig ist. Es finden sich sowohl in den medizinischen als auch in den ökonometrischen Lehrbüchern jeweils Hinweise auf die jeweils „andere“ Methode [20, 26]. Dies ist auch durchaus plausibel. So können z.B. die an Anfang von Abschnitt 3.1.1 erwähnten Gründe für die Verwendung von non-parametrischen Methoden natürlich durchaus auch in der Ökonometrie zutreffen. Ein „Kompromiss“ zwischen non-parametrischen und parametrischen Methoden ist das Cox-Modell. Dieses wird sowohl in der Medizin als auch in der Ökonometrie gerne verwendet [20, 38, 26]. Dem Cox-Modell liegt folgende Formel zugrunde (siehe z.B. Klein und Moeschberger [26]):

$$\lambda(t, x) = \lambda_0(t) exp(x'\beta) \tag{3.3}$$

Der Schwerpunkt liegt bei diesem Modell in der Schätzung des Parametervektors β, $\lambda_0(t)$ wird als Baseline-Hazardrate, die für alle Individuen gleich ist, bezeichnet. Da die Hazardrate für zwei Individuen parallel verlaufen, wird dieses Modell auch als „Proportional-Hazards-Modell“ bezeichnet.

3.2 Prognosemodelle: Klimaszenario

Bei den bisher vorgestellten Modellen in den Abschnitten 3.1.1 und 3.1.2 ging es zwar auch darum Überlebenszeiten und Verweildauern für die Zukunft zu schätzen. Der Fokus lang jedoch nicht auf der Prognose des exakten zukünftigen Zustandes einer Person i, Schwerpunkt war eher die Bestimmung der Determinanten, die die jeweiligen Verweildauern beeinflussen. In diesem Abschnitt wird nun ein Modell erörtert, das explizit Werte für die Zukunft schätzt, also prognostiziert. Die Methode der statistischen Prognose wird zunächst allgemein kurz vorgestellt und anschließend am Beispiel eines Klimaszenarios veranschaulicht.
Abraham [7] klassifiziert in seinem Buch „Statistical Methods for Forecasting“ Prognosemethoden in qualitativ bzw. quantitativ und (ähnlich zu der Einteilung der Modelle in Abschnitt 2.4) in deterministisch und stochastisch. Qualitative Methoden bezeichnet er auch als „subjektiv“ und eher intuitiv. Quantitative Prognosen basieren nach seiner Einteilung auf mathematisch-statischen und werden in deterministisch und stochastisch unterteilt. Als Beispiel für das Vorkommen von deterministischen Modelle nennt er die Physik mit festen „Gesetzen“, stochastische Modelle sieht er dagegen eher in den Sozialwissenschaften, bei denen ein „Rauschen“ in den Daten mit modelliert wird: Es gibt eine Vielzahl von Prognosemodellen wie z.B. parametrische Regressionsmodelle, „Autoregressiv Moving Average“ (ARMA)-Modelle oder Modelle mit einem non-parametrischen Glättungsterm. Bei Interesse werden diese bei Abraham [7] ausführlich vorgestellt.
Da der Schwerpunkt dieser Arbeit jedoch die statistische Modellierung im allgemeinen ist, werden sie hier nicht näher erläutert, es werden stattdessen praxisbezogene Beispiele der Modellierung von Klimazeitreihen und eine Klimaprognose vorgestellt. Als Klima werden Wettererscheinungen bezeichnet, die über einen längeren Zeitraum (i.d.R. 30 Jahre) bestehen [5]. In den letzten Jahrzehnten konnte eine Veränderung des Klimas beobachtet werden (z.B. trockenere Sommer und gestiegene Hochwassergefahr in Bayern), die Ursachen für diese Veränderungen liegen dabei sowohl an natürlichen Ursachen als auch am Einfluss des Menschen [5]. Die Kenntnis des zukünftigen Klimas ist von entscheidender Bedeutung. So werden z.B. durch das Klima im starken Maße die Verbreitung und das Wachstum von Pflanzen in der Land- und Forstwirtschaft beeinflusst [1]. Da Deutschland einer der größten Agrarproduzenten der EU ist [1], sind Strategien für die zukünftige Bepflanzung von Agrarflächen und damit die Kenntnis des Klimas von entscheidender Bedeutung. Eine ausführliche Behandlung der statistischen Modellierung von Klimazeitreihen findet sich in der Disseration von Trömel [36]. Dabei werden die Klimazeitreihen als Realisation eines instatinonären, stochastischen Prozesses betrachtet, die von einem Parametervektor $\boldsymbol{\Theta}$ abhängen, der als Funktion der Zeit modelliert wird. Der Parametervektor $\boldsymbol{\Theta}$ setzt sich zusammen aus einen konstanten Wert, eine (oder die Summe) mehrerer Trendfunktionen, eine zeitlich konstante oder veränderliche saisonale Komponente und eine niederfrequente Komponente. Dabei beschreibt die Trendkomponente längerfristige Änderungen, die saisonale Komponete berücksichtigt jahreszeitlich bedingte Schwankungen und die niederfrequente Komponete erlaubt mehrere lokale Minima bzw. Maxima in einem Beobachtungszeitraum. Als parametrische Verteilungen werden Normal-, Gumbel- und Weibullverteilung gewählt.

Da diese Arbeit sich jedoch nicht mit der Modellierung zukünftiger Klimazeitreihen beschäftigt, sondern z.B. für Deutschland Werte aus den Jahren 1901 bis 2000 betrachtet werden hier keine Detailergebnisse vorgestellt.
Eine Prognose für den Klimawandel bis in das Jahr 2100 wurde im Januar vom Massachusetts Institute of Technology (MIT) erstellt [29]. Dabei wurden verschiedene physikalische und ökonomische Unsicherheitsfaktoren betrachtet (z.B. Erwärmung der Ozeane oder Ausstoß von Kohlenstoffdioxid). Die Simulation der Analysen erfolgte dabei hauptsächlich mit Monte-Carlo-Methoden. Damit können Zufallszahlen aus einer analytisch nicht zugänglichen Posteriori-Verteilung gezogen werden [18]. Dies verlangt zunächst die Festlegung einer Priori-Verteilung. Für die gewählten Parameter wurden diese Priori-Verteilungen entweder durch frühere Studien begründet oder es wurden flache Priori-Verteilungen angenommen. Als Ergebnis aus der Simulation wird beispielsweise eine Erwärmung der Erdtemperatur um 5.1°C im Jahr 2100 prognostiziert.

4 Fazit

Die vorliegenden Kapitel haben gezeigt, dass es nicht ein allgemeingültig „bestes“ Modell gibt. Es gibt keine „generelle Modellstrategie“ [27]. Je nach Ausgangsbedingungen oder Zielsetzung kann ein anderes Modell besser sein, wie z.B. die unterschiedlichen Modellierungsmöglichkeiten für die Verweildaueranalyse zeigen (siehe auch z.B. Lehmann [27]). Desweiteren wird die Entscheidung für eine bestimmte Vorgehensweise in der Modellierung auch von verschiedenen Inferenzschulen abhängen (vergleiche Abschnitt 1.4). Auch Dempster [14] weist darauf hin, dass in einigen Fällen die Wahrscheinlichkeiten in einem stochastischen Modell durch empirische Daten hinreichend genau bestimmt werden können, in anderen Fällen eher durch den Annahme einer Priori-Verteilung. Mehrfach wurde in dieser Arbeit darauf hingewiesen, dass ein Modell immer nur ein zweckgebundenes Abbild der Realität ist, niemals eine 100%-ige Nachbildung. Auch Kaplan [24] merkt an, dass ein Modell immer einem bestimmten Zweck dient. Ein Modell beantwortet immer nur spezifische Fragestellungen. Dazu passt auch gut ein Zitat von dem Physiker Niels Bohr, das sich z.B. bei Dempster (S. 250) [14] findet: „It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature.“. George E. P. Box [2], Professor an der Universität von Wisconsin, geht mit folgendem Zitat sogar noch etwas weiter: „All modells are wrong but some are useful.“. Umso wichtiger ist daher die Wahl des richtigen Modells für die konkrete und spezifische Fragestellung. In Abschnitt 2.3.2 wird ein Überblick über einige formale Kriterien der Modellwahl gegeben. Jedoch dürfen diese Methoden nicht als alleiniges Auswahlkriterium gelten. Mindestens genauso wichtig für eine gute Modellierung sind Kenntnisse über Mechanismen und Kausalitäten, die hinter dem Modell liegen [14, 13]. Die Modellierung von Kausalität setzt allerdings bereits die Kenntnis einer Theorie voraus [40]. Auch in der MIT-Studie wurden die Priori-Verteilungen teilweise mit vorhergehenden Studien begründet. Diese Theorie hat allerdings nicht immer etwas mit Statistik zu tun, sie bezieht sich vielmehr auf die jeweilige Fragestellung, die mit der Modellierung beantwortet werden soll. An keiner Stelle des Prozesses der Modellierung darf ein Modell „für sich stehen“. So fordert z.B. der Soziologe Ziegler [40], dass aufgrund des theoretischen Modells eindeutig angebbar sein muss, welche statistischen Parameter zu berechnen sind und welche Konsequenzen sich aus den Parameterwerten für die theoretischen Aussagen ergeben. Auch Lehmann [27] empfiehlt, dass man sich bei der Modellierung auf Dinge beschränken sollte, von denen man etwas versteht.

Literaturverzeichnis

[1] *Folgen des Klimawandels für Land- und Forstwirtschaft.* http://edoc.hu-berlin.de/miscellanies/klimawandel-28044/75/PDF/75.pdf, Abruf: 16. Dezember. 2010

[2] *George E. P. Box.* http://en.wikiquote.org/wiki/George_Box, Abruf: 19. Januar. 2011

[3] *Grundsicherung für Arbeitssuchende: Verweildauern von Hilfsbedürftigen, Bericht der Bundesagentur für Arbeit.* http://statistik.arbeitsagentur.de/cae/servlet/contentblob/11884/publicationFile/3344/Sonderbericht-Verweildauer.pdf, Abruf: 18. November. 2010

[4] *Internet Encylopedia of Philosophy.* http://www.iep.utm.edu/models, Abruf: 20. November. 2010

[5] *Klimawandel, Warum ändert sich unser Klima?, Bayrisches Landesamt für Umwelt.* http://www.lfu.bayern.de/umweltwissen/doc/uw_79_warum_aendert_sich_unser_klima.pdf, Abruf: 16. Dezember. 2010

[6] *Monatsbericht Oktober 2010 der Bundesagentur für Arbeit.* http://statistik.arbeitsagentur.de/cae/servlet/contentblob/101072/publicationFile/42755/Monatsbericht-201010.pdf, Abruf: 18. November. 2010

[7] Abraham, B. ; Ledolter, J. : *Statistical Methods for Forecasting.* New York : John Wiley & Sons, 1983

[8] Altman, D. G. ; Royston, P. : What Do We Mean by Validating a Prognostic Model? In: *Statistics in Medicine* 19 (2000), S. 453–473

[9] Barnett, V. : The Study of Outliers: Purpose and Modell. In: *Applied Statistics* 27 (1978), S. 242–250

[10] Bayarri, M. J. ; Berger, J. O.: The Interplay of Bayesian and Frequentist Analysis. In: *Statistical Science* 19 (2004), S. 58–80

[11] Cox, D. : Role of Models in Statistical Analysis. In: *Statistical Science* 5 (1990), S. 169–174

[12] Dawid, A. ; Stone, M. : The Functional-model Basis of Fiducial Inference. In: *The Annals of Statistics* 10 (1982), S. 1054–1067

[13] Dempster, A. P.: Causality and Statistics. In: *Journal of Planning and Inference* 25 (1990)

[14] Dempster, A. : Logicist Statistics I. Models and Modeling. In: *Statistical Science* 13 (1998), S. 248–276

[15] Efron, B. : Controversies in the Foundations of Statistics. In: *The American Mathematical Monthly* 85 (1978), S. 231–246

[16] Efron, B. : Why Isn't Everyone a Bayesian? In: *The American Statistican* 40 (1986), S. 1–5

[17] Fahrmeir, L. : *Lebensdauer- und Ereignisanalyse, Vorlesungsskript.* `http://www.statistik.lmu.de/~chris/survival/Skript07.pdf`, Abruf: 20. Januar. 20110

[18] Fahrmeir, L. ; Kneib, T. ; Lang, S. : *Regression - Modelle, Methoden und Anwendungen.* Heidelberg : Springer, 2009

[19] Fisher, R. : The Nature of Probability. In: *Centennial Review* 2 (1958)

[20] Greene, W. H.: *Econometric Analysis.* New Jersey : Pearson Education, 2003

[21] Hastie, T. ; Tibshirani, R. ; Friedman, J. : *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, 2008

[22] Held, L. : *Methoden der statistischen Inferenz.* Heidelberg : Spektrum, 2008

[23] Härdle, W. K. ; Schulz, R. ; Wang, W. : Prognose mit nichtparametrischen Verfahren. In: *SFB 649 Discussion Paper* 041 (2010)

[24] Kaplan, D. T.: *Statistical Modeling.* 2009

[25] Kaplan, E. L. ; Meier, P. : Nonparametric Estimation from Incomplete Observations. In: *Journal of the American Statistical Association* 53 (1958), S. 457–481

[26] Klein, J. P. ; Moeschberger, M. L.: *Survival Analysis: Techniques for Censored and Truncated Data.* New York : Springer, 2005

[27] Lehmann, E. : Model Specification: The Views of Fisher and Neyman, and Later Developments. In: *Statistical Science* 5 (1990), S. 160–168

[28] Ludwig-Mayerhofer, W. : Statistische Modellierung von Verlaufsdaten in der Analyse sozialer Probleme, Teil II: Datenauswertung. In: *Soziale Probleme* 5 (1994)

[29] Sokolov, A. ; Stone, P. ; Forest, C. ; Prinn, R. ; Sarofim, M. ; M. Webster, S. P. ; C.A. Schlosser, D. K. ; Dutkiewicz, S. ; Reilly, J. ; C. Wang, B. F. ; Melillo, J. ; Jacoby, H. : *Probabilistic Forecast for 21st Century Climate Based on Uncertainties in Emissions (without Policy) and Climate Parameters.* `http://globalchange.mit.edu/files/document/MITJPSPGC_Rpt169.pdf`, Abruf: 04. Dezember. 2010

[30] Stachowiak, H. : *Allgemeine Modelltheorie.* Wien : Springer, 1973

[31] *Kapitel* Zur Geschichte des Modelldenkens und des Modellbegriffs. In: Stachowiak, H. : *Modelle - Konstruktion der Wirklichkeit.* München : Wilhelm Fink, 1983, S. 17–146

[32] *Kapitel* Modelle und Mathematik. In: Stachowiak, H. : *Modelle - Konstruktion der Wirklichkeit.* München : Wilhelm Fink, 1983, S. 171–221

[33] Stein, P. ; Noack, M. : *Skript zur Ereignisanalyse.* `http://soziologie.uni-duisburg.de/personen/stein/veranstaltungen/smfl/Ereignisanalyse.pdf`, Abruf: 18. November. 2010

[34] Therneau, T. : *Survival analysis, including penalised likelihood.* `http://cran.r-project.org/web/packages/survival/survival.pdf`, Abruf: 18. November. 2010

[35] Toutenburg, H. : *Moderne nicht-parametrische Verfahren der Risikoanalyse.* Heidelberg : Physica, 1992

[36] Trömel, S. : *Statistische Modellierung von Klimazeitreihen.* Frankfurt am Main, Johann Wolfgang Goethe Universität, Dissertation, August 2004

[37] Waldmann, E. : *R-Code „Vergleich homo- und heteroskedastische Fehlerstruktur", Tutorium zur Vorlesung Schätzen und Testen I im Wintersemester 2009/10.* `http://www.statistik.lmu.de/~semwiso/schaetzentesten1-ws0910/blaetter/Tutorium4.r`, Abruf: 16. Dezember. 2010

[38] Winterhanger, H. : *Determinanten der Arbeitslosigkeit - Neue Erkenntnisse aus der IEB?* `http://ftp.zew.de/pub/zew-docs/dp/dp06077.pdf`, Abruf: 22. November. 2010

[39] Ziegler, A. ; Lange, S. ; Bender, R. : Überlebenszeitanalyse: Eigenschaften und Kaplan-Meier Methode. In: *Statistikserie in der DMW* (2007)

[40] Ziegler, R. : *Theorie und Modell.* München : R. Oldenbourg, 1972