Erweiterung des Cox-Proportional-Hazards-Modells um latente Faktoren und latente Klassen

Dissertation

Zur Erlangung der Würde des Doktors der

Wirtschafts- und Sozialwissenschaften

des Fachbereichs BWL

der Universität Hamburg

vorgelegt von

Lena Herich

aus Bückeburg

Hamburg, August 2012

Vorsitzender: Prof. Dr. K. Peters
Erstgutachter: Prof. Dr. K. Wegscheider
Zweitgutachter: Prof. Dr. M. Clement

Datum der Disputation: 23.10.2012

Inhaltsverzeichnis

Kapitel 1

Einleitung

In Situationen mit komplexen Beziehungen zwischen Kovariaten sind Strukturgleichungsmodelle ein nützliches Werkzeug, um Zusammenhänge abzubilden. Außerdem können multikollineare oder ähnliche Parameter zu latenten Variablen zusammengefasst werden. Diese Erweiterungen sind auch für Fragestellungen in der Ereigniszeitanalyse möglich. Dazu wird das Standard-Modell der Überlebenszeitanalyse, die Cox-Propotional-Hazards-Regression, mit einem Strukturgleichungsmodell kombiniert. Dieser Ansatz ist relativ neu und die Modelle wurden bislang nur selten verwendet.
Die Anwendungsmöglichkeiten dieser Modellklasse sind vielfältig. In der Wissenschaft treten häufig Fragestellungen auf, bei denen ein Aspekt durch mehrere ähnliche Parameter gemessen wird. In der Ökonomie können das Kundenprofile sein, die an unterschiedlichen Kaufgewohnheiten festgemacht werden. Ein Beispiel aus der Medizin sind Diagnosen, die aufgrund verschiedener Kennzeichen und Symptome getroffen werden. In den Sozialwissenschaften werden oft Persönlichkeitsprofile auf der Grundlage von separat gemessenen Teilaspekten definiert.
Wird der Einfluss dieser Effekte auf die Zeit bis zu einem Ereignis untersucht, so werden die beobachteten Kennzahlen standardmäßig als einzelne Parameter in der Cox-Regression modelliert. Häufig werden die resultierenden Effekte jedoch so interpretiert, als ob sie ein verborgenes latentes Phänomen beschreiben.

In dieser Dissertation sollen die Anwendungsmöglichkeiten der Cox-Regression in Kombination mit einem Strukturgleichungsmodell demonstriert werden. Die betrachteten latenten Variablen können kategorielles oder kontinuierliches Skalenniveau besitzen. Anhand von Beispielen sollen die Vor- und Nachteile der Modellansätze im Vergleich zur Standard-Cox-Regression diskutiert werden. Die Prognosegüte der Modelle kann anhand von verschiedenen Kennzahlen verglichen werden. Des Weiteren ist es möglich, die Belastbarkeit der neuen Modelle mithilfe des Bootstrapping-Verfahrens zu untersuchen.

Kapitel 2 beginnt mit einem Überblick über die Literatur zu Überlebenszeitmodellen, in denen latente Variablen verwendet werden.
In Kapitel 3 wird eine Einführung in die Ereigniszeitanalyse gegeben, insbesondere wird die Cox-Proportional-Hazards-Regression vorgestellt. Anschließend werden in Kapitel 4 die Grundlagen von Strukturgleichungsmodellen erläutert. Dabei geht es vor allem um die Definition von latenten kontinuierlichen Variablen. Diese werden auch Faktoren genannt.
In Kapitel 5 wird das Cox-Proportional-Hazards-Modell mit einer kontinuierlichen latenten

Variablen vorgestellt. An dieser Stelle wird das erste Anwendungsbeispiel eingeführt: Der prognostische Wert der Herzfrequenzvariabilität für die kardiale Mortalität. Das vorgestellte Modell wird auf den Datensatz angepasst. Insbesondere wird untersucht, welcher Informationsgewinn mit dieser Analyse im Vergleich zur Standard-Cox-Regression erzielt werden kann. Abschließend wird der Datensatz verwendet, um die Reliabilität des Modells mithilfe der Bootstrap-Methode zu evaluieren.
In Kapitel 6 wird eine kurze Einführung in die latente Strukturanalyse gegeben. Dabei handelt es sich um Strukturgleichungsmodelle, bei denen die latente Variable kategorielles Skalenniveau besitzt. Anstelle eines latenten Faktors können damit latente Klassen definiert werden. Diese Modellklasse kann auch als Submodell von Mischverteilungsverfahren verstanden werden.
In Kapitel 7 wird das Cox-Proportional-Hazards-Modell mit latenten Klassen formal beschrieben. Hier wird noch einmal auf das Beispiel aus Kapitel 5 zurückgegriffen, um die Anwendungsmöglichkeiten des Modells zu verdeutlichen. Schließlich wird die Modellstabilität evaluiert.
Das 8. Kapitel betrachtet eine praktische Fragestellung aus dem Bereich der Ökonomie. Es liegt der Datensatz eines deutschen Mobilfunkanbieters vor, dessen primäres Geschäftsmodell der Vertrieb von Prepaid-Karten ist. Untersucht werden soll der prognostische Wert verschiedener Kundenparameter auf die Nutzungszeit der erworbenen Telefonkarte. Dazu werden zwei Survivalmodelle mit latenten kategoriellen Variablen angepasst: Die Cox-Regression mit latenten Klassen, wie in Kapitel 7 beschrieben, und die Cox-Regression mit einem Latente-Klassen-Growth-Modell. Die Prognosegüte der Analysen wird mithilfe von verschiedenen Fit-Indizes beurteilt und mit der Lösung der Standard-Cox-Regression verglichen.
Im 9. Kapitel wird eine Zusammenfassung über die Ergebnisse gegeben. Anhand der Resultate soll die Frage beantwortet werden, an welchen Stellen Ereigniszeitmodelle mit latenten Variablen einzusetzen sind und welchen Informationsgewinn sie bringen können.

Die Modelle werden ausführlich erläutert. Auf das Vorgehen der Schätzverfahren wird in allgemeiner Weise eingegangen. Das genaue Vorgehen unterscheidet sich in Abhängigkeit von der verwendeten Software. Programme für die Analyse von Strukturgleichungsmodellen sind Amos, Lisrel, EQS, Stata und Mplus. Die Berechnung einer latenten Klassenanalyse ist mit einer Vielzahl von Programmen möglich, dazu gehören Latent Gold, Stata und Mplus. Die kombinierten Survivalmodelle mit einem latenten Faktor oder latenten Klassen können nur mit Mplus berechnet werden. An geeigneter Stelle sind Anmerkungen zum Vorgehen in Mplus gemacht oder entsprechende Literaturverweise aufgeführt.

Kapitel 2

Literaturübersicht

2.1 Modellierung von Überlebenszeitprozessen mit latenten Variablen

Viele Studien betrachten als Endpunkt die Zeit bis zu einem Ereignis. Das am häufigsten verwendete Modell zur Analyse dieser Fragestellungen ist das Cox-Proportional-Hazards-Modell. In der Vergangenheit wurde das Modell durch zahlreiche Ansätze erweitert. Eine der neusten Entwicklungen beschäftigt sich mit der Berücksichtigung von latenten Variablen.
Artikel, in denen Ereigniszeitprozesse mit Strukturgleichungsmodellen kombiniert werden, stammen von den Entwicklern des Softwareprogramms Mplus, B. Muthen und T. Asparaouhov sowie K. Larsen, einem Schüler von Muthen. Erste Ansätze auf diesem Gebiet finden sich davor nur in zwei Publikationen einer Gruppe um H. Lin und C. E. McCulluch.
Zunächst soll eine kurze Übersicht über die Artikel zu diesem Thema gegeben werden. Anschließend werden die einzelnen Publikationen genauer vorgestellt, dabei wird in gebotener Kürze auf die Anwendungsbeispiele eingegangen. Insbesondere werden die verwendeten Modelle beschrieben. Auf die zugehörigen Schätzverfahren und den Prozess der Modellanpassung soll nur am Rande eingegangen werden. Zum Abschluss wird eine Übersicht gegeben, welche Fragestellungen mithilfe von Ereigniszeitmodellen mit latenten Variablen beantwortet wurden.

Lin, H., Turnbull, B. et al. verwenden in dem Artikel *"Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer"* [LTMS02], siehe auch [MLST02], den Teildatensatz einer Präventionsstudie zum Prostatakrebs. Als Outcome werden longitudinale Messungen des prostataspezifischen Antigens und das Auftreten eines Karzinoms betrachtet. Es werden latente Klassen spezifiziert, um die Heterogenität in der untersuchten Population abzubilden. Die Verlaufsmessungen werden mit einem linearen gemischten Modell mit einem klassenspezifischen Effekt analysiert. Zur Modellierung des Überlebenszeitprozesses wird eine Cox-Regression mit einer klassenspezifischen Baseline-Hazard-Funktion angepasst. Bedingt auf die latente Klassenzugehörigkeit werden die beiden Prozesse als unabhängig angenommen. Das Modell wurde von C. Proust-Lima and J.M. Taylor später in dem R-Packet "lcmm" implementiert, siehe u.a [PLT09].

Larsen stellt in dem Artikel *"Joint analysis of time to event and multiple binary indicators of latent classes"* [Lar04] ein Modell vor, in dem der Einfluss einer kategoriellen latenten

Variable durch den Achsenabschnittsparameter in einem Cox-Proportional-Hazards-Modell berücksichtigt wird. Die latenten Klassen werden durch kategorielle Indikatorvariablen definiert, dafür wird eine latente Klassenregression verwendet.

In dem Artikel *"The Cox PH Model with a continuous latent variable measured by multiple binary indicators"* [Lar05] betrachtet Larsen statt einer kategoriellen, eine kontinuierliche latente Variable. Diese wird durch binäre Indikatorvariablen mit einem Item-Response-Modell definiert. Der latente Faktor wird in der Cox-Regression als Kovariate berücksichtigt. Larsen verwendet in beiden Artikel die Daten einer Studie aus der Geriatrie zur Veranschaulichung der vorgestellten Methoden.

In der Publikation *"Continuous time survival in latent Variable Models"*[AMM06] ordnen Muthen und Asparahouv die von Larsen vorgestellten Modelle in einen weiteren Rahmen ein. Vorgestellt wird ein sogenanntes Multilevel-Survival-Mixture-Modell. In dieser Modellformulierung kann die Hazard-Funktion kontinuierliche latente Variablen als Kovariaten enthalten, die durch kategorielle oder kontinuierliche Items gemessen werden. Gleichzeitig bietet der Mixture-Teil des Modells die Möglichkeit, mithilfe von latenten Klassen Subgruppen in der Population zu modellieren. Es gibt verschiedene Möglichkeiten, den Einfluss dieser latenten kategoriellen Variablen auf die Ereigniszeit zu berücksichtigen. In der Hazard-Funktion können die Parameterschätzer der Kovariaten klassenspezifisch geschätzt werden, außerdem kann ein klassenspezifischer Achsenabschnittsschätzer berücksichtigt werden. Dazu kommt noch ein Multilevel-Teil. Das heißt, dass alle Parameterschätzer des Modells auch clusterspezifisch geschätzt werden können.

Der Artikel *"Continuous-Time Survival Analysis in Mplus"* [MA06] von Muthen und Asparahouv ist im Wesentlichen eine Ergänzung zu [AMM06], in dem die Implementierung der verschiedenen Modelle in Mplus erläutert wird. Auf diesen Artikel soll an dieser Stelle nicht weiter eingegangen werden.

Die neueste Publikation auf diesem Gebiet von Muthen, Asparahouv, Baye et al. *"Applications of Continuous time Survival Analysis in latent Variable Models for the Analysis of Oncology Randomized Clinical Trial Data using Mplus"* [MAB$^+$09] untersucht den Einfluss der Lebensqualität auf die Prognose von Patienten mit einem Pleuramesotheliom. Die Lebensqualität wird anhand verschiedener Items zur Baseline und im Verlauf gemessen. Es werden die bereits bekannten Methoden aus dem Artikel [AMM06] zur Modellierung des Einflusses latenter Variablen auf die Überlebenszeit angewendet. Für die Analyse der longitudinalen Daten werden Growth-Mixture-Modelle [MA08] angepasst. Dabei werden verschiedene Klassen für unterschiedliche Verläufe der Lebensqualität im Zeitverlauf geschätzt. Abschließend wird auch dieses Modell mit der Überlebenszeitanalyse kombiniert.

Ein kombiniertes Modell mit latenten Klassen

Der erste Artikel, in dem bei der Modellierung eines Überlebenszeitprozesses eine latente kategorielle Variable Berücksichtigung findet, stammt von 2002. Lin et al. [LTMS02] untersuchen die Daten eine retrospektive Teilstudie des "Nutritional Prevention of Cancer Trials" (NPC). Es liegen Verlaufsmessungen des prostataspezifischen Antigens (PSA) vor, einem Biomarker für Prostatakrebs. Als Zielgröße wird einerseits der Verlauf des PSA über die Zeit,

andererseits die Zeit bis zum Auftreten eines Karzinoms betrachtet.
In bisherigen Modellierungsansätzen wurde bei der Überlebenszeitanalyse für alle Patienten eine gemeinsame Baseline-Hazard-Funktion angenommen. Die Autoren argumentieren, dass diese Annahme in einer heterogenen Population nicht gerechtfertigt ist. Es besteht insbesondere die Vermutung, dass die betrachtete Population aus verschiedenen Subgruppen besteht, die sich in ihrem PSA-Verlauf unterscheiden.
Für die Modellierung wird ein Joint-Latent-Class-Modell vorgeschlagen. Dazu gehört zunächst eine latente Klassenanalyse zur Bestimmung der unbeobachteten Klassenzugehörigkeit. Die Identifizierung der Klassen geschieht mit einem multinominalen Logit-Modell. Die longitudinalen Messungen des PSA werden mit einem linearen gemischten Modell mit klassenspezifischen Effekten untersucht. Als zweiter Endpunkt wird das Eintreten eines Prostatakarzinoms mit einem Cox-PH-Modell betrachtet. Der Einfluss der latenten Klassen wird hier durch eine klassenspezifische Baseline-Hazard-Funktion modelliert.
Das lineare gemischte Modell und die Cox-Regression sind in dem vorliegenden Artikel noch weiter an die spezifische Fragestellung angepasst, das Cox-Modell beinhaltet unter anderem einen Frailty-Term. Bei der Modellierung wird die Annahme getroffen, dass die longitudinalen Messungen und das Überlebenszeitmodell in Abhängigkeit von den latenten Klassen voneinander unabhängig sind. Die Parameter für das gesamte Modell werden gemeinsam mit der semiparametrischen Maximum-Likelihood (ML)-Methode geschätzt.

Cox-PH-Modell mit latenten Klassen

Larsen verwendet in seinem Artikel [Lar04] ebenfalls eine kategorielle latente Variable als Prädiktor für die Zeit bis zum Eintreten eines Ereignisses. Das latente Konstrukt wird hier durch eine Reihe dichotomer Variablen gemessen.
In beiden Artikeln [Lar04], [Lar05] werden Daten der "Woman´s Health and Ageing Study" (WHAS) verwendet, welche die Ursachen und Verläufe von körperlicher Funktionsunfähigkeit bei älteren Menschen untersucht. Zunächst werden fünf Items auf einem Fragebogen betrachtet, die verschiedene Aspekte der schweren Bewegungsunfähigkeit messen. Es stellt sich die Frage nach dem prognostischen Wert dieser Variablen auf die Überlebenszeit.

Es wird eine latente Klassenregression verwendet, dass heißt, eine latente Klassenanalyse mit Kovariaten, um den Zusammenhang zwischen den fünf dichotomen Indikatorvariablen, manifesten Variablen und einer kategoriellen latenten Variable zu modellieren.
Um den Zusammenhang zwischen beobachteten Kovariaten und den latenten Klassen zu untersuchen, wird, wie bei Lin et al., eine multinomiale logistische Regression verwendet. Hierbei wird die Annahme getroffen, dass die Kovariaten direkt auf die latente Variable wirken und nicht auf die beobachteten Items. Diese Annahme kann durch die Anpassung eines erweiterten Modells getestet werden.
Für die Ereigniszeitanalyse wird ein Cox-PH-Modell benutzt. Der Einfluss der latenten Variable wird durch einen klassenspezifischen Achsenabschnittsschätzer im parametrischen Teil modelliert. Die Likelihood-Funktion für das gesamte Modell setzt sich aus den Dichtefunktionen der verschiedenen Teile zusammen.
Larsen stellt in der Arbeit klar die Vorteile heraus, die sich bieten, wenn die gemessenen Items nicht einzeln oder gemeinsam in ein Cox-PH-Modell aufgenommen werden, sondern vorher zu einer latenten Variable zusammengefasst werden. Die wesentlichen Gründe liegen darin, dass die einzelnen Indikatoren jeweils nur Aspekte der Bewegungsunfähigkeit messen; außer-

dem können mithilfe einer latenten Variable Messfehler berücksichtigt und Kollinearitäten zwischen den Indikatoren vermieden werden.

Cox-PH-Modell mit einem latenten Faktor

Für das beschriebene Anwendungsbeispiel aus der Geriatrie stellt sich Larsen in einem zweiten Artikel [Lar05] die Frage, welchen Einfluss eine kontinuierliche latente Variable auf die Zeit bis zu einem Ereignis besitzt. Gemessen wird diesmal die körperliche Funktionsfähigkeit, die durch vier binäre Items eines Fragebogens gemessen wird.
Die Skalierung der latenten Variablen erfordert einen anderen Ansatz zur Modellierung des Zusammenhangs zwischen latenter Variable und beobachteten Indikatoren. Verwendet wird ein Modell der Item-Response-Theorie, das 2-parametrische logistische Modell nach Birnbaum [Bir68]. Es wird wieder davon ausgegangen, dass der Einfluss der Kovariaten nicht auf die gemessenen Items wirkt, sondern nur auf die latente Variable.
Für die Betrachtung der Überlebenszeit wird das Cox-PH-Modell verwendet. Neben einem Vektor von Kovariaten wird in den parametrischen Teil der Hazard-Funktion der Vektor der latenten Variablen und ein Interaktionsterm von manifesten und latenten Variablen eingefügt. Die Likelihood-Funktion für die Berechnung setzt sich wieder aus den Dichtefunktionen der verschiedenen Teile des Modells zusammen. Die Schätzung der Parameter und das Verfahren der Modellanpassung sind in beiden Artikeln von Larsen ähnlich. Die Berechnung der Modelle geschieht mithilfe des Expectation-Maximization (EM)-Algorithmus.

Es werden verschiedene Tests zur Evaluation der Modellannahmen vorgeschlagen und durchgeführt. Die Modellanpassung geschieht jeweils schrittweise. Im Fall der kategoriellen latenten Variable wird zunächst die optimale Anzahl Klassen gesucht und es werden die Kovariaten bestimmt, deren Einfluss bei der latenten Klassenregression signifikant sind. Auch bei dem Modell mit einer kontinuierlichen latenten Variablen wird zunächst, unabhängig von der Betrachtung der Ereigniszeitvariablen, das Messmodell spezifiziert. In einem zweiten Schritt wird dann das gesamte Modell mit latenter Variable, Cox-Regression und zugehörigen Kovariaten angepasst. Auswahlkriterium bei der Entscheidung für ein Modell ist das AIC beziehungsweise das Ergebnis eines Likelihood-Ratio-Tests.
Bei der latenten Klassenregression können die ermittelten Klassen als verschiedene Schweregrade der Bewegungsunfähigkeit interpretiert werden. Die Vorteile des erweiterten Modells sieht der Autor unter anderem darin, dass die Ergebnisse von Fragebogendaten in der Überlebenszeitanalyse verwendet werden können.
Im Item-Response-Modell wird die kontinuierliche latente Variable der körperlichen Funktionsfähigkeit als Effektschätzer interpretiert. Mit dem erweiterten Modell wird zum einen das Problem von Kollinearitäten zwischen Kovariaten in einer Cox-Regression gelöst, zum anderen lassen sich die Wirkungspfade der Kovariaten nachvollziehen.

Cox-PH-Modell mit latenten Klassen und Faktoren

Muthen und Asparahouv stellen in ihrem Artikel [AMM06] einen allgemeinen multivariaten und mehrstufigen Rahmen für die Survivalanalyse für stetige Zeit vor. Inbegriffen ist die gemeinsame Modellierung der Ereigniszeitanalyse und kontinuierlichen und kategoriellen beobachteten und latenten Variablen. Die Zeit bis zu einem Ereignis kann mit verschiedenen

Hazard-Funktionen modelliert werden, dabei werden Rechtszensierungen berücksichtigt.
Kernstück der Arbeit ist ein sogenanntes Multilevel-Latent-Variable-Mixture-Modell. Ein Multilevel-Mixture-Modell für ein Strukturgleichungsmodell findet sich bereits in [MA08], diese Formulierung wird hier um das Cox-PH-Modell erweitert.
Der latente Faktor wird mit einem Messmodell spezifiziert, der Zusammenhang zu anderen latenten kontinuierlichen Variablen wird mit einem Strukturmodell bestimmt. Falls es sich bei den Indikatoren des latenten Faktors um kategorielle Variablen handelt, wird die Latent-Response-Variable-Formulierung verwendet [MA02], dass heißt, es wird eine zugrunde liegende normalverteilte latente Variable definiert. Für die Definition der latenten Klassen wird ein multinomiales Logit-Modell verwendet.
In der Hazard-Funktion werden die latenten Faktoren als Kovariaten berücksichtigt. Im parametrischen Teil können zudem die Einflüsse der manifesten und der latenten Variablen auf die Ereigniszeit klassenspezifisch geschätzt werden. Für die latenten Klassen können vollkommen unrestringierte Baseline-Hazard-Funktionen geschätzt werden. Alternativ kann der Einfluss der latenten Klassen, wie bei Larsen, durch einen klassenspezifischen Achsenabschnitt in der Cox-Regression modelliert werden. Das Modell kann mehrstufig aufgebaut werden. Alle Parameter können entweder als feste oder als clusterspezifische Effekte geschätzt werden.
Das Modell ist im Wesentlichen eine Kombination der Modelle von Larsen, so können hier gleichzeitig latente Klassen und kontinuierliche latenten Variablen in der Ereigniszeitanalyse berücksichtigt werden. Durch die Definition der latenten Klassen bietet sich die Möglichkeit, die Homogenität einer Population infrage zu stellen. Dieser Teil wird als Mixture-Modell bezeichnet.

Im übrigen Teil des Artikels werden Erweiterungen dieses Rahmenwerkes betrachtet. Es werden Frailty Modelle vorgeschlagen, um den Zusammenhang zwischen zwei oder mehr Survivalprozessen zu modellieren. Dabei wird in die jeweiligen Hazard-Funktionen, neben unterschiedlichen nichtparametrischen Baseline-Hazards, ein gemeinsamer parametrischer Teil definiert. Des Weiteren wird auf Modelle mit zeitabhängigen Kovariaten, den Zusammenhang zwischen der Ereigniszeitanalyse in kontinuierlicher und diskreter Zeit und Daten mit Bindungen, sogenannten “Ties”, eingegangen.

Cox-Modell mit einer zeitabhängigen Kovariate, latenten Klassen und Faktoren

In dem neusten Artikel von Muthen und Asparahouv zu dem Thema werden die Daten einer randomisierten Studie aus der Onkologie verwendet. Es werden Patienten mit einem forgeschrittenem Pleuramesotheliom betrachtet. Verglichen werden bestmögliche unterstützende Behandlungsmaßnahmen gegenüber einer Pemetrexed-Chemotherapie. Betrachtet wird die progressionsfreie Überlebenszeit.
In dem vorliegenden Artikel soll untersucht werden, ob die Lebensqualität der Patienten einen Einfluss auf die Zeit bis zu einem Ereignis besitzt. Zum einen geht es um den prognostischen Wert der Messungen zum Studienbeginn, zum anderen stellt sich die Frage, ob unterschiedliche Entwicklungen der Lebensqualität im Zeitverlauf einen Einfluss auf die progressionsfreie Überlebenszeit besitzen. Die Messung der Lebensqualität basiert auf Patientenbefragungen mit dem “Lung Cancer Symptom Scale Mesothelioma Fragebogen” (LCSS). Das Instrument besteht aus neun Items, die jeweils auf einer Skala von 0-100 gemessen werden.
Da die Annahme proportionaler Hazards für den Einfluss der Variable Behandlungseffekt nicht erfüllt ist, wird eine Hazard-Funktion gewählt, die zeitabhängige Einflüsse von Kovaria-

ten erlaubt. Mit diesem Modell wird dann, unter Berücksichtigung des Behandlungseffektes, der Einfluss der Baseline-Variablen auf die Überlebenszeit untersucht.
Es wird diskutiert, wie die einzelnen Items der Lebensqualität zum Baseline-Zeitpunkt mit einem latenten Variablenmodell zusammengefasst werden können. Verglichen werden die Ergebnisse einer Faktorenanalyse, latenter Klassenanalyse und einer Faktor-Mixture-Analyse. Das letztgenannte Modell berücksichtigt gleichzeitig eine kategorielle und eine kontinuierliche latente Variable, wie in [AMM06]. In diesem Modell wird die Annahme der vollkommenen Unabhängigkeit der Items innerhalb einer Klasse entspannt. Items können innerhalb einer Klasse korrelieren, die Stärke ist durch einen Faktor und die zugehörigen Ladungen beschrieben. Die Ergebnisse der verschiedenen Modelle, für eine unterschiedliche Anzahl von Faktoren und latenten Klassen, werden mithilfe verschiedener Fit-Indizes, wie dem CFI und dem BIC, verglichen. Die besten Ergebnisse erzielt die Faktor-Mixture-Analyse mit einem Faktor und zwei latenten Klassen.
Nachdem die Entscheidung für das latente Variablenmodell getroffen ist, wird das Faktor-Mixture-Modell mit dem nichtproportionalen Hazard-Modell kombiniert, um zu untersuchen, ob die Lebensqualitätsmessungen, zusätzlich zu den anderen Kovariaten, einen signifikanten Einfluss auf die progressionsfreie Überlebenszeit besitzen.
Es ergibt sich eine ähnliche Modellformulierung wie in [AMM06]. Die nichtproportional spezifizierte Hazard-Funktion verwendet als Kovariaten eine Reihe manifester Variablen, die latenten Klassen und verschiedene Interaktionen.
Für das Beispiel aus der Onkologie stellt sich heraus, dass für Patienten der Behandlungsgruppe die Mitgliedschaft in einer bestimmten Klasse einen signifikanten Einfluss auf die progressionsfreie Überlebenszeit besitzt. Patienten, die der Klasse mit insgesamt höherer Lebensqualität angehören, besitzen eine bessere Prognose für den Krankheitsverlauf.
In dem Artikel werden außerdem eine Reihe verschiedener Modelle verglichen, die den Einfluss einer longitudinal gemessenen Variable auf die Überlebenszeit berücksichtigen können. Dabei handelt es sich um sogenannte Growth-Survivalmodelle, s. auch [MA08]. Diese Analysen können ebenfalls mit dem Cox-PH-Modell kombiniert werden. Der Artikel zeigt, wie die Modelle an den gegebenen Datensatz angepasst werden können.

Zusammenfassung

Im Wesentlichen betrachten die vorgestellten Artikel zwei Fragestellungen. Wie kann die Heterogenität einer Population bei der Überlebenszeitanalyse adäquat berücksichtigt werden? Ist es möglich, den Einfluss nicht direkt messbarer Variablen auf die Zeit bis zu einem Ereignis zu ermitteln? Bei Larsen [Lar05] werden die Modelle außerdem dazu verwendet, direkte und indirekte Wirkungspfade der Kovariaten auf die Überlebenszeit zu untersuchen.
Die Antwort darauf können Survivalmodelle mit latenten Klassen und latenten Faktoren geben. Die latenten Klassen können durch Indikatorvariablen zum Baseline-Zeitpunkt, mit einer latenten Klassenanalyse (LCA), oder durch longitudinal gemessene Indikatoren, mithilfe der Latenten-Klassen-Growth-Analyse (LCGA), bestimmt werden. Kontinuierliche latente Variablen werden mit der Faktorenanalyse bestimmt. Werden die zugehörigen Items nominal oder kategoriell gemessen, so wird die Item-Response-Theorie oder die Latent-Response-Variablen-Formulierung angewendet; beide Ansätze geben äquivalente Ergebnisse [MA02]. Die Ereigniszeitanalyse wird mit dem Cox-PH-Modell modelliert, nur bei Muthen et al. [MAB$^+$09] ist die Annahme proportionaler Hazards nicht erfüllt und es wird eine Cox-Regression mit einer zeitabhängigen Kovariate angepasst. Die Methoden sind relativ neu, die vorgestellten Artikel

sind die ersten, die ihre Anwendung zeigen. Die Berechnung der Modelle geschieht bei Lin et al. mit Splus, Larsen verwendet, so weit wie möglich, Mplus, die komplexeren Analysen werden ebenfalls in Splus bzw. R programmiert. Muthen und Asparahouv beseitigen den Mangel an Software für die Überlebenszeitanalyse mit latenten Variablen durch die Implementierung der Modelle in Mplus.

Mit den verfügbaren Modellen ist ein praktischer Rahmen für die Berücksichtigung von latenten Variablen in der Überlebenszeitanalyse gegeben. Die Eigenschaften und Verwendungsmöglichkeiten dieser Methoden sollen in den folgenden Kapiteln anhand von praktischen Beispielen ausgelotet werden.

Kapitel 3

Survivalanalyse

3.1 Einführung

In der Überlebenszeitanalyse wird als Zielvariable die Zeit bis zum Auftreten eines Ereignisses betrachtet. Die Zeit kann kontinuierlich (Stunden, Tage, ...) oder kategoriell (Anzahl der Schuljahre, etc.) gemessen werden. Ein Ereignis kann beliebig definiert sein, beispielsweise kann es sich um Tod eines Patienten, den Arbeitsplatzwechsel oder das Ende eines Vertragsverhältnisses handeln. Da die Survivalanalyse besonders in der Medizin verwendet wird, handelt es sich häufig um negative Ereignisse, die im Englischen als "failure" bezeichnet werden. In der vorliegenden Arbeit wird immer von einem genau definierten Endpunkt ausgegangen, miteinander konkurrierende Ereignisse werden nicht untersucht.

Eines der Hauptprobleme in der Überlebenszeitanalyse ist häufig das Fehlen exakter Information über den Ereigniszeitpunkt. Dies kann verschiedene Gründe besitzen, vgl. [KM03]. Zum Beispiel ist es möglich, dass der Beobachtungszeitraum endet, bevor jedes Individuum ein Ereignis gehabt hat. Häufig geschieht es auch, dass Personen aus unbekannten Gründen im Verlauf einer Studie ausscheiden. Für die Schätzung der Überlebenszeitfunktion wird daher zu jedem Individuum, neben der Beobachtungszeit, mithilfe einer Indikatorvariablen die Information aufgenommen, ob eine Zensierung oder ein Ereignis stattgefunden hat. Dabei wird die Annahme getroffen, dass die Zensierung unabhängig von dem Auftreten des Ereignisses ist.
Es gibt verschiedene Arten von Zensierungen. Bei intervallzensierten Daten fehlt die Information, wann genau in einem Zeitintervall die Zensierung stattgefunden hat. Bei Linkszensierungen fehlt die Information, wann ein Individuum, das zum Beispiel eine Krankheit diagnostiziert bekommen hat, sich vorher infiziert hat. Am häufigsten treten Rechtszensierungen auf. Bei diesen ist der Beobachtungszeitraum auf der rechten Seite der Zeit abgeschnitten und es fehlt die Information, ob ein Individuum danach noch ein Ereignis bekommt. In diesem Fall ist nur bekannt, dass die Überlebenszeit größer als die Beobachtungszeit ist [KM03, KK05]. Im Folgenden sollen ausschließlich Rechtszensierungen betrachtet werden.
Mit Überlebenszeitanalyse kann nun unter anderem das Risiko eines Individuums geschätzt werden, zu einem bestimmten Zeitpunkt das Ereignis von Interesse zu bekommen. Dabei werden alle Individuen betrachtet, die bis zu dem entsprechenden Zeitpunkt noch kein Ereignis hatten und sich unter Risiko befinden. Zur Beschreibung der Prozesse wird das Prinzip der Survival-Funktion und Hazard-Funktion verwendet, dass im folgenden Abschnitt beschrieben

wird. Die Konzepte werden hier in Anlehnung an die Bücher von Kalbfleisch und Prentice [KP80], Klein und Moeschberger [KM03] sowie Kleibaum und Klein erläutert [KK05].

3.2 Grundlegende Definitionen und Modellspezifikation

Es bezeichne t die Zeit seit dem Beginn der Beobachtung. Diese wird im Folgenden immer als stetig angenommen. Des Weiteren bezeichne U_i eine Zufallsvariable für den Ereigniszeitpunkt eines Individuums i.

Die Notation im Fall von rechtszensierten Daten ist wie folgt: Neben der Überlebens- oder Ereigniszeit U_i wird die Zensierungszeit C_i der Individuen aufgenommen. Dabei ist die genaue Lebenszeit eines Studiensubjektes nur bekannt, wenn diese kleiner oder gleich der Zensierungszeit ist. Es bezeichne $T_i = min(U_i, C_i)$ die Beobachtungszeit eines Individuums i. Ein Zensierungsindikator $D \in (0, 1)$ gibt an, ob ein Ereignis stattgefunden hat. Die Informationen für ein Individuum können durch ein Paar von Zufallsvariablen (T_i, D_i) zusammengefasst werden.

Es gibt verschiedene Möglichkeiten, die Wahrscheinlichkeitsverteilung von T zu spezifizieren. In der Überlebenszeitanalyse ist es üblich, die Survival-Funktion $S(t)$ und die Hazard-Hunktion $h(t)$ anzugeben, zum Verständnis ist es auch interessant, die Dichteverteilung $f(t)$ zu betrachten. Ist eine der drei Formen bekannt, so können die beiden anderen daraus berechnet werden.

Die Überlebenszeitfunktion (**Survival-Funktion**) ist definiert als die Wahrscheinlichkeit, dass ein Individuum an einem spezifizierten Zeitpunkt noch lebt. Formell heißt das:

$$S(t) = P(T > t). \tag{3.1}$$

Aus dieser Definition folgt, dass es sich bei $S(t)$ um eine nicht steigende, linksstetige Funktion handelt, mit der Eigenschaft $S(0) = 1$ und $S(t), t \to \infty = 0$. Hier wird die Wahrscheinlichkeit des rechten Endes der sonst in der Statistik üblicherweise verwendeten kumulierten Verteilungsfunktion angegeben. Entsprechend ergibt sich der Zusammenhang zu $F(t)$, der Wahrscheinlichkeitsfunktion von T.

$$S(t) = 1 - F(t) = 1 - P(T \leq t) \tag{3.2}$$

Da es sich bei T um eine stetige Zufallsvariable handelt, kann die Verteilung auch durch eine Dichtefunktion angegeben werden:

$$f(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{3.3}$$

$$= -dF(t)/dt \tag{3.4}$$

bzw.

$$F(t) = \int_0^t f(\tau) d\tau. \tag{3.5}$$

Gebräuchlicher in der Survivalanalyse ist jedoch die Verwendung der sogenannten **Hazard-Funktion**. Sie gibt, ähnlich wie die Dichtefunktion, die Wahrscheinlichkeit an, dass zu einem

bestimmten Zeitpunkt das Ereignis eintritt, jedoch unter der Bedingung, dass die Individuen bis zu diesem Zeitpunkt überlebt haben und sich noch unter Risiko befinden. Es gilt:

$$h(t) \quad = \quad lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.6}$$

Wichtig ist dabei zu beachten, dass die Wahrscheinlichkeit durch das Zeitintervall Δt geteilt wird, es sich hier folglich um eine Rate handelt. Die Hazard-Funktion gibt eine lokale, einem bestimmten Zeitpunkt zugehörige Beschreibung des Prozesses an und kann Werte zwischen 0 und ∞ annehmen. Weiterhin gilt:

$$h(t) \quad = \quad \frac{f(t)}{S(t)}. \tag{3.7}$$

Den direkten Zusammenhang zwischen Hazard-Funktion und Survival-Funktion kann man leicht herleiten. Ausgehend von $S(t)$ folgt, dass

$$h(t) = \frac{-d}{dt} log(S(t)) \tag{3.8}$$

und andersherum erhält man durch Integration der rechten Seite und mit $S(0) = 1$ entsprechend:

$$S(t) = exp\left(- \int_0^t h(\tau) d\tau \right), \tag{3.9}$$

wobei $\int_0^t h(t) d\tau$ auch kumulierte Hazard-Funktion genannt und mit $H(t)$ bezeichnet wird.

3.3 Konstruktion der Likelihood-Funktion

Bei der Ermittlung der Koeffizienten in der Überlebenszeitanalyse wird üblicherweise die partielle Likelihood-Funktion verwendet. Es bezeichne wie zuvor U die Überlebenszeit, C die Zeit bis zur Zensierung und T die Beobachtungszeit, die für jedes Individuum als $min(U, C)$ definiert ist. Im Folgenden wird angenommen, dass sowohl die Beobachtungszeiten als auch die Zensierungszeiten unabhängig und identisch verteilt sind. Im Fall von rechtszensierten Daten gibt es zwei Möglichkeiten, vgl. [KM03]: Bei einer Zensierung nimmt die Indikatorvariable D den Wert 0 an und der Beitrag zur Likelihood-Funktion ergibt sich durch:

$$\begin{aligned} P(T, D = 0) &= P(T = C, D = 0) \cdot P(D = 0) = P(D = 0) \\ &= P(U > C) = S(C). \end{aligned}$$

Das Ergebnis ist die Survival-Funktion zum Zeitpunkt C, dass heißt die Wahrscheinlichkeit, bis zum Zeitpunkt der Zensierung überlebt zu haben.
Die zweite Möglichkeit ist, dass ein Ereignis eintritt. In diesem Fall ist der Indikator $D = 1$ und die Beobachtungszeit T entspricht der Ereigniszeit U. Es folgt:

$$\begin{aligned} P(T, D = 1) &= P(T = U | D = 1) \cdot P(D = 1) \\ &= P(U = T | U \leq C) \cdot P(U \leq C) \\ &= \left[\frac{f(t)}{1 - S(C)} \right] [1 - S(C)] = f(t). \end{aligned}$$

Der Beitrag zur Likelihood ist die Dichtefunktion, welche die approximative Wahrscheinlichkeit angibt, dass ein Ereignis zum Zeitpunkt t eintritt. Diese beiden Ausdrücke können zusammengefasst werden. Es ergibt sich:

$$P(t, D) = [f(t)]^D \, [S(t)]^{1-D} \, . \tag{3.10}$$

Im Folgenden werden nun die Paare von Zufallsvariablen (t_i, D_i), $i = 1, \ldots, n$ betrachtet. Mithilfe von $f(t_i) = h(t_i)S(t_i)$ aus Gleichung 3.7 folgt für die zugehörige Likelihood:

$$L = \prod_{i=1}^{n} [h(t_i)]^{D_i} \, exp \, [-H(t_i)] \, . \tag{3.11}$$

Die Konstruktion der Likelihood für Daten mit einer anderen Art von Zensierungsmechanismus funktioniert auf eine ähnliche Weise [KM03].

3.4 Cox-Proportional-Hazards-Modell

Die Hazard-Funktion kann parametrisch modelliert werden, zum Beispiel als Weibull-, loglogistische oder Exponentialfunktion. Die Schwierigkeit besteht jedoch darin, das richtige Modell zu wählen [KK05]. Stattdessen wird in der Überlebenszeitanalyse häufig ein semiparametrisches Modell verwendet, bei dem nur für den Einfluss der Kovariaten eine funktionale Form bestimmt wird, die Baseline-Hazard-Funktion jedoch unspezifiziert bleibt.
Das besonders verbreitete Modell mit dieser Eigenschaft ist unter dem Namen **Cox-Regression** oder **Cox-Proportional-Hazards-Modell** bekannt [Cox72]. Es ist definiert als:

$$h(t, X) \quad = \quad h_0(t) * exp(\sum \beta X_i).$$

Die Hazardfunktion besteht in diesem Modell aus zwei Faktoren. Zum einen $h_0(t)$, dem Baselinehazard, und zum anderen $exp(\beta X_i)$, mit dem der Einfluss der Kovariaten modelliert wird. Dabei bezeichnet $X = (X_1, \cdots, X_q)$ einen Vektor von q erklärenden Variablen, die sowohl kontinuierlich als auch kategoriell seinen können, und β den Vektor der zugehörigen Regressionskoeffizienten.
Hier ist im Besonderen zu bemerken, dass nur der Baseline-Hazard eine Funktion von t ist und die Annahme gemacht wird, dass der Einfluss der Kovariaten unabhängig von der Zeit ist. Durch den Exponenten ist gewährleistet, dass die geschätzten Hazards in jedem Fall nicht negativ sind, und die Rate in einem möglichen Wertebereich von $0 \leq h(t, X) < \infty$ liegt.
Das Cox-PH-Modell verdankt seine Popularität folgender Eigenschaften, die sich aus den Modellannahmen ergeben [KK05]:

- Die Ergebnisse sind in den meisten Fällen eine gute Näherung für die des korrekten parametrischen Modells.
- Auch ohne Spezifizierung von $h_0(t)$ können gute Schätzungen für die Regressionskoeffizienten, Hazard Ratios und adjustierte Survivalkurven ermittelt werden.
- Für die meisten Datensituationen ist das Cox-Modell eine sichere Entscheidung, es handelt sich um ein sehr robustes Modell.

Das **Hazard Ratio** (HR) wird dann verwendet, wenn das Risiko von verschiedenen Individuen zu einem Zeitpunkt verglichen werden soll. Es ist definiert als der Quotient der Hazard-Funktionen zweier Individuen, die sich in den Werten der erklärenden Variablen unterscheiden, bei denen aber davon ausgegangen wird, dass sie die gleiche Baseline-Hazard-Funktion besitzen. Es folgt:

$$HR = \frac{h(t, X^*)}{h(t, X)}.$$

Diese Definition gilt auch für parametrische Modelle. Der Vorteil des Cox- Modells liegt darin, dass sich für das Hazard Ratio ein konstanter Wert ergibt, da

$$\begin{aligned} HR(t) &= \frac{h_0(t) exp\left(\sum \beta X^*\right)}{h_0(t) exp\left(\sum \beta X\right)} \\ &= \exp\left(\sum_{i=1}^{q} \beta(X_i^* - X_i)\right). \end{aligned}$$

Hier kürzt sich mit $h_0(t)$ der Einfluss der Zeit aus der Gleichung heraus. Dabei muss gelten, dass die Ausprägungen der Kovariaten unabhängig von der Zeit sind. Von dieser Eigenschaft leitet sich der Name Cox-Proportional-Hazards-Modell ab.

Betrachte man zur Veranschaulichung einmal zwei Gruppen, die sich nur aufgrund einer einzelnen bivariaten Kovariate unterscheiden. Dabei stehe der Wert 1 für die Behandlungs- und 0 für die Kontrollgruppe. Mithilfe des Hazard Ratio kann nun der Effekt der Therapie im Vergleich zur Kontrolle quantifiziert werden, vgl. [KK05]. Es ergibt sich:

$$HR = \exp\left(\beta(1-0)\right) = \exp\left(\beta\right).$$

Dieser Wert bleibt für die beiden betrachteten Individuen konstant über die Zeit. Liegt das HR über 1, ist das relative Risiko für Patienten der Therapie im Verhältnis zu Kontrollgruppe erhöht, liegt der Wert unter 1, so ist das relative Risiko entsprechend verringert. Im multivariaten Modell mit verschiedenen erklärenden Variablen ist eine ähnliche Interpretation möglich. Dabei kann die Frage beantwortetet werden, welchen Einfluss eine Kovariate besitzt, wenn man davon ausgeht, dass der Effekt aller anderen Variablen als fix angenommen wird.

3.4.1 Schätzung

Partieller-Likelihood-Ansatz

Zur Schätzung der Parameter im Cox-PH-Modell kann die partielle Likelihood-Funktion verwendet werden. Wie bisher ist für eine beobachtete Stichprobe von n, für jedes Individuum j, die Information über die Beobachtungszeit T_j, ein Zensierungsindikator D_j und ein Vektor von Kovariaten oder Risikofaktoren $X_j = (X_{j1}, \ldots, X_{jp})$ bekannt. Im Folgenden bezeichnen $t_1, t_2, \ldots, t_D$ die geordneten Ereigniszeiten.
Es wird angenommen, dass, bei Kenntnis der Kovariaten X, die Zensierungen unabhängig von der Ereigniszeit sind. Außerdem soll zur Vereinfachung der Fall betrachtet werden, dass zu

jedem Zeitpunkt maximal ein Ereignis auftritt, also keine "Ties" vorkommen. Die Konstruktion der partiellen Likelihood in dem Fall, dass Bindungen vorkommen, wird zum Beispiel bei [KM03] beschrieben. Das Risk Set $R(t)$, zu einem Zeitpunkt t, ist als die Menge der Individuen definiert, die kurz vor einem Ereignis oder einer Zensierung noch unter Beobachtung stehen. Die partielle Likelihood für das Cox-Modell sieht dann folgendermaßen aus:

$$L(\beta) = \prod_{i=1}^{D} \frac{exp(\sum_{k=1}^{q} \beta_k X_{ik})}{\sum_{j \in R(t_i)} exp(\sum_{k=1}^{q} \beta_k X_{jk})}. \tag{3.12}$$

Betrachtet wird die bedingte Wahrscheinlichkeit, dass zu einem Zeitpunkt t ein bestimmtes Individuum i mit Kovariaten X_i ein Ereignis hat, gegeben, dass überhaupt ein Individuum aus dem Risk Set ein Ereignis erlebt. Die Likelihood-Funktion ist das Produkt dieser bedingten Wahrscheinlichkeiten über alle Zeitpunkte hinweg, an denen ein Ereignis auftritt. In dieser Form kommt die zeitabhängige Baseline-Hazard-Funktion nicht mehr vor, da sie sich herausgekürzt hat. An dieser Stelle ist zu bemerken, dass die obige Likelihood-Funktion nicht die genaue Verteilung der Ereignis- und Zensierungszeiten berücksichtigt, sondern nur die Reihenfolge in der sie auftreten. Nur wenn sich die Reihenfolge ändert, führt das zu anderen Risk Sets zu den Ereigniszeitpunkten und damit auch zu einer veränderten Likelihood-Funktion.
Nun können die Regressionskoeffizienten $\beta_1, \ldots, \beta_q$ geschätzt werden, die die gegebene Likelihood-Funktion $L(\beta)$ maximieren. Da es sich bei der Bildung des Logarithmus um eine monotone Transformation handelt, führt es zu dem gleichen Ergebnis, wenn stattdessen die Log-Likelihood-Funktion maximiert wird, also:

$$logL(\beta) = \sum_{i=1}^{D} \sum_{k=1}^{q} \beta_k X_{ik} - \sum_{i=1}^{D} ln(\sum_{j \in R(t_i)} exp(\sum_{k=1}^{q} \beta_k X_{jk})). \tag{3.13}$$

Die Log-Likelihood-Funktion hat den Vorteil, dass leichter die partiellen Ableitungen hinsichtlich der β's berechnet werden können. Die ML-Schätzer werden dann durch die Lösung dieser q nichtlinearen Gleichungen ermittelt. Die Statistik-Programme verwenden dazu iterative Methoden.

Profile-Likelihood-Ansatz

Eine andere Möglichkeit zur Schätzung der Parameter im Cox-Modell bietet die Profile-Likelihood-Methode. Es sei wieder für jedes Individuum die Information über die Beobachtungszeit, Zensierung und einen Vektor von Kovariaten gegeben: (T_j, D_j, X_j). Dann ergibt sich die vollständige Likelihood der Daten unter Berücksichtigung der Zensierungen durch [KM03]:

$$L(\beta, h_0(t)) = \prod_{j=1}^{n} [h_0(T_j)]^{D_j} \left[exp(\beta^T X_j)\right]^{D_j} exp(-H_0(T_j) exp(\beta^T X_j)). \tag{3.14}$$

Im Folgenden werden die Parameterschätzer als fix angesehen und die gegebene Likelihood als Funktion von $h_0(t)$ maximiert. Die zu maximierende Likelihood ist dann:

$$L_\beta(h_0(t)) = \left[\prod_{i=1}^{D} h_0(t_i) exp(\beta^T X_i)\right] exp\left[-\sum_{j=1}^{n} H_0(T_j) exp(\beta^T X_j)\right]. \tag{3.15}$$

Wie bisher bezeichnen $t_1 < \ldots < t_D$ die geordneten Ereigniszeiten, sodass der Zensierungsindikator D_i verschwindet. Nun wird zwischen Zeitpunkten mit und ohne Ereignisse unterschieden. Zu Zeitpunkten, an denen kein Ereignis eintritt, ist die Funktion maximal, wenn $h_0(t) = 0$.
Der Profile-Maximum-Likelihood-Schätzer für die Zeitpunkte $i = 1, \ldots, D$, an denen ein Ereignis eintritt, ist [KM03, Joh83]:

$$\hat{h}_{0i}(t) = \frac{1}{\sum_{j \in R(t_i)} exp(\beta^T X_j)}. \tag{3.16}$$

Mithilfe dieses Schätzers ist es möglich, die kumulative Baseline-Hazard-Funktion zu schätzen, das Ergebnis ist der sogenannte Breslow Schätzer:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} exp(\beta^T X_j)}. \tag{3.17}$$

Dieser Schätzer gilt für den Fall, dass zu jedem Zeitpunkt maximal ein Ereignis geschieht. Es gibt alternative Vorschläge zur Berechnung der partiellen Likelihood für Daten mit Bindungen zwischen den Ereigniszeiten. Die Schätzer von Breslow [Bre74] und Efron [Efr77] können "Ties" berücksichtigen und sind in den meisten gängigen Statistikprogrammen enthalten.
Setzt man den Schätzer der kumulierten Baseline-Hazard-Funktion in die vollständige Likelihood ein, so erhält man die Profile-Likelihood. Diese ist proportional zur partiellen Likelihood, die in Abschnitt 3.4.1 hergeleitet wurde. Die Koeffizientenschätzer β können damit wie gewohnt berechnet werden.

Kapitel 4

Strukturgleichungsmodelle

Das Cox-Regressionsmodell ist das Standardmodell in der Ereigniszeitanalyse. Es gibt verschiedene Erweiterungsmöglichkeiten, wie zum Beispiel die Berücksichtigung zeitabhängiger Kovariaten oder die Betrachtung multipler Endpunkte. In jedem Fall wird jedoch davon ausgegangen, dass alle Kovariaten perfekt und ohne Fehler gemessen werden können. Dies ist eine Annahme, die man gerade in der Psychologie und den Wirtschafts- und Sozialwissenschaften nur sehr selten treffen kann. Anstelle von "manifesten", dass heißt direkt messbaren Variablen, kommt es häufig vor, dass sogenannte "latente" Variablen untersucht werden sollen. Als latent bezeichnet man Variablen, die nicht direkt beobachtbar sind, sondern nur indirekt durch mehrere Items gemessen werden können. Beispiele für solche Konstrukte sind Intelligenz, Depression oder Motivation.
Im Folgenden soll eine Erweiterung des Cox-PH-Modells betrachtet werden, die es ermöglicht, latente Variablen als in der Ereigniszeitanalyse zu berücksichtigen. Zuvor soll eine Einleitung in die Modellierung mit latenten Variablen durch Strukturgleichungsmodelle in Anlehnung an die Bücher von Jöreskog und Sörbom [JS89], Schumacker und Lomax [SL04] sowie Reinecke [Rei05] gegeben werden.
Mit Strukturgleichungsmodellen, im Englischen "Structural Equation Models" oder "SEMs", werden latente Konstrukte definiert, es können die Zusammenhänge zwischen latenten Variablen untereinander und Verbindungen zwischen latenten und manifesten Variablen untersucht werden.
Im Allgemeinen wird dabei von theoriegeleiteten Hypothesen ausgegangen, die anhand der Daten einer empirischen Untersuchung überprüft werden sollen. Ausgangspunkt dieser Modelle sind Varianzen und Kovarianzen bzw. Korrelationen zwischen den manifesten Variablen. Mithilfe dieser wird überprüft, inwieweit eine implizierte Struktur zwischen den Variablen von den beobachteten Daten unterstützt wird. Um Strukturgleichungsmodelle zu verstehen, ist es notwendig, einen Blick in die Entstehung dieser Modellklasse zu werfen.

4.1 Entstehungsgeschichte

4.1.1 Regressionsanalyse

Die ersten Schritte in Richtung Strukturgleichsmodelle setzte Pearson 1896 mit der formalen Definition der linearen Regression [SNL07]. Dabei handelt es sich um ein Verfahren, dass seinen häufigen Einsatz und weite Verbreitung auch seiner Einfachheit verdankt. Im einfachen

bzw. multiplen Regressionsmodell werden respektive die Beziehungen von einer oder mehrerer unabhängiger manifester Variablen auf eine abhängige manifeste Variable untersucht. Diese Beziehung wird in einem Gleichungssystem ausgedrückt:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \epsilon_i. \tag{4.1}$$

Den üblichen Konventionen folgend werden manifeste abhängige Variablen mit y bezeichnet und manifeste unabhängige Variablen mit x. Die Indizes $i = 1, \ldots, p$ stehen für die einzelnen Beobachtungen, bei einem Stichprobenumfang p. In dieser Gleichung bezeichnen $\beta_1, \ldots, \beta_q$ die zu schätzenden Regressionskoeffizienten. Um einen konstanten Term in der Gleichung zu behalten, wird häufig x_{1i} gleich 1 gesetzt. Dieses Gleichungssystem lässt sich auch in Matrixschreibweise formulieren:

$$y = X\beta + \epsilon. \tag{4.2}$$

Hier sind y und ϵ $(p \times 1)$ Vektoren, β ist ein $(q \times 1)$ Vektor und X ist eine $(p \times q)$ Matrix. Die Regressionskoeffizienten sind unbekannt und müssen geschätzt werden, die zugehörigen Parameterschätzer werden mit $\hat{\beta}$ bezeichnet. Zur Lösung dieses Gleichungssystems wird die "Methode der Kleinsten Quadrate" verwendet. Umformen des Gleichungssystems führt auf den OLS (Ordinary Least Squares)- Schätzer für die Parameter [Rei05]:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{4.3}$$

Hier ist zu beachten, dass zur Lösung des Systems die Inverse der Matrix X gebildet werden muss. Dies beinhaltet, dass die Determinante der Matrix nicht 0 sein kann. Ist das nicht der Fall, so können Informationen nicht mehr eindeutig einer bestimmten Variable zugeordnet werden, man spricht von Multikollinearität.
Wichtig bei der linearen Regression ist es, sich der Annahmen bewusst zu sein, die dieser Methode zu Grunde liegen und auch für die darauf aufbauenden Modelle gelten. Letztlich ist selbst die Regressionsanalyse nichts anderes als eine spezielle Form eines Strukturgleichungsmodells [Gei10].
Bei der linearen Regression wird von einem linearen Zusammenhang zwischen den Parametern ausgegangen. Für die Störterme ϵ_i wird vorausgesetzt, dass ihr Erwartungswert gleich 0 ist. Es wird angenommen, dass die Störterme weder untereinander noch mit den Regressoren korrelieren, außerdem sollen sie frei von Heteroskedastizität sein. Die Störterme werden als normalverteilt angenommen. Für die Regressoren wird vorausgesetzt, dass sie linear unabhängig sind.

4.1.2 Pfadanalyse

Eine Weiterentwicklung der Regressionsanalyse gelang dem Genetiker Wright ab 1920 mit der Pfadanalyse [Rei05]. Ziel dieser Methode ist es, komplexere Beziehungen zwischen manifesten Variablen zu untersuchen. Mit diesem Verfahren können die Zusammenhänge zwischen mehreren abhängigen und unabhängigen Variablen formuliert werden.

Zur Verdeutlichung der Beziehungen zwischen den betrachteten Variablen werden Pfaddiagramme verwendet. In diesen Graphiken werden mithilfe von Pfeilen die Beziehungen zwischen

den verschiedenen Variablen verdeutlicht. Dabei gelten die folgenden Regeln, um die Darstellungsform zu vereinheitlichen [Jör02]. Die Unterscheidung zwischen latenten und manifesten Variablen ist erst im folgenden Abschnitt bei den Strukturgleichungsmodellen von Bedeutung. In der Pfadanalyse wird davon ausgegangen, dass alle Variablen fehlerfrei messbar sind.

- Latente Variablen werden von Kreisen oder Ellipsen eingeschlossen, manifeste Variablen werden von Rechtecken umschlossen. Fehlerterme werden ebenfalls in das Pfeildiagramm eingezeichnet, jedoch ohne Umrandung.

- Ein einseitig gerichteter Pfeil zwischen zwei Variablen kennzeichnet eine kausale Beziehung von einer Variablen auf die andere. Ein zweiseitig gerichteter Pfeil bedeutet, dass eine Korrelation zwischen diesen beiden Variablen besteht.

- Pfeile können mit den jeweiligen Pfadkoeffizienten beschriftet werden. Die Indizierung folgt der Regel, dass zunächst das Ziel und dann der Ursprung genannt wird.

Die Pfadanalyse formuliert die Zusammenhänge zwischen den Variablen mithilfe von Regressionsgleichungen. Es gilt weiterhin die Annahme, dass die Beziehungen zwischen den Variablen linear sind. Neben den Voraussetzungen der linearen Regression muss gelten, dass die Residuen verschiedener Regressionsgleichungen nicht miteinander korrelieren. Außerdem wird angenommen, dass die Residuen und die unabhängigen Variablen unkorreliert sind.

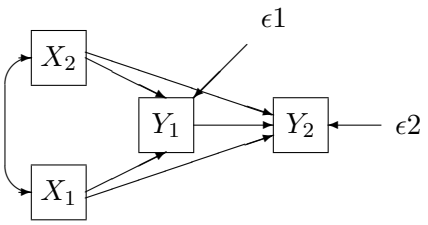

Abbildung 4.1: Pfaddiagramm

In Abbildung 4.1 ist ein Beispiel für ein Pfaddiagramm dargestellt, vgl. [Rei05]. Hier bezeichnen X_1 und X_2 exogene Variablen und Y_1 und Y_2 endogene Variablen.
Die Zusammenhänge zwischen diesen Variablen können mithilfe von Regressionsgleichungen formuliert werden. In der Literatur findet sich zur Berechnung der Pfadkoeffizienten der Ansatz der Dekomposition der Korrelationen. Für eine genauere Ausführung siehe [Rei05]. Auf diese Weise kann die Beziehungsstruktur zwischen den Variablen offengelegt werden. Die Korrelationen zwischen den betrachteten Variablen können in direkte und indirekte Effekte zerlegt werden. Für die Berechnung der Pfadkoeffizienten werden keine Originaldaten benötigt, es genügt, die Korrelationen zwischen den einzelnen Variablen zu kennen. Bei der Berechnung eines solchen Modells mit einem Computerprogramm muss nur die Korrelationsmatrix eingegeben und die Pfadstruktur spezifiziert werden.

4.1.3 Konfirmatorische Faktorenanalyse

Die Ursprünge der Faktorenanalyse liegen zu Beginn des 20. Jahrhunderts in den Arbeiten des Psychologen Charles Spearman [SL04]. Dieser entwickelte die Idee, dass stark korrelierende Items zusammengenommen eine Variable im Hintergrund messen. Bei der Faktorenanalyse wird die Annahme verlassen, dass alle Variablen manifest und direkt messbar sind. Stattdessen werden latenten Variablen untersucht, die nur indirekt beobachtbar sind. Diese werden auch "Faktoren" genannt.
Während bei der explorativen Faktorenanalyse im Vorfeld keine Vorstellungen über die Zusammenhänge zwischen Indikatoren und latenten Konstrukten vorliegen, müssen bei der konfirmatorischen Faktorenanalyse konkrete Hypothesen über die Zusammenhangsstruktur getroffen werden. In einem zweiten Schritt wird überprüft, ob diese Struktur bestätigt werden kann. Häufig wird die konfirmatorische Faktorenanalyse bei Strukturgleichungsmodellen zur Definition eines Messmodells verwendet.
Hier ist anzumerken, dass bei der konfirmatorischen Faktorenanalyse und bei Strukturgleichungsmodellen davon ausgegangen wird, dass alle betrachteten Variablen metrisches Skalenniveau besitzen. Sind die beobachteten Items kategoriell, ist dies besonders zu berücksichtigen, siehe dazu Abschnitt 4.2.2.
Im Folgenden wird auf das Verfahren der konfirmatorischen Faktorenanalyse eingegangen. Das grundlegende Modell sieht dabei wie folgt aus [SL04]:

$$y \quad = \quad \Lambda_y \eta + \epsilon. \tag{4.4}$$

Es bezeichne η einen Vektor latenter Variablen $(m \times 1)$ und y einen Vektor manifester Variablen $(p \times 1)$. Im Allgemeinen gilt $p > m$. Des Weiteren beschreibt Λ_y eine Matrix von Faktorladungen und ϵ den $(p \times 1)$ Vektor der Residuen.
Die Vorstellung bei diesem Modell ist, dass jede beobachtete Variable y eigentlich eine lineare Funktion einer oder mehrerer latenter Variablen ist. Die Faktorladungen können dabei ähnlich wie die Regressionskoeffizienten bei der linearen Regression interpretiert werden. Für eine Änderung des Faktors um eine Einheit gibt die Faktorladung die erwartete Änderung der Variable y an. Ein möglicher Fehler der Vorhersage wird mit dem Residualterm ϵ ausgedrückt.
In Gleichung 4.4 wird die Annahme getroffen, dass alle Variablen in ihrer Abweichung vom Mittelwert gemessen werden, und $E(\eta) = 0$, $E(y) = 0$ und $E(\epsilon) = 0$ gilt. Daraus resultiert die günstige Eigenschaft, dass die Kovarianzmatrix zwischen zwei dieser Zufallsvariablen als Erwartungswert der Produkte der Vektoren ausgedrückt werden kann. So gilt für die Matrix der Kovarianzen zweier Vektoren $(y_1, \cdots, y_p)$ [Lon83]:

$$Cov(y, y) = E(yy^T). \tag{4.5}$$

Die Grundlage der Berechnungen bei der Faktorenanalyse ist die Varianz-Kovarianzmatrix oder einfach Kovarianzmatrix der manifesten Variablen. Es wird zwischen der beobachteten Kovarianzmatrix einer Stichprobe S (Stichprobenmatrix) und einer zu dem spezifizierten Modell gehörigen Kovarianzmatrix Σ (Populationsmatrix) unterschieden. Für die Kovarianzmatrix des Modells gilt nach Gleichung 4.5:

$$\Sigma = E(yy^T). \tag{4.6}$$

Dabei handelt es sich um eine symmetrische $(p \times p)$ Matrix. Die Elemente der Matrix sind die Kovarianzen zwischen den einzelnen manifesten Variablen y_i und y_j. Würde es sich bei

den y zusätzlich um standardisierte Variablen mit einer Varianz von 1 handeln, so wäre Σ die zugehörige Korrelationsmatrix. Um die Beziehung zwischen der Kovarianzmatrix und den unbekannten Parametern darzustellen, kann Gleichung 4.4 für die Indikatoren y eingesetzt werden. Es folgt:

$$\Sigma = E\left[(\Lambda_y \eta + \epsilon)(\Lambda_y \eta + \epsilon)^T\right]. \tag{4.7}$$

Durch Ausmultiplizieren, Anwendung des Erwartungswertoperators und unter Berücksichtigung der Annahme, dass die manifesten Variablen y und der Fehlerterm ϵ unkorreliert sind, ergibt sich daraus die sogenannte Kovarianzgleichung:

$$\Sigma = \Lambda \Phi \Lambda^T + \Theta. \tag{4.8}$$

Hier ist die Kovarianz der latenten Variablen mit $\Phi = E\left[\eta\eta^T\right]$ und die Kovarianz der Residuen mit $\Theta = E\left[\epsilon\epsilon^T\right]$ bezeichnet. Diese Gleichung zerlegt die beobachtete Kovarianz einer Stichprobe in die unbekannten Parameter λ, Φ und Θ. Die Idee für die Schätzung ist es, die Parameter so zu wählen, dass die Abweichung der beobachteten Kovarianzmatrix von der geschätzten minimal wird.

4.2 Strukturgleichungsmodelle

Strukturgleichungsmodelle ermöglichen eine Kombination der konfirmatorischen Faktorenanalyse und der Pfadanalyse. Man kann sich diese wie eine Regression auf latenter Ebene vorstellen, mit dem Vorteil, dass Messfehler angemessen berücksichtigt werden [Gei10]. Wesentliche Beiträge auf diesem Gebiet stammen von Jöreskog (1973), Keesling (1972) und Wiley (1973). Die am breitesten verbreitete Formulierung ist die Formulierung von Jöreskog, auf der die Software LISREL (Linear Structural Relations Model) basiert. In dem Modell wird angenommen, dass alle beobachteten Variablen kontinuierlich gemessen sind. Später wurde das Modell erweitert, um auch dichotome und ordinale Indikatorvariablen berücksichtigen zu können, siehe Abschnitt 4.2.2 .

Das allgemeine Strukturgleichungsmodell besteht aus zwei Teilen, mehreren Mess- und einem Strukturmodell [SL04]. Im Messmodell wird spezifiziert, wie die latenten Variablen durch manifeste Variablen gemessen werden. Das Verfahren entspricht dem der konfirmatorischen Faktorenanalyse. Mit dem Strukturmodell werden die Beziehungen der latenten Variablen untereinander beschrieben. Betrachte man beispielsweise eine latente Variable Intelligenz. Im Messmodell wird spezifiziert, durch welche Items diese Variable bestimmt wird, dass könnten verschiedene Tests sein, die unterschiedliche Aspekte der Intelligenz untersuchen. Eine Fragestellung kann darin bestehen, den Zusammenhang zwischen Intelligenz und einer zweiten latenten Variable wie Motivation zu untersuchen, die ebenfalls durch ein Messmodell bestimmt ist. Dazu wird ein Strukturmodell spezifiziert, welches die Intelligenz als Funktion der Motivation beschreibt. Ein Residualterm gibt den Anteil der abhängigen latenten Variable an, der nicht durch diese Beziehung zu erklären ist.

Im Folgenden soll auf die konkrete Formulierung von Strukturgleichungsmodellen eingegangen werden. Bei der Modellformulierung wird zwischen verschiedenen Arten von Variablen unterschieden. Eine latente Variable, die von einer anderen latenten Variable abhängig ist, wird als endogen bezeichnet, η, ansonsten handelt es ich um eine exogene latente Variable, ξ. Für die beobachteten Variablen verhält es sich ähnlich, wie bereits bei der Regressionsanalyse werden diese mit x und y bezeichnet.

4.2.1 Modellspezifikation

Fasst man das Mess- und das Strukturmodell zusammen, ergibt sich folgendes Strukturgleichungsmodell. Im Folgenden wird das Modell nach der Formulierung von Jöreskog vorgestellt [JS89].

Strukturmodell:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{4.9}$$

Messmodell:

$$y = \Lambda_y\eta + \epsilon \tag{4.10}$$
$$x = \Lambda_x\xi + \delta \tag{4.11}$$

Dabei bezeichne η einen $(m \times 1)$ Vektor der latenten endogenen Variablen und ξ einen $(n \times 1)$ Vektor der latenten exogenen Variablen. Die manifesten Variablen werden durch einen $(p \times 1)$ Vektor y und einen $(q \times 1)$ Vektor x wiedergegeben.
Es bezeichne B eine $(m \times m)$ Koeffizientenmatrix, welche die Beziehungen zwischen den η beschreibt, und Γ $(m \times n)$ eine Koeffizientenmatrix für den direkten Einfluss von ξ auf η. Es wird davon ausgegangen, dass keine Beziehungen zwischen einer latenten endogenen Variable und sich selbst besteht, demzufolge sind alle Elemente auf der Hauptdiagonalen von B gleich 0.
Die Faktorladungen für den Zusammenhang zwischen y und η ergeben den $(p \times m)$ Vektor Λ_y, für den Zusammenhang zwischen x und ξ ist der $(q \times n)$ Vektor Λ_x gegeben.
Die Vektoren ζ $(m \times 1)$, ϵ $(p \times 1)$, δ $(q \times 1)$ geben die Residuen im Strukturmodell beziehungsweise in den Messmodellen an.

Für die Schätzung wird wie bei der Faktorenanalyse versucht, die Populationskovarianzmatrix in die unbekannten Parameter zu zerlegen. Dazu ist es notwendig, zunächst die zum Modell gehörigen Kovarianzmatrizen zu betrachten.
Die Kovarianz der latenten exogenen Variablen ξ wird in der $(n \times n)$ Matrix Φ abgebildet, die der zugehörigen Residuen ζ in der $(m \times m)$ Matrix Ψ. Die Kovarianzen der manifesten Variablen sind in der $(p \times p)$ Matrix $\Sigma_{xx} = E(xx^T)$ und der $(q \times q)$ Matrix $\Sigma_{yy} = E(yy^T)$ aufgeführt. Die Kovarianzen der zugehörigen Residuen werden durch die $(p \times p)$ Matrix Θ_ϵ und die $(q \times q)$ Matrix Θ_δ angegeben.

In diesem Modell gelten eine Reihe von Annahmen für die Beziehungen der Parameter untereinander [SNL07]. Zunächst gilt weiterhin, dass alle Variablen als Abweichungen von ihrem arithmetischen Mittel gemessen werden. Es wird davon ausgegangen, dass die Faktorladungen

und die Fehlerterme einer Gleichung miteinander unkorreliert sind und die Messfehler über die Gleichungen nicht miteinander korrelieren. Zudem gilt, dass die exogenen latenten Variablen und Messfehler unkorreliert sind. Keine der Strukturgleichungen ist redundant, d.h. $\dot{B} = (I - B)^{-1}$ existiert.

Die erste Annahme ist insbesondere von Bedeutung, weil daraus folgt, dass die Kovarianz zweier Vektoren dem Erwartungswert ihres Produktes entspricht, vergleiche Gleichung 4.5. Die Populationskovarianzmatrix des Modells sieht wie folgt aus:

$$\Sigma = E \begin{pmatrix} yy^T & yx^T \\ xy^T & xx^T \end{pmatrix}.$$

Dabei handelt es sich um eine $((p + q) \times (p + q))$ Matrix. Um diese Matrix in die unbekannten Parameter zu zerlegen, werden zunächst die manifesten Variablen x und y durch die Gleichungen aus 4.10 ersetzt. Nach Anwendung des Erwartungswertoperators und einigen Umformungen unter Berücksichtigung der Modellannahmen folgt daraus:

$$\Sigma = \begin{pmatrix} \Lambda_y Cov(\eta\eta^T)\Lambda_y^T + \Theta_\epsilon & \Lambda_y Cov(\eta, \xi)\Lambda_x^T \\ \Lambda_x Cov(\xi, \eta)\Lambda_y^T & \Lambda_x Cov(\xi\xi)\Lambda_x^T + \Theta_\delta \end{pmatrix}. \tag{4.12}$$

In diesem Ausdruck müssen nun die Kovarianzmatrizen der latenten Variablen eingesetzt werden. Diese können mit Hilfe des Strukturmodells ermittelt werden. Durch einfache Umformungen und unter Berücksichtigung der Modellannahmen ergibt sich [Lon83]:

$$\begin{array}{ll} Cov(\eta, \eta) = \dot{B}(\Gamma\Phi\Gamma^T + \Psi)\dot{B}^T & Cov(\eta, \xi) = \dot{B}\Gamma\Phi \\ Cov(\xi, \eta) = \Phi\Gamma^T\dot{B}^T & Cov(\xi, \xi) = \Phi \end{array}$$

mit $\dot{B} = (I - B)^{-1}$. Einsetzen in Gleichung 4.12 ergibt die gewünschte Form:

$$\Sigma = \begin{pmatrix} \Lambda_y\dot{B}(\Gamma\Phi\Gamma^T + \Psi)\dot{B}^T\Lambda_y^T + \Theta_\epsilon & \Lambda_y\dot{B}\Gamma\Phi\Lambda_x^T \\ \Lambda_x\Phi\Gamma^T\dot{B}^T\Lambda_y^T & \Lambda_x\Phi\Lambda_x^T + \Theta_\delta. \end{pmatrix}. \tag{4.13}$$

Diese Gleichung ermöglicht es, die Varianzen und Kovarianzen zwischen den manifesten Variablen auf die Einflüsse der einzelnen Parameter aufzuteilen. Wenn das theoretische Modell tatsächlich für alle Beziehungen zwischen den manifesten Variablen Rechnung trägt, sollte die originale Stichprobenkovarianzmatrix perfekt durch die geschätzte Populationskovarianzmatrix reproduziert werden können.

In Abbildung 4.2 ist ein Strukturgleichungsmodell abgebildet. Ein ähnliches Beispiel findet sich bei [Rei05]. Das Pfaddiagramm passt zum einleitenden Beispiel, wenn als latente abhängige Variable die Intelligenz und als latente unabhängige Variable die Motivation eingesetzt werden, die jeweils durch drei Items gemessen werden.

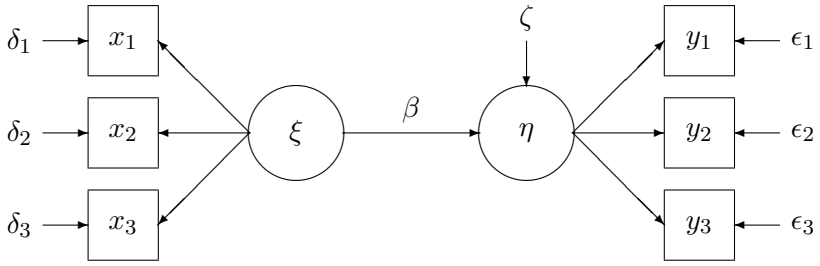

Abbildung 4.2: Beispiel für ein Strukturgleichungsmodell

Vereinfachtes LISREL Modell

Das oben formulierte Modell ist eine sehr allgemeine und weit gefasste Formulierung, die eine große Klasse von Modellen beinhaltet sowie zahlreiche Teilmodelle, mit denen gearbeitet werden kann. Für den überwiegenden Teil der Fragestellungen genügt jedoch die folgende reduzierte Form, bei der angenommen wird, dass $x = \xi$ [JS89]. Somit entfällt die latente unabhängige Variable. Stattdessen wird im Strukturmodell ein Vektor von Kovariaten x berücksichtigt.

Strukturmodell:

$$\eta = B\eta + Hx + \zeta \tag{4.14}$$

Messmodell:

$$y = \Lambda_y \eta + \Delta x + \epsilon \tag{4.15}$$

Es gelten die gleichen Bezeichnungen wie in den Gleichungen 4.9-4.10. In diesem Modell wird ein $(q \times 1)$ Vektor von Kovariaten x berücksichtigt. Es bezeichne H eine $(m \times q)$ Koeffizientenmatrix für den Einfluss der direkt gemessenen Variablen auf die latenten Variablen. Des Weiteren ist Δ eine Matrix der Parameterschätzer für die Regression von y auf x. Die Kovarianzmatrizen besitzen die gleiche Notation wie im kompletten Strukturgleichungsmodell.

4.2.2 Berücksichtigung von ordinalen Indikatorvariablen

Bei der Definition von Strukturgleichungsmodellen wurde bislang angenommen, dass die abhängigen Variablen metrisches Skalenniveau besitzen. Dies ist jedoch nicht immer der Fall. Es kommt häufig vor, dass manifeste Variablen betrachtet werden, die nur einer geringe Anzahl von Kategorien und keine äquidistanten Abstufungen zwischen den Ausprägungen besitzen. Eine Möglichkeit, ordinale Variablen in Strukturgleichungsmodellen zu verwenden, bieten

Latent-Response-Variablen (LRV). Diese Methode beruht auf der Idee, dass eine kategorielle manifeste Variable durch eine kontinuierlich Variable im Hintergrund generiert wurde. Eine andere Möglichkeit, kategorielle Variablen zur Definition von latenten Variablen zu verwenden, bietet die Item-Response-Theorie (IRT). Die beiden Modellformulierungen sind statistisch äquivalent [MA02]. Im Folgenden wird das Prinzip der LRV-Formulierung nach [Jör02] beschrieben. Mplus verwendet die LRV-Parametrisierung, IRT Schätzer können durch Transformation der Ergebnisse ermittelt werden [MA02].
Bei der LRV-Methode wird angenommen, dass einer beobachteten ordinalen Variablen u_j eine kontinuierliche latente Variable u_j^* unterliegt. Diese verborgene Variable besitzt einen Wertebereich zwischen $-\infty$ und ∞. Schwellenwerte τ unterteilen die kontinuierliche Variable u_j^* in Regionen, die den beobachteten Kategorien zugeordnet werden können.
Es wird angenommen, dass eine Indikatorvariable u_j insgesamt s Kategorien besitzt, $v = 1, 2, \cdots, s$, dann folgt:

$$u_j = v \;\Leftrightarrow\; \tau_{v-1} < u_j^* < \tau_v. \tag{4.16}$$

Dabei gilt für die Schwellenwerte:

$$-\infty < \tau_0 < \tau_1 < \cdots < \tau_s = +\infty.$$

Im Prinzip kann jede stetige Verteilung für u_j^* gewählt werden. Üblich sind als Verteilungsannahmen entweder die Normalverteilung oder die logistische Verteilung [MA02].
Die Wahrscheinlichkeit für die Kategorie v der abhängigen Variablen u_j ergibt sich durch:

$$\pi_v = P(u_j = v) = P(\tau_{v-1} < u_j^* < \tau_v) = F(\tau_v) - F(\tau_{v-1}). \tag{4.17}$$

Wobei $F(.)$ eine kumulierte Verteilungsfunktion ist, also üblicherweise die der Normalverteilung Φ. Formt man die Gleichung um, so können die Schwellenwerte zwischen den verschiedenen Kategorien berechnet werden [Jör02]

$$\tau_v = F^{-1}(\pi_1 + \pi_2 + \cdots + \pi_v), \tag{4.18}$$

für $v = 1, \cdots, s-1$. Hier bezeichnet F^{-1} das Inverse der Funktion. Eine Schätzung für die Schwellenwerte τ_v kann erzielt werden, indem man für die Wahrscheinlichkeiten $\pi_1, \pi_2, \cdots, \pi_v$ die prozentualen Anteile einsetzt, mit denen die jeweiligen Kategorien aufgetreten sind [Jör02]. Auf diese Weise ergeben sich die Schätzer $\hat{\pi_v}$ für die Schwellenwerte der einzelnen Kategorien von u_j.

Die Formulierung kann erweitert werden, um den Einfluss von Kovariaten x auf die kontinuierliche Variable u_j^* zu untersuchen. Es bezeichne x einen $(q \times 1)$ Vektor von Kovariaten. Es wird eine Regression der Latent-Response-Variablen u_j^* auf x definiert [Jör02]:

$$u_j^* = \alpha' + \gamma' x + z, \tag{4.19}$$

wobei α' den Achsenabschnitt, γ' einen Vektor von Regressionskoeffizienten und z einen Fehlerterm bezeichnet. Nimmt man an, dass der Fehlerterm z normalverteilt ist mit einem Mittelwert von 0 und einer Varianz ς^2, so ergibt sich ein Probit Modell. Es folgt, dass u_j^* bedingt auf x normalverteilt ist:

$$u_j^* \sim N(\alpha' + \gamma' x, \varsigma^2). \tag{4.20}$$

Die Wahrscheinlichkeit für die Ausprägung einer bestimmten Kategorie v oder niedriger ist dann, bedingt auf x:

$$\pi_v = \Phi\left(\frac{\tau_v - \alpha' - \gamma' x}{\varsigma}\right). \tag{4.21}$$

Dabei gelte $v = 1, \cdots, s-1$. Die lineare Regression von u_j^* auf x ist äquivalent zu einer Probit Regression für u_j. Alternativ kann man annehmen, dass u^* logistisch verteilt ist. In diesem Fall entspricht die lineare Regression für u^* einer logistischen Regression für u_j [Jör02].
Die LRV-Formulierung ist deshalb praktisch, da sie es ermöglicht, die linearen Beziehungen zwischen den u_j^* und den anderen Variablen im Strukturgleichungsmodell beizubehalten. Muthen formuliert in [Mut83] ein allgemeines Modell, dass auch dichotome und geordnete kategorielle Indikatorvariablen berücksichtigen kann. Das Strukturgleichungsmodell ist wie in den Gleichungen 4.14-4.15 spezifiziert. An der Stelle der manifesten Variablen y steht jedoch eine Latent-Response-Variable y^*. Im Fall von kontinuierlichen Variablen y gelte $y = y*$. Im Fall von kategoriellen y wird y^* wie oben definiert.

4.2.3 Identifikation

Ein Aspekt, der in jedem Strukturgleichungsmodell beachtet werden muss, um die Identifikation des Modells zu gewährleisten, ist die Metrik der latenten Variablen. Ohne eine eindeutige Skalierung ist es nicht möglich, gleichzeitig die Faktorladungen und die Varianzen einer latenten Variable zu schätzten. In der Praxis wird daher meistens eine der Faktorladungen auf 1 gesetzt. Dadurch wird für die latente Variable die Maßeinheit der zugehörigen manifesten Variable abzüglich ihres Fehlerterms verwendet. Es macht Sinn, hier die Variable auszuwählen, welche den latenten Faktor am besten repräsentiert [Büh11]. Die andere Möglichkeit, die Skalierung festzulegen, ist die Standardisierung der latenten Variablen. Dies kann Sinn machen, wenn die betrachteten Items sehr unterschiedliche Varianzen aufweisen [Gei10]. In diesem Fall kann der Faktor später als Effektstärkemaß interpretiert werden.
Nach der Modellspezifikation ist es möglich, Restriktionen in das Modell einzufügen, die sich aus theoretischen Überlegungen ergeben. Die Parameter werden dann unterschieden in fixe Parameter, die bereits im Voraus auf einen bestimmten Wert gesetzt wurden, beschränkte Parameter, die in einem bestimmten Wertebereich liegen müssen, und freie Parameter, die noch vollkommen unbekannt sind [SL04].
Bevor die Schätzung der Parameter im Strukturgleichungsmodell vorgenommen wird, muss das Problem der Modellidentifikation gelöst werden. Bei der Lösung von Strukturgleichungsmodellen wird, mathematisch gesehen, nichts anderes getan, als simultan eine Reihe linearer Gleichungen zu lösen. Es stellt sich also die Frage, ob auf Basis der beobachteten Stichprobenmatrix und des theoretischen Modells, das durch die Populationsmatrix spezifiziert ist, überhaupt ein eindeutiges Set von Parametern gefunden werden kann. In der praktischen Anwendung kann diese Frage schnell mithilfe verschiedener Regeln überprüft werden, dazu gehören die sogenannte “Order Condition” und die “Rank Condition” [SL04]. Bei der ersten Bedingung wird überprüft, ob die Zahl eindeutiger Werte in der Stichprobenmatrix S größer oder gleich der Anzahl freier zu schätzender Parameter im Modell ist. Diese Bedingung ist notwendig, aber nicht hinreichend, daher sollte in einem zweiten Schritt die “Rank Condition” geprüft werden. Diese Regel untersucht, ob die Determinante der Matrix ungleich 0 ist. Ist diese Bedingung erfüllt, so ist das Strukturgleichungsmodell mit hoher Wahrscheinlichkeit identifiziert.

4.2.4 Schätzung

Nachdem die Populationskovarianzmatrix in Abhängigkeit der unbekannten Parameter ausgedrückt wurde, müssen diese nun geschätzt werden.
Die Idee der Parameterschätzung ist die gleiche wie bei der Faktorenanalyse. Die unbekannten Parameter werden so gewählt, dass die geschätzte Kovarianzmatrix $\hat{\Sigma}$ die Stichprobenmatrix S im Rahmen der Restriktionen bestmöglich wiedergibt [SL04]. Als Maß für den Grad der Anpassung werden sogenannte Diskrepanzfunktionen oder Fit Funktionen $F(S, \hat{\Sigma})$ verwendet. Einige dieser Funktionen sind die Maximum-Likelihood (ML) Diskrepanzfunktion, Unweighted-Least-Squares(ULS)-, General-Least-Squares(GLS)- und Ordinary-LeastSquares(OLS)- Diskrepanzfunktionen. Im Folgenden soll die Maximum-Likelihood(ML)-Diskrepanzfunktion, die vermutlich gebräuchlichste der Methoden, vorgestellt werden. Die formale Beschreibung ist nicht erschöpfend, vielmehr wird versucht, die allgemeine Idee der Diskrepanzfunktion zu vermitteln. Die ML-Fitfunktion ist gegeben durch [Rei05]:

$$F_{ML} = log(Det(\hat{\Sigma})) + tr(S\Sigma^{-1}) - log(Det(S)) - (p + q). \tag{4.22}$$

Hierbei bezeichne Det die Determinante und $tr()$ die Spur einer Matrix. p und y geben die Anzahl manifester Variablen x und y an. Damit die ML-Fit-Funktion geschätzt werden kann, dürfen die Determinanten von Σ und S nicht 0 sein, sonst sind die entsprechenden Matrizen singulär und können nicht invertiert werden. Das Model besitzt dann einen perfekten Fit an die Daten, wenn die Stichprobenmatrix exakt mit der Populationsmatrix übereinstimmt. Ist das Modell überidentifiziert, so können die Parameterschätzer mithilfe einer iterativen Prozedur berechnet werden. Dabei wird die Diskrepanzfunktion schrittweise durch eine immer bessere Schätzung der Parameter minimiert.
Bei der Maximum-Likelihood-Schätzung wird angenommen, dass die manifesten Variablen multivariat normalverteilt sind. Wenn diese Annahme erfüllt ist, zeichnen sich die Schätzer durch ihre asymptotische Konsistenz und Effizienz aus. Außerdem können Signifikanztests für den Gesamtmodellfit durchgeführt werden. Die Maximum-Likelihood-Schätzer sind skaleninvariant. Die Schätzung der Parameter auf Basis der Korrelationsmatrix führt zu den gleichen Ergebnissen wie die Berechnung mithilfe der Kovarianzmatrizen [Rei05]. In der Praxis können für die Lösung von Strukturgleichungsmodelle Programme wie LISREL, Mplus, Amos und EQS verwendet werden. Details zur Schätzung eines Strukturgleichungsmodells in Mplus sind bei [Mut04b] gegeben.

4.2.5 Modellprüfung

Nachdem das Strukturgleichungsmodell spezifiziert und geschätzt wurde, gilt es das resultierende Modell zu überprüfen. Im Folgenden sollen eine Reihe von Verfahren und Modellgütemaße vorgestellt werden, mit denen der Fit des Modells an die beobachteten Daten evaluiert werden kann. Die Übersicht ist keineswegs vollständig, sondern orientiert sich an den Verfahren, die in Kapitel 5 dieser Arbeit verwendet werden, um das Survivalmodell mit latenten Faktoren zu überprüfen.
Zum einen gibt es die Möglichkeit, die einzelnen Parameter direkt zu betrachten. Im ersten Schritt sollte analysiert werden, ob die Größe und Richtung der geschätzten Parameter plausibel ist. Außerdem kann mithilfe der Wald-Statistik überprüft werden, ob der Einfluss einzelner Parameter signifikant ist [SNL07]. Dieser Test nach A. Wald (1943) untersucht, ob

ein Parameterschätzer ungleich 0 ist und in das Modell einbezogen werden sollte. Ist der Wald-Test nicht signifikant, so kann die zugehörige Variable aus dem Modell entfernt werden. Es sollte beachtet werden, dass die statistische Signifikanz abhängig von der Stichprobengröße ist. Daher sollten immer auch theoretische Überlegungen bei der Bewertung von Parametern berücksichtigt werden.
Des Weiteren kann nach der Anpassung eines Strukturgleichungsmodells die standardisierte oder unstandardisierte Residualmatrix betrachtet werden. Diese berechnet sich aus der Differenz der Korrelationsmatrix aller Variablen vor und nach der Modellanpassung. In der Residualmatrix ist zu erkennen, bei welchen Variablen es eventuelle Abweichungen gibt. Ein großes Residuum bei einer Variablen weist auf Fehlspezifikationen hin, die modifiziert werden sollte. Liegen insgesamt hohe Residuen vieler Variablen vor, so ist das ein Hinweis darauf, dass das ganze Modell falsch spezifiziert wurde.
Zusätzlich gibt es eine Reihe von globalen Fitindizes, die sich auf die Anpassungsgüte des gesamten Modells beziehen und Vergleiche zwischen verschiedenen Modellen ermöglichen.

X^2-Statistik

Ein grundlegendes Maß zur Beurteilung der allgemeinen Anpassung eines Modells ist die X^2-Statistik. Bei der Maximum-Likelihood-Schätzung ist dieses Maß als das $(n-1)$-fache des minimalen Wertes der zugehörigen Fit-Funktion definiert [SL04]. Die X^2-Quadrat-Statistik trifft eine Aussage darüber, ob sich die Stichprobenmatrix von der geschätzten Populationsmatrix unterscheidet. Ein signifikanter X^2-Wert indiziert einen Unterschied. Ist der Unterschied nicht signifikant, spricht das dafür, dass sich die Matrizen ähnlich sind und das theoretische Modell die beobachteten Varianz-Kovarianzstrukturen gut nachbildet.
Die X^2-Statistik besitzt als Gütekriterium jedoch einige gravierende Nachteile. Das Kriterium ist sehr sensitiv bezüglich der Stichprobengröße, bei steigendem N wird automatisch häufiger ein signifikanter Unterschied beobachtet. Ein weiteres Problem besteht darin, dass das Kriterium sensitiv auf Abweichungen von der multivariaten Normalverteilung der manifesten Variablen reagiert. In der Praxis wird selten die X^2-Statistik selbst als Modelgütemaß verwendet, sondern Modifikationen dieser Größe. Für Daten, bei denen der Verdacht besteht, dass die Annahme der multivariaten Normalverteilung nicht erfüllt ist, können sogenannte robuste X^2-Tests verwendet werden, wie die Satorra-Bentler-X^2-Statistik. Dabei handelt es sich um eine adjustierte X^2-Statistik, die versucht, den Bias zu korrigieren, der durch Schiefe in der Verteilung der Daten entsteht [Mut04b].

Likelihood-Ratio-Test

Für den Vergleich von zwei Modellen kann der Likelihood-Ratio-Test verwendet werden. Der Test ist auch unter der Bezeichnung X^2-Differenz-Test bekannt. Mit dem Likelihood-Ratio-Test kann ein Ausgangsmodell, mit einer Reihe von Parametern, und ein zweites Modell, in dem einigen dieser Parameter Restriktionen auferlegt wurden, verglichen werden. Diese Restriktionen können zum Beispiel so aussehen, dass mehrere Parameter auf 0 gesetzt werden. Man spricht in diesem Fall von ineinander geschachtelten oder genesteten Modellen. Auf diese Art und Weise kann untersucht werden, ob die ausgeschlossenen Variablen einen statistisch signifikanten Einfluss auf das Modell haben.

Zunächst werden dazu beide Modelle berechnet und die jeweiligen X^2-Statistiken notiert. Das unrestringierte Modell hat dabei immer einen größeren Wert, mit dem Test wird überprüft, ob die Änderung zum restringierten Modell signifikant ist. Die verwendete Teststatistik berechnet sich wie folgt [Büh11]:

$$LR = X_r^2 - X_u^2.$$

Es bezeichne X_r^2 die X^2-Statistik des restringierten Modells und X_u^2 die des unrestringierten. Die Differenz ist wiederum X^2-verteilt mit $(df_r - df_u)$ Freiheitsgraden. Erzielt der Likelihood-Ratio-Test ein signifikantes Ergebnis, so sollte das unrestringierte Modell vorgezogen werden. Der Likelihood-Ratio-Test ist allerdings, genau wie die X^2-Statistik, sensitiv bezüglich der Stichprobengröße. In großen Stichproben können unbedeutende Differenzen als signifikant gefunden werden, während in kleinen Stichproben bedeutende Abweichungen unerkannt bleiben können. Ist die Annahme der multivariaten Normalverteilung der manifesten Variablen nicht erfüllt, so kann der Sartorra-Bentler-X^2-Differenz-Test verwendet werden. Das Vorgehen zur Berechnung des Sartorra-Bentler-X^2-Differenz-Tests ist in [SB01] beschrieben, siehe auch [Mut04b].

CFI

Ein weiteres Gütemaß ist der Comparative Fit Index (CFI). Dieses Maß vergleicht den Fit des berechneten Modells mit dem eines Nullmodells. So wird ein Modell bezeichnet, in dem alle Variablen unkorreliert sind. Der CFI ist wie folgt definiert [SL04]:

$$CFI = 1 - \frac{X_n^2 - df_n}{X_0^2 - df_0}. \tag{4.23}$$

Hier bezeichnen X_n^2 die X^2-Statistik des zu untersuchenden Modells und X_0^2 die X^2-Statistik des Nullmodells, mit den zugehörigen Freiheitsgraden df_n und df_0. Der Comparative Fit Index liegt definitionsgemäß zwischen 0 und 1. Der Fit des Modells ist umso besser, je näher der Wert an 1 liegt, häufig wird als Schwellenwert für einen guten Fit der Wert 0.95 genannt [Gei10].

AIC

Das AIC, Akaikes Informationskriterium (1971), besitzt den Vorteil, dass es auch für den Vergleich von Modellen verwendet werden kann, die nicht ineinander geschachtelt sind, sofern diese auf dem gleichen Datensatz beruhen. Die Berechnung erfolgt wie folgt [Aka87]:

$$AIC = X^2 + 2k, \tag{4.24}$$

wobei X^2 die X^2-Statistik und k die Anzahl freier, im Sinne von zu schätzenden Parametern im statistischen Modell bezeichnet. Das AIC als Maß für die Güte des Modells kann heuristisch auch so erklärt werden, dass die X^2-Statistik für eine erhöhte Komplexität des Modells bestraft wird [SNL07]. Der absolute Wert des AIC hat keine Aussage über die Güte des Modells, sondern ist nur im Vergleich mit anderen Modellen für den gleichen Datensatz

zu interpretieren. Das Modell mit dem niedrigeren AIC ist das bessere Modell. Es gibt keine Tests dafür, ob der Unterschied zwischen zwei Modellen signifikant ist, aber es existieren verschiedene Regeln für die Interpretation der Ergebnisse.

BIC

Das Bayesianische Informationskriterium (BIC) nach G. Schwarz (1987) ist wie folgt definiert [Sch78]:

$$BIC = X^2 + k * ln(N), \tag{4.25}$$

wobei X^2 die X^2-Statistik, k die Anzahl freier Parameter im statistische Modell und N die beobachtete Stichprobengröße bezeichnet. Das BIC erhöht den Strafterm mit steigendem Stichprobenumfang.
Eine Modifikation dieses Wertes ist das für die Stichprobengröße adjustierte BIC. Hier wird $ln(N)$ durch $ln[(N+2)/24]$ ersetzt [Mut04b]. Damit wird der Strafterm für zusätzliche Parameter etwas geringer als beim einfachen BIC. Das Modell mit dem kleinsten BIC ist vorzuziehen. Wie beim AIC handelt es sich hier nicht um absolute Maße. Das BIC ist nur im Vergleich von Modellen zu verwenden, die auf dem gleichen Datensatz beruhen. Nylund [NAM07] empfiehlt, auf der Basis von Simulationsstudien, zum Vergleich von Latenten-Klassen-Modellen und Faktor-Mixture-Modellen insbesondere das BIC.

RMSEA

Der RMSEA (Root-Mean-Square-Error of Approximation) ist ein weiteres Maß zur Beurteilung des Modellfits, das wie folgt definiert ist, vgl. [Büh11]:

$$RMSEA = \sqrt{\frac{X^2 - df}{N * df}}. \tag{4.26}$$

Hier bezeichnet X^2 die X^2-Statistik, N die beobachtete Stichprobengröße und df die Anzahl von Freiheitsgeraden. In dem Fall, dass $df > X^2$ wird der Zähler auf 0 gesetzt. Der RMSEA betrachtet die Abweichung der Stichprobenmatrix von der geschätzten Populationsmatrix. Der Wert des Maßes ist umso höher, je größer der Unterschied ausfällt und desto komplexer das Modell ist. Der RMSEA liegt zwischen 0 und 1. Eine gute Anpassung ist gegeben, wenn der Wert kleiner oder gleich 0.05 ist.

Kapitel 5

Survivalanalyse mit einem latenten Faktor

5.1 Modellspezifikation

Im Folgenden wird die Cox-Regression mit einem latenten Faktor nach der Formulierung von Muthen [AMM06] vorgestellt. Diese Formulierung ist eine Verallgemeinerung des Modells von Larsen (2005) [Lar05], das ausschließlich binäre Indikatoren des latenten Faktors berücksichtigt. Das Modell bietet die Möglichkeit, bei der Überlebenszeitanalyse den Einfluss einer oder mehrerer kontinuierlicher latenter Variablen auf die Zeit bis zu einem Ereignis oder einer Zensierung zu untersuchen. Zur Modellierung des Survival-Prozesses wird das Cox-PH-Modell verwendet. In der Hazard-Funktion wird neben den manifesten Variablen ein latenter Faktor als Kovariate berücksichtigt. Ein Strukturgleichungsmodell beschreibt die Beziehungen zwischen den manifesten und latenten Variablen. Die Schätzung der Parameter geschieht simultan mithilfe der Maximum-Likelihood-Methode.
Es bezeichne T_i die Ereigniszeitvariable zu jedem Individuum i. In dem Fall, dass die manifesten Variablen kategorielles Skalenniveau besitzen, wird die Latent-Response-Variablen-Formulierung verwendet und eine kontinuierliche Variable y^* im Hintergrund definiert [Jör02, Mut04b]. Es bezeichne y_{ji} die j-te kategorielle abhängige Variable mit $v = 1, \ldots, s$ geordneten Kategorien. Dann gilt:

$$y_{ji} = s \quad \Leftrightarrow \quad \tau_v < y^*_{ji} < \tau_{v+1}.$$

Für eine kontinuierliche Variable gelte $y^*_{ji} = y_{ji}$. Das Modell besitzt die folgende Form:

$$y^*_i = \Lambda\eta_i + \epsilon_i \tag{5.1}$$
$$\eta_i = B\eta_i + Hx_i + \zeta_i \tag{5.2}$$

$$h_i(t) = h_0(t) * exp(\beta x_i + \kappa\eta_i). \tag{5.3}$$

Die erste Gleichung beschreibt das Messmodell. Hier bezeichnet y^*_i den Vektor aller abhängigen Variablen. Es wird angenommen, dass die zugehörigen latenten Variablen η_i normalverteilt sind. Die zweite Gleichung gibt das Strukturmodell an. Eine latente Variablen kann durch andere latente Variablen oder einen Vektor von Kovariaten x_i beeinflusst werden. Hier ist Λ die Matrix der Faktorladungen, B die Koeffizientenmatrix für die Beziehungen zwischen den

latenten Variablen und H die Koeffizientenmatrix für den direkten Einfluss von x auf η.
Es wird angenommen, dass die Residuen ϵ und ζ normalverteilt sind, den Mittelwert 0 besitzen und nicht mit anderen Variablen korrelieren.
Die untere Gleichung gibt die Hazard-Funktion $h(t)$ an, die neben den manifesten Kovariaten x zusätzlich von den latenten Variablen η beeinflusst werden kann. Die zugehörigen Koeffizientenschätzer sind mit β und κ gegeben. Die Hazard-Funktion beinhaltet keinen Achsenabschnittsschätzer, da dieser in dem Modell nicht identifizierbar wäre.

5.1.1 Schätzung

Es wird bei der Cox-PH-Regression von nicht informativen, rechtszensierten Daten ausgegangen. Damit liegen zu jedem Individuum i, zusätzlich zu der Ereigniszeitvariablen T_i , ein Zensierungsindikator D_i vor. Die vollständige Likelihood der Daten im Cox-Modell unter Berücksichtigung der Zensierungen wird wie in Abschnitt 3.4.1 berechnet. Die Baseline-Hazard-Funktion wird in dem Modell mit dem Profile-Likelihood-Ansatz als komplett unrestringierte Stufenfunktion geschätzt [AMM06]. Die Definition der Likelihood für das Mess- und Strukturmodell ist im technischen Appendix von Mplus beschrieben [Mut04b].
Für die Schätzung des Survivalmodells mit latenten Variablen wird die gemeinsame Likelihood der beobachteten Variablen $[T_i, D_i, y_i, |x_i]$ betrachtet. Die Schätzung geschieht mit der Maximum-Likelihood-Methode unter Verwendung des EM-Algorithmus. Details zu dem Verfahren sind bei [Mut04b] und [Lar05] zu finden. Nähere Erläuterungen zur Verwendung des EM-Algorithmus sind bei [DLR77] gegeben.

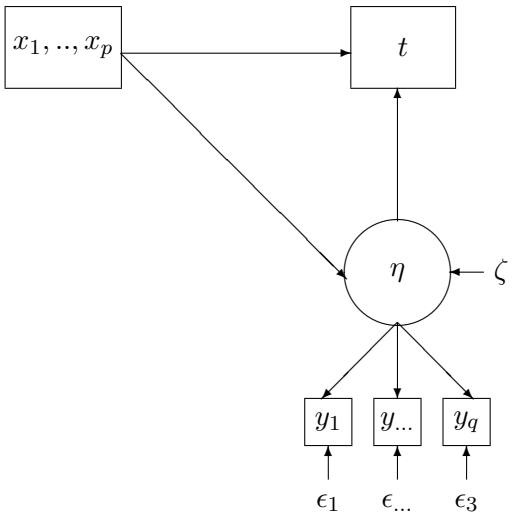

Abbildung 5.1: Schematisches Pfaddiagramm für das Survivalmodell mit einer latenten Variable

5.1.2 Modellanpassung

In Anlehnung an die Artikel von Larsen [Lar05] und Muthen [MAB+09] soll die Anpassung des Modells in mehreren Schritten geschehen. Zunächst werden unabhängig voneinander ein

latentes Variablenmodell und ein Survivalmodell angepasst. Im zweiten Schritt werden die Modelle kombiniert. Auf diese Weise kann im gesamten Modell nachvollzogen werden, wie stark sich die Definition der Faktoren durch die Berücksichtigung von Kovariaten und der Ereigniszeit verändert.
Im ersten Schritt werden für die gemessenen Items mehrere Faktorenanalysen berechnet. Mit verschiedenen Modellgütemaßen, wie dem BIC oder CFI, wird das Modell mit dem besten Fit ermittelt. Dabei können und sollten auch inhaltliche Gesichtspunkte berücksichtigt werden, sodass die resultierenden Faktoren eine sinnvolle Interpretation besitzen.
Im folgenden Schritt kann untersucht werden, ob es Kovariaten gibt, die einen signifikanten Einfluss auf die Definition des latenten Faktors bzw. der latenten Faktoren besitzen.
Gleichzeitig wird das Survivalmodell angepasst. Es sollte überprüft werden, ob die Proportional-Hazards-Annahme für alle Kovariaten erfüllt ist. Dies kann z.B. mithilfe der Schönfeld Residuen geschehen, siehe [KK05]. Ist die Annahme für eine oder mehrere Variablen nicht erfüllt, so können diese in der Cox-Regression zeitabhängig modelliert werden. Auch hier können Fitmaße helfen, das optimale Modell zu finden.
Wenn das latente Variablenmodell gewählt ist, kann es mit der Cox-Regression kombiniert werden. Ausgehend vom gesättigten Modell werden nun schrittweise die Kovariaten entfernt, deren Einfluss auf das Ereignis nicht signifikant ist. In dem Modell können auch Korrelationen der Kovariaten untereinander berücksichtigt werden. Korrelationen werden aus dem Modell entfernt, wenn sie nicht signifikant sind oder sich die Modellgütemaße (AIC, BIC) verbessern, nachdem sie aus dem Modell genommen wurden.
Es gibt verschiedene Möglichkeiten, den Fit des finalen Modells zu überprüfen. Zunächst können eine Reihe von Maßen zur Beurteilung der Modellgüte herangezogen werden, wie der CFI. Des Weiteren werden bei Larsen [Lar05] erweiterte Modelle mit zusätzlichen Kovariaten angepasst. Mithilfe des Likelihood-Ratio-Tests können diese Modelle mit dem finalen Modell verglichen werden.

5.2 Ein Anwendungsbeispiel: Herzfrequenzvariabilität als Prädiktor für kardiale Mortalität

Um die vorgestellte statistische Methode zu verdeutlichen, wurde ein Anwendungsbeispiel gewählt. Es handelt sich dabei um ein Beispiel aus der Medizin, vom Prinzip wäre das Verfahren auf beliebige andere Bereiche übertragbar, in denen mit Cox-PH-Modellen gearbeitet wird. Historisch gesehen finden sich die Problemstellungen, die mit Strukturgleichungsmodellen gelöst werden, insbesondere im Bereich der Sozialwissenschaften und der Psychologie.

Für diese Arbeit wird ein Routinedatensatz aus einer kardiologischen Klinik verwendet. Untersucht werden soll der prognostische Wert der Herzfrequenzvariabilität (HRV) auf das Risiko der Patienten, an einem koronaren Tod zu versterben. Die Herzfrequenzvariabilität eines Patienten ist nicht direkt messbar. Es existiert eine Vielzahl von Maßen, die jeweils Teilaspekte des Verhaltens abbilden. Diese Parameter beschreiben beispielsweise den mittleren Abstand zwischen zwei aufeinanderfolgenden Herzschlägen (sogenannte RR-Intervalle) oder die Standardabweichung dieser Intervalle.
In dieser Arbeit soll darauf verzichtet werden, näher auf die medizinischen Zusammenhänge einzugehen. Nähere Information zur Herzfrequenzvariabilität als Marker für kardiale Mortalität sind in [ESC96] zu finden.

Insgesamt liegen Daten von 508 Patienten vor. Die mittlere Follow-Up Zeit beträgt 3.25 Jahre. In dieser Zeit wurden 42 Ereignisse beobachtet. Die Daten wurden bereits in einer Arbeit von 2000 (unveröffentlicht) mit Standardverfahren (u.a. Clusteranalyse und Cox-Regressionen) analysiert. Der Artikel empfiehlt zehn verschiedene Parameter zur Messung der Herzfrequenzvariabilität. Zusätzlich wird die Kovariate ventrikuläre Extrasystolen (lves) aufgenommen, von der bekannt ist, dass sie eine starke Assoziation mit dem Ereignis Koronartod besitzt. Die zehn HRV-Parameter sind kontinuierlich gemessen. Da es sich um ganz unterschiedliche Maße handelt, variieren die Skalen sehr stark. Hier werden die vorgeschlagenen Transformationen der vorausgehenden Untersuchung übernommen. Zusätzlich werden die Parameter *inslml* und *islml* transformiert. Ob eine Logarithmische (l)-, Wurzel (r)- oder Kehrwerttransformation (i) angewendet wurde, um Normalverteilung zu erreichen, ist an dem ersten Buchstaben des Variablennamens zu erkennen.

Parameter	Mittelwert	SD
nmeannn	883.99	146.19
rhf	9.78	5.34
lfn	0.72	46.11
ef	46.11	13.42
rnspsdnn	5.21	1.13
nhrvidx	35.37	13.99
rspsdnn	8.75	1.66
rlf5	1.04	0.52
inslml	1.19	0.33
islml	1.33	0.19
lves	1.76	0.92

Tabelle 5.1: Übersicht der HRV-Parameter

Standardmäßig lassen sich die Effekte der Variablen auf die Überlebenszeit mit univariaten und multiplen Cox-PH-Modellen untersuchen. Das Ergebnis sind adjustierte oder unadjustierte Hazard Ratios für die einzelnen Parameter. Im folgenden Abschnitt soll diese Methode auf den Datensatz angewendet werden. Die Ergebnisse dienen als Ausgangspunkt für die nachfolgenden Modelle. Um einen Vergleich mit den Ergebnissen aus späteren Analysen zu ermöglichen wurden die Einflussvariablen z-standardisiert.

5.2.1 Analyse mit dem Cox-PH-Modell

Im ersten Schritt wird mit Hilfe der Schönfeld-Residuen [KK05] untersucht, ob die Annahme proportionaler Hazards für die einzelnen Parameter erfüllt ist. Das Ergebnis ist, dass die PH-Annahme für alle betrachteten Variablen angenommen werden kann.
Tabelle 5.2 und 5.3 zeigen die Ergebnisse der univariaten und der multiplen Cox-Regressionen. In der univariaten Cox-Regression sind nahezu alle Parameter signifikant mit der Ereigniszeit assoziiert. Die berechneten Hazard Ratios geben einen Schätzer für den Einfluss der einzelnen Parameter. Eine höhere Herzfrequenzvariabilität bedeutet in jedem Fall ein verringertes

Parameter	Schätzer	SE	p-Wert	Haz. Ratio
znmeannn	-0.68	0.21	0.001	0.51
zrhf	-0.30	0.26	0.250	0.74
zlfn	-0.72	0.11	<0.001	0.49
zef	-1.17	0.18	<0.001	0.31
zrnspsdnn	-0.43	0.20	0.026	0.65
znhrvidx	-0.51	0.20	0.012	0.60
zrspsdnn	-0.66	0.18	<0.001	0.52
zrlf5	-1.07	0.26	<0.001	0.34
zinslml	-0.55	0.39	0.154	0.58
zislml	-0.73	0.17	<0.001	0.48
zlves	0.81	0.15	<0.001	2.25

Tabelle 5.2: Univariate Cox-Regressionen

Parameter	Schätzer	SE	p-Wert	Haz. Ratio
znmeannn	-0.23	0.22	0.290	0.79
zrhf	0.91	0.38	0.016	2.48
zlfn	0.02	0.21	0.935	1.02
zef	-0.81	0.21	<0.001	0.44
zrnspsdnn	0.56	0.29	0.050	1.75
znhrvidx	0.16	0.37	0.662	1.17
zrspsdnn	-0.64	0.29	0.026	0.53
zrlf5	-1.11	0.64	0.084	0.33
zinslml	0.34	0.19	0.072	1.40
zislml	-0.46	0.22	0.036	0.63
zlves	0.54	0.16	0.001	1.72

Tabelle 5.3: Multiple Cox-Regression

Risiko. Bei der multiplen Cox-Regression zeigen die Hälfte der Parameter einen signifikanten Einfluss auf die Ereigniszeit, zwei weitere verfehlten das Signifikanzniveau nur knapp. Die zugehörigen Parameterschätzer unterscheiden sich in ihrer Richtung. Adjustiert für die anderen Variablen ist bei einigen HRV-Parametern ein Anstieg um 1 Standardabweichung (SD) mit einem erhöhten, bei anderen mit einem verringerten Risiko assoziiert. Der Einfluss der Kovariate zlves auf die Überlebenszeit ist auch unter Berücksichtigung aller HRV-Parameter hoch signifikant. Ein Anstieg um eine SD ist mit einer erhöhten Mortalität von 72% assoziiert.

Die Cox-Regression vermag es jedoch nicht vollständig, den Zusammenhang zwischen der Herzfrequenzvariabilität und der kardialen Mortalität aufzuklären. Die einzelnen Parameter messen unterschiedliche Aspekte des untersuchten Komplexes und sind hoch korreliert. Dies führt zu Multikollinearität im Cox-PH-Modell. Parameter, die möglicherweise eine starke Assoziation mit dem Ereignis besitzen, können so die Signifikanz verfehlen. Ein häufiges Vorgehen bei der Cox-Regression ist die schrittweise rückwärtige Selektion von nicht-signifikanten Parametern, bis sich nur noch signifikante Einflussvariablen im Modell befinden. Bei diesem

Vorgehen müsste hier ein Großteil der HRV-Parameter aus dem Modell entfernt werden. Auch die Interpretation muss sich auf die Effekte der einzelnen Indikatoren beschränken, von denen bekannt ist, dass sie erst zusammengenommen in der Lage sind, das Verhalten vollständig zu beschreiben. Ein explizites Maß für die Herzfrequenzvariabilität ist nicht verfügbar.
Eine weitere Einschränkung der vorliegenden Analyse ist, dass die Beziehungen zwischen den unabhängigen Parametern nicht modelliert werden können. Im betrachteten Beispiel stellt sich z.B. die Frage, ob die Kovariate zlves nicht nur mit dem Ereignis Koronartod, sondern auch mit der Herzfrequenzvariabilität zusammenhängt.

Jeder der betrachteten Parameter repräsentiert einen unterschiedlichen Aspekt der Herzfrequenzvariabilität. Diese Annahme unterstützt die Definition einer oder mehrerer latenter Variablen. Im folgenden Abschnitt soll das Survivalmodell mit latenten Variablen angepasst werden.

5.2.2 Analyse mit dem Survivalmodell mit latenten Variablen

Das Modell besteht aus drei Teilen. Es gibt eine Kovariate, die zehn standardisierten HRV-Parameter sowie einen Ereignisindikator und eine Zeitvariable. Es wird der direkte Einfluss der Kovariate auf die Ereigniszeit betrachtet. Die HRV-Parameter werden zu einer oder mehreren latenten Variablen zusammengefasst und es wird nur der Einfluss der latenten Variablen auf die Zeit bis zum Ereignis geschätzt. Eine signifikante Beziehung zwischen der Kovariate und den latenten Variablen kann im Modell berücksichtigt werden.
Die Anpassung des Modells geschieht schrittweise, wie in Abschnitt 5.1.2 beschrieben. Zunächst wird das optimale Messmodell für die HRV-Parameter gesucht. Es werden Faktorenanalysen (CFA) mit einer unterschiedlichen Anzahl von latenten Variablen angepasst.

Modell	Log-Likelihood	# freie Parameter	BIC
CFA 1 F	-6194	30	12576
CFA 2 F	-6001	39	12244

Tabelle 5.4: Modellanpassung

Das Ergebnis der Modellanpassung ist in Tabelle 5.4 aufgeführt. Die erklärte Varianz in dem Modell mit einem latenten Faktor beträgt 40.3%, im Zwei-Faktorenmodell erklärt die erste latente Variable 35.1% der Varianz und die zweite weitere 19.6%. Die Lösung der CFA mit zwei Faktoren zeigt im Vergleich das kleinste BIC und damit den besten Fit.
Tabelle 5.5 zeigt das Ergebnis der konfirmatorischen Faktorenanalyse mit zwei latenten Variablen. Zur Identifizierbarkeit des Modells wird die Varianz der latenten Faktoren F1 und F2 auf 1 gesetzt. Alle Faktorladungen werden frei geschätzt. Bei der Modellierung wird keine Korrelation zwischen den Faktoren zugelassen. Die Faktorladungen in der Tabelle sind folglich standardisiert. Dadurch sind die Parameterschätzer vergleichbar. Es kann eine Aussage darüber getroffen werden, durch welche Variablen die Faktoren primär definiert sind.
In Tabelle 5.5 ist zu erkennen, dass F1 primär durch zrnspsdnn, znhrvidx und zrspsdnn definiert ist. Dabei handelt es sich um Parameter, welche die gesamte Verteilung der RR-Intervalle während der Tages (zrspsdnn) und der Nacht- (zrnspsdnn) Periode beschreiben. Ein typischer Parameter dafür ist auch hrvidx, ein Schätzer für die gesamte Herzratenvariabilität [ESC96]. Faktor 2 wird insbesondere durch zrlf5 und zislml, aber auch durch zrhf und

	Variable	Parameterschätzer	SE	p-Wert
F1 by	zef	0.11	0.06	0.053
F1 by	znmeannn	0.60	0.06	<0.001
F1 by	zrhf	0.51	0.10	<0.001
F1 by	zlfn	0.24	0.05	<0.001
F1 by	zrnspsdnn	0.89	0.02	<0.001
F1 by	znhrvidx	0.78	0.03	<0.001
F1 by	zrspsdnn	0.82	0.06	<0.001
F1 by	zrlf5	0.62	0.10	<0.001
F1 by	zinslml	-0.16	0.05	0.003
F2 by	zef	0.20	0.06	0.001
F2 by	znmeannn	0.28	0.08	<0.001
F2 by	zrhf	0.56	0.13	<0.001
F2 by	zlfn	0.25	0.06	<0.001
F2 by	zrnspsdnn	0.23	0.08	0.003
F2 by	znhrvidx	0.29	0.07	<0.001
F2 by	zrspsdnn	0.20	0.12	0.096
F2 by	zrlf5	0.69	0.14	<0.001
F2 by	zinslml	0.55	0.11	<0.001
F2 by	zislml	0.72	0.14	<0.001

Tabelle 5.5: Faktorenanalyse

zinslml bestimmt. Dabei handelt es sich um Parameter, die hauptsächlich die Variabilität von Herzschlag zu Herzschlag repräsentieren. Nahezu alle Faktorladungen sind signifikant.
Im nächsten Schritt wird die Faktorenanalyse mit zwei latenten Variablen um den Einfluss der Kovariaten erweitert. Es zeigt sich, dass die Kovariate zlves signifikant mit F2 assoziiert ist, jedoch nicht mit F1. Die PH-Annahme ist für die Kovariate erfüllt. Als Survivalmodell kann das Proportional-Hazards-Modell verwendet werden.
Im nächsten Schritt wird das Survivalmodell mit den latenten Variablen geschätzt. Es wird spezifiziert, dass die latenten Faktoren F1 und F2 nicht miteinander korrelieren. Tabelle 5.6 zeigt das Ergebnis des finalen Modells. Der erste Teil der Tabelle beschreibt die Definition der Faktoren durch die HRV-Parameter. Die berechneten Parameterschätzer sind die jeweiligen Faktorladungen. Die meisten der Faktorladungen sind statistisch signifikant. An dieser Stelle kann untersucht werden, wie stark sich die Faktorladungen durch die zusätzliche Berücksichtigung des Einflusses der Kovariaten und der Ereigniszeit geändert haben. Bei dieser Analyse haben sich die Faktorladungen im Vergleich zu den Ergebnissen in Tabelle 5.5 kaum verändert.
Im unteren Teil von Tabelle 5.6 ist der Einfluss der Kovariate auf die latente Variable F2 angegeben. Der angegebene Parameterschätzer ist der Koeffizient einer linearen Regression. Da sowohl die Kovariate als auch der latente Faktor standardisiert sind, bezieht sich der Koeffizient auf eine Änderung der Variablen in Standardabweichungen. Der Einfluss von zlves auf Faktor 2 ist negativ. Das heißt, dass Patienten mit niedrigen Werten von zlves im Allgemeinen einen hohen F2 Faktorscore besitzen.
Im nächsten Abschnitt der Tabelle sind die Parameterschätzer β der Faktoren in der Cox-

	Parameter	Schätzer	SE	p-Wert	Haz. Ratio
F1 by	zef	0.09	0.05	0.086	
F1 by	znmeannn	0.59	0.06	<0.001	
F1 by	zrhf	0.48	0.08	<0.001	
F1 by	zlfn	0.22	0.05	<0.001	
F1 by	zrnspsdnn	0.89	0.04	<0.001	
F1 by	znhrvidx	0.78	0.04	<0.001	
F1 by	zrspsdnn	0.80	0.06	<0.001	
F1 by	zrlf5	0.58	0.09	<0.001	
F1 by	zinslml	-0.15	0.05	0.004	
F2 by	zef	0.23	0.05	<0.001	
F2 by	znmeannn	0.27	0.08	<0.001	
F2 by	zrhf	0.57	0.09	<0.001	
F2 by	zlfn	0.28	0.07	<0.001	
F2 by	zrnspsdnn	0.23	0.08	0.003	
F2 by	znhrvidx	0.29	0.07	<0.001	
F2 by	zrspsdnn	0.23	0.10	0.022	
F2 by	zrlf5	0.73	0.12	<0.001	
F2 by	zinslml	0.46	0.15	0.002	
F2 by	zislml	0.62	0.14	<0.001	
F2 on	zlves	-0.26	0.06	<0.001	
Time on	F1	-0.42	0.18	0.018	0.66
Time on	F2	-0.85	0.26	0.001	0.43
Time on	zlves	0.64	0.16	<0.001	1.90

Tabelle 5.6: Survivalmodell mit latenten Variablen, standardisiert

Regression aufgeführt. Zusätzlich ist das Hazard Ratio angegeben. Ein Anstieg der latenten Variablen F1 um eine Standardabweichung ist mit einer um 34% verringerten Mortalität assoziiert. Faktor 2 verhält sich ähnlich: Ein Anstieg des Faktorscores um eine Standardabweichung verringert das Risiko um 57%.
Die letzte Zeile von Tabelle 5.6 gibt den Einfluss der Kovariate auf die Ereigniszeit an. Der Parameterschätzer der loglinearen Regression kann so interpretiert werden, dass eine Zunahme von zlves um 1 Standardabweichung mit einem 1.9-fach erhöhtem Risiko assoziiert ist. In Abbildung 5.2 ist das Ergebnis des Survivalmodells mit den latenten Variablen in einem Pfaddiagramm dargestellt.

5.2.3 Vergleich der Modelle

Das Cox-Modell verdankt seine weite Verbreitung seiner Einfachheit. Der Einfluss der Parameter auf das Ereignis kann einzeln oder simultan untersucht werden. In einem univariaten Modell ist zu beachten, dass die einzelnen Parameter jeweils nur Teile der Herzfrequenzvariabilität messen. Werden alle Parameter gemeinsam in einer multiplen Cox-Regression analysiert, ergeben sich Schwierigkeiten durch starke Kollinearitäten zwischen den Variablen. Die hohe Korrelation zwischen den einzelnen Indikatoren ist nicht erstaunlich, wenn man bedenkt, dass die Indikatoren unterschiedliche Aspekte der gleichen Diagnose messen. Dieses Problem kann

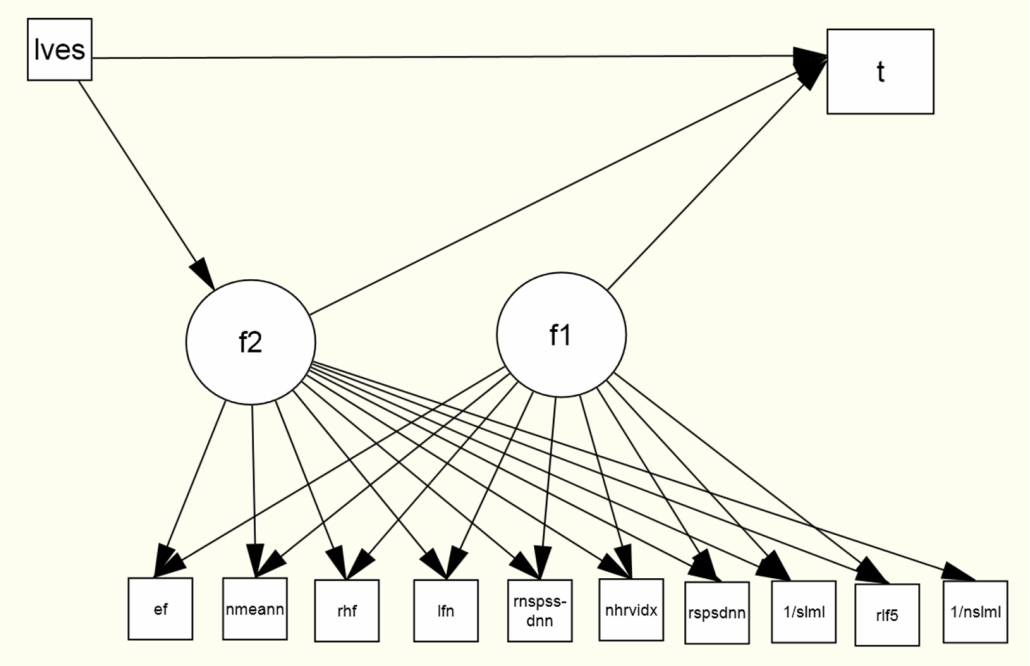

Abbildung 5.2: Survivalmodell mit latenten Faktoren für den HRV-Datensatz

mit dem Survivalmodell mit latenten Variablen überwunden werden.
Eine weitere Limitation der Standard-Cox-Regression ist, dass Beziehungen zwischen den Kovariaten nicht modelliert werden können. Das Cox-Modell ist in dieser Hinsicht wenig flexibel. Im Survivalmodell mit latenten Variablen kann gleichzeitig der Einfluss eines Parameters auf die Überlebenszeit und auf das latente Konstrukt geschätzt werden. Das macht die simultane Schätzung der Modellparameter mit dem Maximum-Likelihood-Ansatz möglich. So können differenzierte Aussagen über direkte und indirekte Beziehungen zwischen Kovariaten und der Überlebenszeit gemacht werden.
Das erweiterte Modell bietet die Möglichkeit, zur Entwicklung von Theorie beizutragen. Anhand von Modellgütekriterien können Annahmen validiert oder abgelehnt werden. Im Gegenzug nimmt der Anwender die Restriktionen in Kauf, die bei der Analyse von Strukturgleichungsmodellen gemacht werden.
Das Survivalmodell mit latenten Variablen ermöglicht es, ein direktes Maß für die Herzfrequenzvariabilität zu berechnen. Außerdem kann die Struktur dieser Variablen untersucht werden. Im vorgestellten Beispiel zeigt sich, dass die Herzfrequenzvariabilität am besten durch zwei latente Faktoren beschrieben wird. An diesem Punkt liegt jedoch eine Herausforderung der Methode. Die Interpretation der Ergebnisse mit latenten Variablen ist in vielen Bereichen, wie z.B. in der Medizin, weitestgehend unüblich.
Zusammenfassend lässt sich sagen, dass mit dem Survivalmodell mit latenten Variablen zusätzliche Fragen beantwortet werden können, die bei der Standardmodellierung offengeblieben sind. Dieser Zugewinn an Information geschieht im Tausch gegen ein komplexeres Modell mit zusätzlichen Annahmen. In solchen Fällen, in denen verschiedene Parameter als Indikatoren für einen Aspekt verwendet werden oder komplexe Strukturen zwischen den Kovariaten bestehen, ist das Survivalmodell mit latenten Variablen in jedem Fall die adäquate Wahl.

5.2.4 Evaluation der Modellstabilität mit Bootstrapping

Das berechnete Modell ist wesentlich komplexer als eine Standard-Cox-Regression. Es bleibt zu untersuchen, wie verlässlich die berechnete Lösung ist. Ein Hinweis für die Reliabilität eines Modells ist es, wenn die Ergebnisse anhand eines anderen Datensatzes repliziert werden können. Im Allgemeinen liegt so ein zweiter Datensatz, der unter den gleichen Voraussetzungen gewonnen wurde, aber nicht vor.
Ein statistisches Verfahren, mit dem die Zuverlässigkeit einer Lösung untersucht werden kann, ist das Bootstrapping. Dabei werden mehrfach Stichproben der gleichen Größe aus dem Originaldatensatz gezogen. Es wird mit Zurücklegen gezogen, sodass die Informationen zu einigen Patienten doppelt oder mehrfach vorkommen können, während anderen Patienten ganz entfallen. Das Modell wird dann für die neu generierten Datensätze berechnet. Auf diese Weise können zahlreiche Schätzer für die einzelnen Parameter ermittelt werden. Mithilfe dieser Ergebnisse können ein Bootstrap-Schätzer für den Standardfehler und ein approximatives Konfidenzintervall berechnet werden. Eine ausführliche Einführung in das Bootstrapping ist bei [ET93] gegeben.
Im Folgenden soll die Bootstrap-Methode auf das Survivalmodell mit latenten Variablen angewendet werden. Es werden insgesamt 250 Stichproben aus dem Originaldatensatz gezogen und genauso oft das Modell berechnet. In Anlehnung an [ET93] werden die beobachteten Schätzer nachfolgend mit $\hat{\theta}$, Bootstrap-Schätzer mit $\hat{\theta}^*$ und Bootstrap-Standardfehler mit $\hat{se}^*$ bezeichnet.

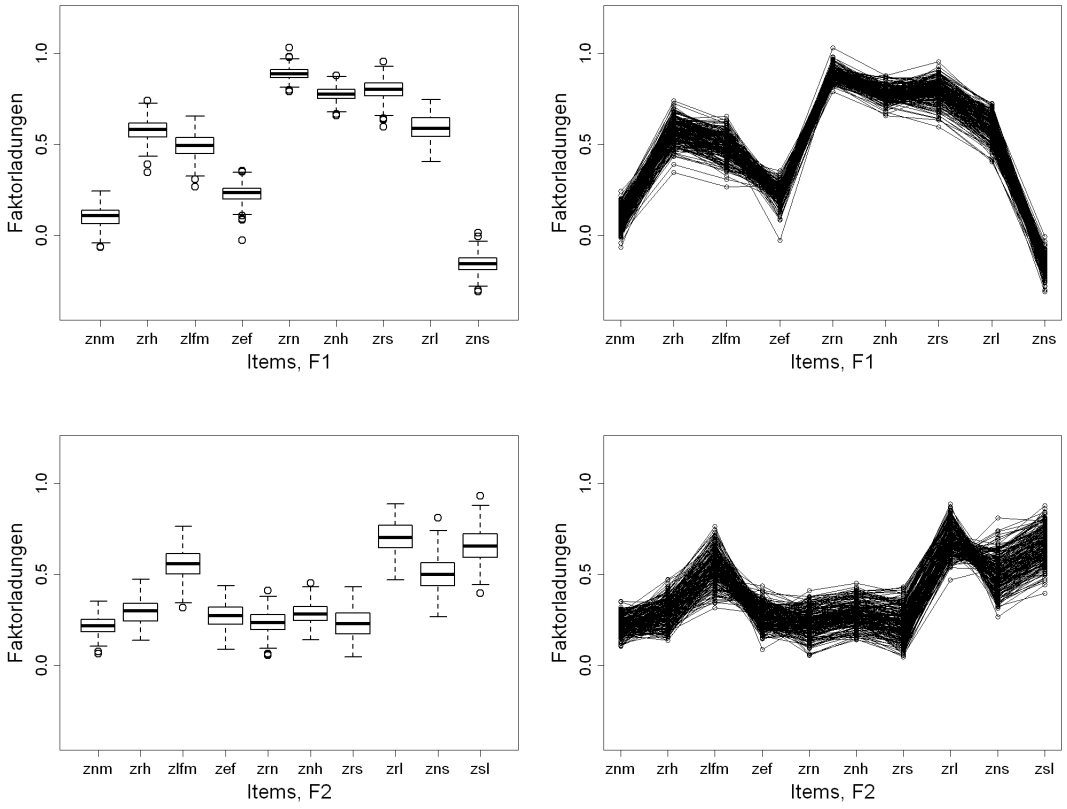

Abbildung 5.3: Bootstrap-Replikationen der Faktorladungen der HRV-Parameter

Abbildung 5.3 zeigt die Bootstrap-Parameterschätzer der Faktorladungen. Die Schätzer aller Items zeigen eine relativ geringe Streubreite. Die Standardfehler der Faktorscores von F1 liegen zwischen 0.04 und 0.07. Die Faktorladungen von F2 sind etwas größer, sie befinden sich zwischen 0.05 und 0.09. Die beobachteten Faktorladungen liegen bei jedem Item sehr nah am zugehörigen Bootstrapschätzer (hier nicht dargestellt). Diese Ergebnisse lassen auf eine geringe Unsicherheit bei der Schätzung der Faktorladungen schließen. Dieser Eindruck zeigte sich auch zuvor, da sich die Ergebnisse der Faktorenanalyse in 5.5 nur geringfügig von denen des gesamten Modells in Tabelle 5.6 unterscheiden. Die Liniendiagramme in Abbildung 5.3 bieten noch einen anderen Blick auf die Ergebnisse. Hier sind die geschätzten Faktorladungen aller 250 Bootstrap-Replikationen abgebildet. Ergebnisse eines Modells sind mit einer Linie verbunden. Eine solche Graphik kann dazu verwendet werden, Ausreißerpfade zu entdecken, die eine andere Definition der latenten Faktoren unterstützen.

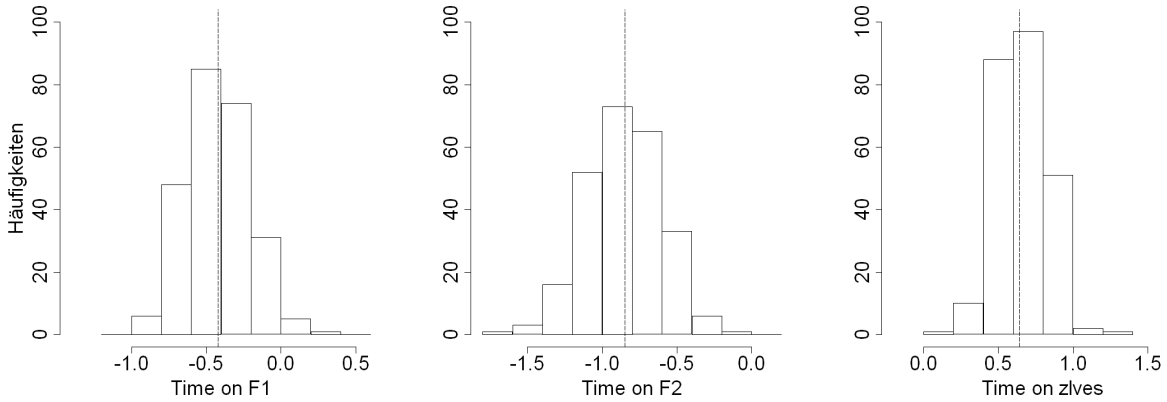

Abbildung 5.4: Histogramm der Bootstrap-Replikationen

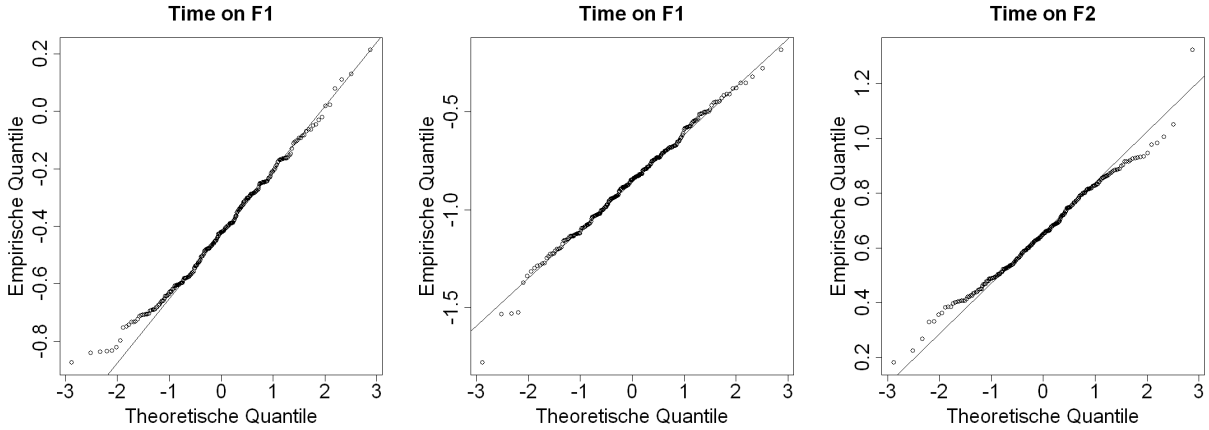

Abbildung 5.5: QQ-Plots für die Bootstrap-Replikationen

Abbildung 5.4 zeigt die Histogramme der Bootstrap-Schätzer für die Koeffizienten von F1, F2 und zlves in der Cox-Regression. Die beobachteten Schätzer aus Tabelle 5.6 sind mit vertikalen Linien gekennzeichnet. In Abbildung 5.5 sind zusätzlich QQ-Plots der Ergebnisse dar-

stellt. Aus diesen Abbildungen kann auf normalverteilte Daten geschlossen werden. Um einen noch klareren Eindruck von der Verteilung zu bekommen, könnten zusätzliche Bootstrap-Stichproben gezogen werden.

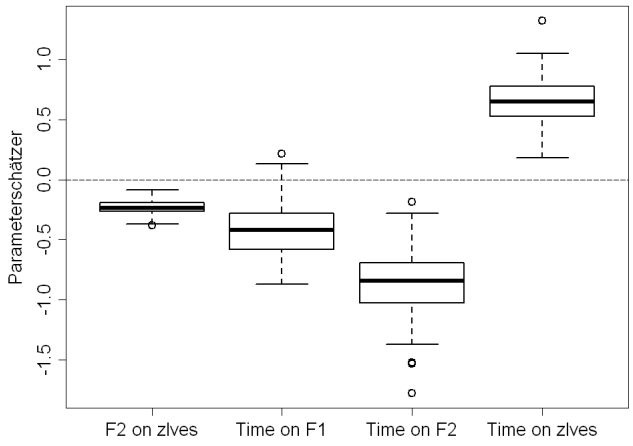

Abbildung 5.6: Bootstrap-Replikationen der Parameterschätzer in der Cox-Regression im Boxplot

In Abbildung 5.6 sind die Bootstrap-Schätzer für die Koeffizienten der Cox-Regression und die Regressionskoeffizienten für den Zusammenhang zwischen zlves und F2 in einem Boxplot dargestellt.

Variable	B.str.-Mittelw. ($\hat{\theta}^*$)	B.str.-SE ($\hat{se}^*$)	Schätzer ($\hat{\theta}$)	95% B.str. - KI	
F2 on zlves	-0.23	0.06	-0.26	-0.37	-0.15
Time on F1	-0.42	0.21	-0.42	-0.83	-0.02
Time on F2	-0.86	0.25	-0.85	-1.34	-0.36
Time on zlves	0.66	0.17	0.64	0.31	0.96

Tabelle 5.7: Übersicht der Bootstrap-Ergebnisse

Tabelle 5.2.4 gibt eine Übersicht über die beobachteten Parameter, Mittelwerte der Bootstrap-Replikationen, Bootstrap-Standardfehler und die daraus berechneten approximativen Konfidenzintervalle.
Die Mittelwerte der Bootstrap-Replikationen liegen für alle vier Parameter nah an den beobachteten Schätzern. Deutliche Unterschiede sind bei der Varianz der Ergebnisse zu erkennen. Die geringste Unsicherheit ist bei der Schätzung des Zusammenhangs zwischen F2 und zlves zu sehen. Der Standardfehler ist so niedrig wie bei den Faktorladungen. Das zugehörige 95%-Bootstrap-Konfidenzintervall ist entsprechend schmal.
Die Bootstrap-Schätzer von F1 und F2 besitzen mehr als dreimal so große Standardfehler. Die Bootstrap-Konfidenzintervalle sind entsprechend breit, schließen die 0 jedoch nicht mit ein. Das bedeutet, dass ein Anstieg des Faktorscores von F1 oder F2 in jedem Fall mit einem verringerten Risiko assoziiert ist. In Abbildung 5.6 ist ein klarer Unterschied in der Größe

der beiden Parameter zu sehen. Die Schätzer von F1 besitzen eine Spannweite von -1 bis zu Werten über 0. Die Parameterschätzer von F2 liegen deutlich tiefer, mit Extremwerten bis zu -1.8.
Der Bootstrap-Schätzer von zlves besitzt ebenfalls einen relativ großen Standardfehler. Das Konfidenzintervall befindet sich deutlich über der 0. Ein Anstieg der Variablen zlves impliziert damit sicher ein erhöhtes Mortalitätsrisiko.
Die vier Parameterschätzer aus Abbildung 5.6 hängen aufgrund der simultanen Schätzung des Modells eng miteinander zusammen. Ändert sich einer der Schätzer, so ist zu erwarten, dass sich auch die anderen Schätzer ändern. Wie sich diese Zusammenhänge gestalten, ist in dem Boxplot nicht zu erkennen. Die Ergebnisse der Bootstrap-Replikationen werden deshalb noch einmal in einem Liniendiagramm dargestellt.

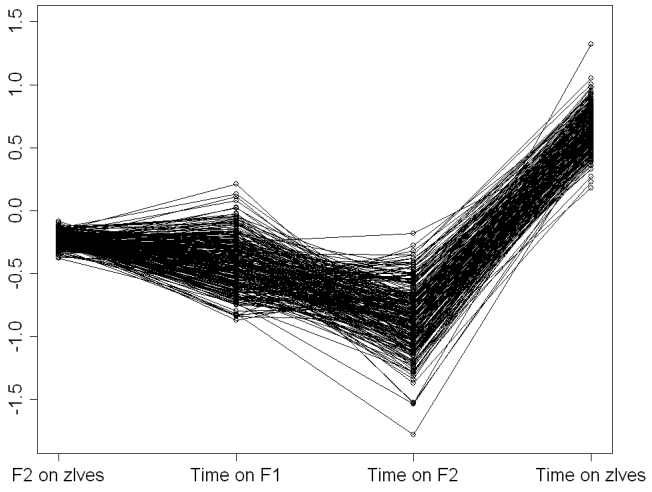

Abbildung 5.7: Bootstrap-Replikationen der Parameter in der Cox-Regression im Liniendiagramm

In Abbildung 5.7 ist wieder zu sehen, dass die Parameterschätzer für den Einfluss der Kovariate auf F2 keine großen Schwankungen zeigen. Damit sind auch keine besonderen Zusammenhänge mit den anderen Parameterschätzern erkennbar. Ähnlich sieht es mit den Beziehungen zwischen den Parameterschätzern von F2 und zlves in der Cox-Regression aus.
Interessanter ist die Beziehung zwischen den Parameterschätzern der beiden latenten Faktoren. Höhere Schätzer von F1 sind mit niedrigeren Werten bei F2 verbunden. Zu dem gemeinsamen Trend ist eine negative Korrelation zu erkennen. Besonders niedrige F1 Werte fallen mit besonders hohen Parameterschätzern von F2 zusammen. In der Abbildung 5.7 sind das die Fälle, bei denen die Linie zwischen F1 und F2 ansteigt. Die Summe der beiden Parameterschätzer ist immer ungefähr gleich groß. Überraschend ist das nicht, da sich das Gesamtrisiko der zehn HRV-Parameter, das in jedem Durchlauf ähnlich ist, auf die zwei Faktoren aufteilen muss. Außerdem sind die beiden latenten Variablen fast durch das gleiche Set von Parametern definiert. In den meisten Fällen wird ein Faktor F1 mit einem höheren und ein Faktor F2 mit einem niedrigeren Parameterschätzer bestimmt. Die Abbildung zeigt jedoch, dass es auch eine geringe Wahrscheinlichkeit dafür gibt, dass die Faktoren andersherum definiert sind.

Kapitel 6

Latente Strukturanalyse

Latente Strukturanalyse ist der Oberbegriff für Modelle, die den Zusammenhang zwischen einer kategoriellen latenten Variablen und manifesten Items unterschiedlicher Skalenniveaus beschreiben. Die Verfahren wurden ursprünglich von P.F. Larzarsfeld und N.W. Henry [LH68] entwickelt. Die latente Profilanalyse und die latente Klassenanalyse sind Spezialfälle dieser Modelle respektive für metrische und kategorielle Items. Es wurde gezeigt, dass die beiden Modelle auch als Submodelle von Mischverteilungsverfahren verstanden werden können, siehe z.B. [Wol70]. Im Folgenden sollen die Verfahren in diesem Kontext erläutert werden.
Mischverteilungsverfahren, im Englischen "Finite Mixture Models", wurden anfänglich dazu entwickelt, um unbekannte Verteilungsformen zu modellieren. Nahezu jede Verteilung kann durch das Zusammenfügen einer genügend großen Anzahl von Normalverteilungen approximiert werden. Eine der ersten Analysen mit Mischverteilungen führte Karl Pearson (1894) durch. Die Parameterschätzer des Modells wurden damals mithilfe der Momentenmethode ermittelt. Der rechnerische Aufwand dabei war jedoch erheblich, sodass das Verfahren zunächst keine weite Verbreitung fand. Später wurde gezeigt, dass die Lösung mit der Maximum-Likelihood-Methode bessere Ergebnisse erzielt, siehe zum Beispiel [Wol70]. Breite Anwendung fanden die Verfahren erst einige Jahre später, dank der Erkenntnis, dass die Anpassung einer Mischverteilung mit der Maximum-Likelihood-Methode als Missing-Data-Problem betrachtet werden kann und mit dem EM-Algorithmus lösbar ist.
Mischverteilungen können auch dazu verwendet werden, um die Heterogenität in einer Gruppe von Individuen zu modellieren. Die Idee dahinter ist, dass es in der betrachteten Population Subgruppen von Individuen gibt, die eine unterschiedliche Verteilung der Response-Variablen besitzen. Diese Kategorien werden als latente Klassen bezeichnet. Die Mitgliedschaft in diesen Subgruppen ist nicht direkt zu beobachten und muss geschätzt werden. Bei der latenten Klassenanalyse und der latenten Profilanalyse ist eine zentrale Annahme, dass die Ausprägungen der manifesten Variablen bedingt auf ihre Klassenzugehörigkeit voneinander unabhängig sind.

6.1 Latente Profilanalyse

Die grundlegende Definition der latenten Profilanalyse wird hier in Anlehnung an McLachlan und Peel [MP00] vorgestellt. Auf der Basis von einem Set kontinuierlicher Variablen y werden die Individuen K latenten Klassen zugeordnet. Diese Subgruppen werden auch latente Profile oder Mixture-Komponenten genannt.

Es bezeichne c eine kategorielle latente Variable mit K verschiedenen Klassen, mit $k = 1, \ldots, K$. Hier bedeutet $c = k$ die Mitgliedschaft in Klasse k. Es werden die Items y_j mit $j = 1, \ldots, r$ beobachtet. Beim latenten Profilmodell wird angenommen, dass die Ausprägungen zweier Items bedingt auf ihre Klassenzugehörigkeit unabhängig voneinander sind. Die Annahme der stochastischen Unabhängigkeit der Items y_j impliziert:

$$P(y_1, \cdots, y_r|c) = \prod_{j=1}^{r} P(y_j|c).$$

Das heißt, die Ausprägung einer Variable y_j hängt allein von der Klassenzugehörigkeit c ab, nicht von den anderen Variablen. Für die Wahrscheinlichkeit eines Antwortmusters $y_j, \cdots, y_r$ folgt damit:

$$P(y_1, \cdots, y_r) = \sum_{c=1}^{K} P(c = k) \prod_{j=1}^{r} P(y_j|c).$$

Hier bezeichne $P(c = k)$ die Zuordnungswahrscheinlichkeit der Individuen zu den Klassen k. Da es sich um kontinuierliche Variablen y_j handelt, liegt es nah, anzunehmen, dass die Items bedingt auf ihre Klassenzugehörigkeit normalverteilt sind. Das allgemeine Modell der latenten Profilanalyse ist dann:

$$f(y) = \sum_{c=1}^{K} P(c = k) \prod_{j=1}^{r} f(y_j|\mu_{jk}, \sigma_{jk}^2). \tag{6.1}$$

Dabei bezeichne $f(y_j|\mu_{jk}, \sigma_{jk}^2)$ die Dichte der Normalverteilung mit Mittelwert μ_{jk} und Varianz σ_{jk}^2. Die Verteilung der Klassen c wird durch eine multinomiale logistische Regression bestimmt:

$$P(c = k) = \frac{exp(\alpha_k)}{\sum_{k=1}^{K} exp(\alpha_k)}. \tag{6.2}$$

Wobei α_K die Referenzkategorie bezeichnet. Die einzelnen α_k's geben die Wahrscheinlichkeiten der übrigen Klassen an. Es ist möglich, den Einfluss von Kovariaten auf die latenten Klassen zu untersuchen. Dazu wird in der multinomialen logistischen Regression ein Vektor x von Kovariaten berücksichtigt.
Im Allgemeinen wird die Analyse dazu verwendet, die Individuen den latenten Subgruppen zuzuordnen. Die posterioren Wahrscheinlichkeiten, den Klassen anzugehören, ergeben sich mit dem Satz von Bayes. Es folgt:

$$P(c = k|y_1, \cdots, y_r) = \frac{P(c = k)P(y_1|c = k) \cdots P(y_r|c = k)}{P(y_1, \cdots, y_r)}.$$

Hier ist anzumerken, dass die Individuen zunächst eine gewisse Wahrscheinlichkeit für mehrere Klassen erhalten können. Die Individuen werden der Klasse zugeordnet, für welche die gegebene Wahrscheinlichkeit am größten ist und daraufhin als vollständiges Mitglied dieser Kategorie angesehen.

6.2 Latente Klassenanalyse

Die latente Klassenanalyse kann ebenfalls als Mischverteilung von allgemeinen linearen Modellen formuliert werden. Im Gegensatz zur latenten Profilanalyse werden ausschließlich kategorielle Items betrachtet.
Das Verfahren wird auch als kategorielles Analogon der Faktorenanalyse bezeichnet. Wie bei der Faktorenanalyse wird eine latente Variable verwendet, um die Beziehung zwischen einer Reihe von manifesten Variablen zu beschreiben. Es wird angenommen, dass die beobachteten Items bedingt auf ihre Gruppenzugehörigkeit voneinander unabhängig sind. Die Korrelation zwischen den Indikatorvariablen wird vollständig durch das latente Konstrukt erklärt und es gibt keine residuale Korrelation zwischen den Items. Im Folgenden wird die allgemeine Modellformulierung vorgestellt [MP00, Mas03].
Es werden die Items u_j, mit $j = 1, \ldots, r$ betrachtet, mit s Kategorien pro Item. Dabei bezeichne c eine kategorielle latente Variable mit K verschiedenen Klassen, mit $k = 1, \ldots, K$. Hier bedeutet $c = k$ die Mitgliedschaft in Klasse k. Die Annahme der stochastischen Unabhängigkeit der beobachteten Items u_j, bedingt auf die Klassenzugehörigkeit, impliziert:

$$P(u_1, \cdots, u_r|c) = \prod_{j=1}^{r} P(u_j|c).$$

Für die Wahrscheinlichkeit eines Antwortmusters $u_j, \cdots, u_r$ folgt dann nach dem Satz der totalen Wahrscheinlichkeit:

$$P(u_1, \cdots, u_r) = \sum_{c=1}^{K} P(c = k) \prod_{j=1}^{r} P(u_j|c).$$

Alle theoretischen Parameter der rechten Seite sind unbekannt und müssen geschätzt werden. Die latenten Klassen können zusätzlich durch Kovariaten x beeinflusst werden. Wie bei der latenten Profilanalyse wird der Zusammenhang mit einer multinomialen logistischen Regression beschrieben:

$$P(c = k|x) = \frac{exp(\alpha_k + \gamma_k x)}{\sum_{k=1}^{K} exp(\alpha_k + \gamma_k x)}. \tag{6.3}$$

Dabei wird üblicherweise die letzte Kategorie, K, als Referenzkategorie verwendet, d.h. es gelte $\alpha_K = 0$ und $\gamma_K = 0$, sodass $exp(\alpha_K + \gamma_K x) = 1$. Dann ergeben sich die Odds einer Kategorien k im Vergleich zur Referenzkategorie wie folgt:

$$\frac{P(c = k|x)}{P(c = K|x)} = exp(\alpha_k + \gamma_k x). \tag{6.4}$$

Für binäre Indikatorvariablen u_j kann die Verteilung durch eine logistische Regression definiert werden:

$$P(u_j = 1|c = k) = \frac{1}{1 + exp(-\nu_{jk})}. \tag{6.5}$$

Hier bezeichnen die ν_{jk} die Logits der u_j's für jede Klasse k. Im Falle einer Variablen u_j mit mehr als zwei geordneten Kategorien wird eine polytome logistische Regression verwendet. Die Formulierung ist bei [Mut04b] zu finden.

Die posterioren Wahrscheinlichkeiten eines Individuums, einer bestimmten Klasse anzugehören, ergeben sich mit dem Satz von Bayes:

$$P(c = k|u_1, \cdots, u_r) = \frac{P(c = k)P(u_1|c = k) \cdots P(u_r|c = k)}{P(u_1, \cdots, u_r)}.$$

Die Individuen werden der Klasse zugeordnet, zu der sie die größte Zugehörigkeitswahrscheinlichkeit besitzen.
In Abbildung 6.1 ist ein schematisches Pfaddiagramm für die latente Klassenanalyse abgebildet. Der Kreis bezeichnet die latente kategorielle Variable, die Rechtecke $u_1, \cdots, u_r$ die manifesten Indikatorvariablen. Die Pfade vom latenten Konstrukt auf die Indikatoren indizieren, dass die beobachteten Antworten vollständig durch die latenten Klassen erklärt werden.

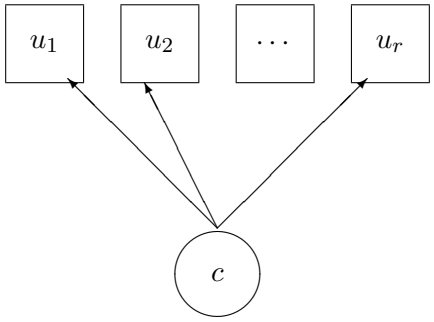

Abbildung 6.1: Schematisches Pfaddiagramm für die latente Klassenanalyse

6.3 Schätzung

Die Schätzung der Modelle wurde wesentlich durch die Arbeit von Goodmann [Goo74] erleichtert. Dieser entwickelte ein Verfahren, dass es ermöglicht, LCA-Modelle mit der Maximum-Likelihood-Schätzung zu lösen. Der Algorithmus wird in der Literatur als Expectation-Maximization (EM)-Algorithmus bezeichnet. Das Verfahren iteriert bei der Berechnung zwischen 2 Schritten. Im E-Schritt werden die Zuordnungswahrscheinlichkeiten $P(c = k|u_j)$ ermittelt, wobei die geschätzten Modellparameter als gegeben angenommen werden. Im M-Schritt werden die Modellparameter nach der ML-Methode geschätzt und die Zuordnungswahrscheinlichkeiten als bekannt vorausgesetzt. Die Schritte werden bis zum Erreichen einer konvergenten Lösung wiederholt [MP00].
In Mplus werden die Parameter mithilfe dieses Verfahrens ermittelt. Technische Details dazu sind bei [Mut04b] gegeben.

Sowohl bei der latenten Profilanalyse als auch bei der latenten Klassenanalyse kann es zu Schwierigkeiten bei der Schätzung kommen, wenn die Likelihood-Funktion mehrere lokale Maxima besitzt. Dies ist kein seltenes Problem und kann verschiedene Gründe haben, die im Folgenden näher betrachtet werden sollen.
Die einfachste Möglichkeit ist, dass zu viele latente Klassen spezifiziert wurden.
Ein anderer Grund kann sein, dass manifeste Variablen innerhalb ihrer latenten Klasse nicht

unabhängig voneinander sind. Dieses Problem wird als “bedingte Unabhängigkeit” oder “lokale Unabhängigkeit” bezeichnet [Ueb12]. Wenn solche Abhängigkeitsstrukturen nicht beachtet werden, führt das häufig dazu, dass zusätzliche latente Klassen geschätzt werden, bis das Modell zu den Daten passt. Diese latenten Klassen besitzen im Allgemeinen aber keine theoretische Basis.
Für die latente Klassenregression schlägt Ubersax [Ueb12] einen Log-Odds-Ratio-Check vor, um bedingte Unabhängigkeiten zwischen den Indikatorvariablen zu entdecken. Für den Fall, dass Items nicht lokal unabhängig sind, werden drei Vorschläge gemacht, um die Annahme der lokalen Unabhängigkeit zu entspannen. Zunächst können zwei abhängige Variablen zu einer Variable mit mehr Kategorien zusammengefasst werden. Alternativ ist es möglich, konditional abhängige Variablen als multiple Indikatoren einer gemeinsamen latenten Variablen zu modellieren. Die letzte Möglichkeit besteht darin, die latente Klassenanalyse als loglineares Modell [VM03] zu formulieren. Mit dieser Formulierung können einfacher bestimmte Typen von Modellen mit bedingten Abhängigkeiten spezifiziert werden.
Bei der latenten Klassenregression ist häufig zu beobachten, dass ein lokales Maximum der Likelihood-Funktion erreicht wird, wenn viele klassenbedingte Antwortwahrscheinlichkeiten auf exakt 0 oder 1 geschätzt werden. Diese sogenannten “Boundary Solutions” können ebenfalls ein Zeichen für die Spezifikation von zu vielen latenten Klassen beziehungsweise eine unvalide Lösung sein [Gei10].
Das übliche Vorgehen zur Vermeidung eines lokalen Maximums bei der Schätzung einer latenten Klassenanalyse oder einer latenten Profilanalyse ist die Wahl von multiplen Startwerten. In den meisten Programmen, mit denen die Modelle gelöst werden können, wie auch bei Mplus, kann eine beliebige Anzahl von zufälligen Startwerte-Sets generiert werden [Ueb12, Mut04b]. Der beste gefundene Log-Likelihood-Wert sollte mindestens durch zwei unterschiedliche Startwertesets erreicht werden.

6.4 Identifikation

Wie bei Strukturgleichungsmodellen ist es wichtig nach der Spezifikation eines Modells zu überprüfen, ob es identifizierbar ist. Dabei wird zwischen intrinsischer und empirischer Identifizierbarkeit eines Modells unterschieden [Ueb12].
Nicht-Identifikation eines Modells bedeutet, dass es verschiedene Sets von Parameterschätzern gibt, die zu dem gleichen Maximum der (Log-)Likelihood-Funktion führen. Intrinsische Nicht-Identifikation bedeutet, dass ein Modell unabhängig vom verwendeten Datensatz nicht lösbar ist. Das ist der Fall, wenn ein Modell zu komplex spezifiziert wurde. Die intrinsische Modellidentifikation kann leicht überprüft werden, wenn man bedenkt, dass die maximal zu schätzende Anzahl von Parametern durch die Freiheitsgrade limitiert ist. Bei der latenten Profilanalyse sind insgesamt $K*(1+2*r)-1$ Parameter zu bestimmen. Damit ist die notwendige Bedingung für ein identifiziertes Modell, dass mindestens genauso viele unterschiedliche Datensätze vorliegen. Bei der latenten Klassenanalyse ist die maximale Anzahl zu schätzender Parameter durch die Anzahl der unterschiedlichen beobachteten Responsepattern minus 1 limitiert [Ueb12].
Empirische Nicht-Identifikation tritt auf, wenn ein Modell aufgrund einer zufälligen besonderen Struktur der Daten nicht lösbar ist. Dies ist keine häufige Erscheinung, aber wahrscheinlicher bei kleinen Stichproben oder wenn nur ein geringer Anteil der möglichen Responsepattern

beobachtet wird [Ueb12].
In beiden Fälle kann man sich im Allgemeinen mit zusätzlichen Restriktionen behelfen und das Modell damit identifizierbar machen. Diese müssen natürlich inhaltlich sinnvoll sein.
Um zu überprüfen, ob ein Modell empirisch identifiziert ist, kann bei der Maximum- Likelihood-Schätzung die Hesse-Matrix betrachtet werden. Ein Modell ist identifiziert, wenn die Matrix vollen Rang besitzt. Typischerweise wird dieser Test jedoch berechnet, wenn der Algorithmus zu einer Lösung konvergiert ist. Diese Lösung besitzt dann keine Aussage und das Modell muss anders spezifiziert und neu berechnet werden.

6.5 Modellprüfung und Interpretation

Die Modellspezifikation bei der latenten Profilanalyse oder der latenten Klassenanalyse sollte immer auf inhaltlichen Hypothesen basieren. Im Allgemeinen ist im Voraus jedoch nicht bekannt, wie viele latente Klassen extrahiert werden sollen. Es gibt verschiedene Verfahren, mit denen überprüft werden kann, wie gut ein Modell zu den gegebenen Daten passt. Üblicherweise werden mehrere Modelle mit einer unterschiedlichen Anzahl latenter Klassen angepasst und unter Betrachtung verschiedener Kriterien das optimale Modell ausgewählt. Mithilfe von Klassifikationsdiagrammen kann beurteilt werden, wie gut die Individuen den Klassen zugeordnet wurden. Zur Interpretation der einzelnen Klassen können Klassenprofildiagramme betrachtet werden. Genauso wichtig wie die statistischen Kriterien ist bei der Modellwahl in jedem Fall die Interpretierbarkeit der Lösung.

Wie bei Strukturgleichungsmodellen können Tests zur Modellgüte mithilfe des Log-Likelihood-Wertes ermittelt werden. Standardmäßig wird bei den berechneten LCA-Modellen der Pearson- und der Likelihood-Ratio-X^2-Test angegeben. Ein signifikanter Wert deutet darauf hin, dass es eine statistisch bedeutsame Abweichung zwischen den beobachteten und den vorhergesagten Antwortmustern gibt. Um zu gewährleisten, dass die Likelihood-Ratio-Statistik tatsächlich X^2-Quadrat verteilt ist, wird eine relativ große Stichprobe benötigt [Gei10]. Ist diese nicht gegeben, so ist der Test wenig aussagekräftig.
Aufgrund der Schwierigkeiten bei der Beurteilung der absoluten Modellgüte wird gewöhnlich der relative Fit betrachtet. Zum Vergleich von Modellen mit einer unterschiedlichen Anzahl von Klassen stehen die in Abschnitt 4.2.5 beschriebenen informationstheoretischen Maße AIC und BIC (adjustiert und unadjustiert) zur Verfügung. Mplus bietet außerdem zwei statistische Tests, den Bootstrap-Likelihood-Ratio Differenztest und den Vuong-Lo-Mendell-Rubin-Test für den Modellvergleich. Eine ausführliche Beschreibung der Verfahren findet sich bei [Gei10].
Weniger ein Maß für die Güte eines Modells, als eine Hilfe bei der Entscheidung für die optimale Anzahl latenter Klassen ist die Entropie. Mit Klassifikationstabellen kann untersucht werden, wie gut sich die geschätzten Klassen voneinander unterscheiden. In den Zeilen der Tabelle sind die Klassen angegeben, denen die Individuen zugeordnete wurden, in den Spalten stehen die mittleren posterioren Wahrscheinlichkeiten der Individuen für die einzelnen Klassen [Mut04b]. Eine gute Klassifizierung ist erreicht, wenn die ermittelte Klassenzugehörigkeit gut zu den geschätzten Wahrscheinlichkeiten für die Klassen passt. Ein zusammenfassendes Maß für diese Eigenschaft ist die Entropie [Mut04b]:

$$E_K = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik} ln \hat{p}_{ik})}{n ln K}. \tag{6.6}$$

Dabei bezeichne $\hat{p}$ die geschätzte Wahrscheinlichkeit für Individuum i, der Klasse k anzugehören. Die Werte können zwischen 0 und 1 liegen, wobei größere Werte eine bessere Klassifikation bedeuten.

Wichtige Modellparameter bei der latenten Profilanalyse sind die erwarteten klassenspezifischen Mittelwerte beziehungsweise bei der latenten Klassenanalyse die klassenbedingten Itemwahrscheinlichkeiten. Nachdem ein Modell geschätzt ist und die latenten Klassen definiert wurden, bieten sie die Möglichkeit, den Subgruppen eine Bedeutung zuzuordnen. Ein nützliches Hilfsmittel dabei sind sogenannte Klassenprofildiagramme. Dabei handelt es sich um Liniendiagramme, auf denen die einzelnen Items auf der X-Achse aufgetragen sind und auf der Y-Achse die zugehörigen mittleren erwarteten Ausprägungen oder die zugehörigen Wahrscheinlichkeiten für jede einzelne Klasse angegeben werden.
Bei der Wahl eines Modells können und sollten mehrere der vorgestellten Verfahren berücksichtigt werden. Beispiele dafür sind bei [Lar04, MAB$^+$09] gegeben.

Kapitel 7

Survivalanalyse mit latenten Klassen

7.1 Modellspezifikation

Im Folgenden wird die Cox-Regression mit latenten Klassen nach der allgemeinen Formulierung von Muthen [AMM06] erläutert. Diese Formulierung verallgemeinert das Modell von Larsen (2004) [Lar04], das nur den Fall von binären Indikatorvariablen betrachtet. Das Modell ermöglicht es, eine latente Klassenvariable als Prädiktor für die Zeit bis zu einem Ereignis oder einer Zensierung zu verwenden. Dazu wird die latente Klassenregression mit einem Cox-PH-Modell kombiniert. Das Modell bietet außerdem die Möglichkeit, den Einfluss von anderen Kovariaten in der Cox-Regression klassenabhängig zu schätzen. Die Definition der latenten Variablen kann durch kategorielle oder kontinuierliche Indikatorvariablen geschehen, dabei können Kovariaten berücksichtigt werden. Die Parameter in dem Modell werden simultan mit der Maximum-Likelihood-Methode geschätzt.
Es werden Individuen i mit den zugehörigen Ereigniszeitvariablen T_i betrachtet. Es bezeichne c eine kategorielle latente Variable mit K verschiedenen Klassen, $k = 1, \ldots, K$. Dabei indiziert $c = k$ die Mitgliedschaft in Klasse k. Es werden die manifesten Variablen u_j mit $j = 1, \ldots, r$ beobachtet. In dem Fall, dass die Indikatorvariablen kategorielles Skalenniveau besitzen, wird eine kontinuierliche Variable u_j^* im Hintergrund definiert [Jör02, Mut04b]. Für die j-te kategorielle Variable u_{ji} mit $v = 1, 2, \ldots, s$ geordneten Kategorien, bedeutet das:

$$[u_{ji} = v | c_i = k] \quad \Leftrightarrow \quad \tau_{kv} < u_{ji}^* < \tau_{kv+1}.$$

Zur Bestimmung der latenten Klassen c werden dann die Latent-Response-Variablen u_{ji}^* verwendet. Im Folgenden gilt, dass die u_{ji}^* normalverteilt sind, häufig wird auch eine logistische Verteilung angenommen [Mut04b]. Bei Mplus wird als Voreinstellung bei einer latenten Klassenanalyse mit Maximum-Likelihood-Schätzung für den Zusammenhang eine logistische Regression geschätzt [Mut04b].
In den folgenden Gleichungen wird der Zusammenhang zwischen dem Vektor der abhängigen Variablen u_i^*, einem Vektor von Kovariaten x_i und den latenten Klassen beschrieben. Außerdem wird die Hazard-Funktion definiert:

$$[u_i^* | c_i = k] = \mu_k + \epsilon_i \tag{7.1}$$

$$P(c_i = k | x_i) = \frac{exp(\alpha_k + \gamma_k x_i)}{\sum_{k=1}^{K} exp(\alpha_k + \gamma_k x_i)} \tag{7.2}$$

$$h_i(t|x_i, c_i = k) \quad = \quad h_{0k}(t) * exp(\iota_k + \beta_k x_i). \tag{7.3}$$

Der ersten beiden Gleichungen gehören zu der latenten Klassenregression. Hier ist ϵ_i ein normalverteilter Fehlerterm mit Mittelwert 0. Ein multinomiales Logit-Modell beschreibt den Zusammenhang zwischen latenten Klassen und Kovariaten. Zur Identifikation gelte $\alpha_K = 0$ und $\gamma_K = 0$.
Die untere Gleichung gibt die Hazard-Funktion für die Ereigniszeit eines Individuums i an. Hier bezeichne $h_{0k}(t)$ die Baseline-Hazard-Funktion, x_i einen Vektor von Kovariaten und β_i den Vektor der zugehörigen Koeffizienten. Der Vektor ι_k bezeichnet den Achsenabschnittsschätzer der Cox-Regression.
Es gibt zwei Möglichkeiten, wie der Einfluss der Klassenvariablen als Prädiktor bei der Überlebenszeitanalyse berücksichtigt werden kann. In beiden Fällen wird die Baseline-Hazard-Funktion als nichtparametrische Stufenfunktion geschätzt. Beim Ansatz von Larsen (2004), unterscheiden sich die Baseline-Hazard-Funktion der einzelnen Klassen jedoch nur durch einen einzigen multiplikativen Faktor. In der obigen Modellformulierung bedeutet das, dass der Einfluss der latenten Klassen allein durch den Achsenabschnitt der Harzardfunktion ι_k berücksichtigt wird. Zur Identifikation gelte $\iota_1 = 0$. Im Gegensatz dazu wird beim Mplus-Ansatz [AMM06] der Baseline-Hazard für jede Klasse vollkommen frei und ohne Restriktionen geschätzt. In diesem Fall kann der Achsenabschnittsschätzer der Hazard-Funktion nicht mehr identifiziert werden.
In beiden Fällen kann der Einfluss der Kovariaten klassenspezifisch geschätzt werden, auch wenn diese Möglichkeit in dem Artikel von Larsen nicht betrachtet wird [Lar04] .
Der Vorteil des Ansatzes von Larsen liegt darin, dass der Klasseneffekt explizit geschätzt und angegeben werden kann. Beim Mplus-Ansatz ist das nicht der Fall, da kein einzelner Parameter den Unterschied zwischen zwei vollkommen unrestringierten Baseline-Hazard-Funktionen ausdrücken kann. Außerdem ist das Modell von Larsen sparsamer. Die zusätzlichen Restriktionen führen dazu, dass wesentlich weniger Parameter geschätzt werden müssen und das Modell leichter identifizierbar ist. Der Vorteil beim Mplus-Ansatz besteht hingegen darin, dass das Modell auch angemessen ist, wenn die Annahme proportionaler Hazards zwischen den Klassen nicht erfüllt ist. Die Evaluation der Annahme proportionaler Hazards zwischen den latenten Klassen ist ein Problem, mit dem sich Larsen in [Lar04] beschäftigt. Muthen verwendet selbst die Methode von Larsen bei der Schätzung eines Mixure-Survivalmodells in [MAB$^+$09].
Der Mplus-Ansatz ist bereits in Mplus Version 5 implementiert, Larsens Ansatz ist kommerziell seit Mplus Version 6 verfügbar.

7.1.1 Schätzung

Bei der Cox-PH-Regression wird von nichtinformativ, rechtszensierten Daten ausgegangen, sodass für jedes Individuum neben der Ereigniszeitvariablen T_i ein Zensierungsindikator D_i gegeben ist. Dann ist die Likelihood der beobachteten Variablen (T_i, D_i, u_i) für ein einzelnes Individuum i:

$$P(T_i, D_i, u_i|x_i) = \sum_{k=1}^{K} P(c_i = k|x_i)P(u_i|c_i = k)P(T_i, D_i|c_i = k, x_i). \tag{7.4}$$

Die vollständige Likelihood der Daten im Cox-Modell unter Berücksichtigung der Zensierungen wird wie in Abschnitt 3.4.1 berechnet. Die Schätzung der Parameter geschieht mit der

Maximum-Likelihood-Methode unter Verwendung des Expectation-Maximization-Algorithmus. Larsen beschreibt das genaue Verfahren in [Lar04] für den Fall von kategoriellen Indikatorvariablen.

Die Berechnung einer latenten Klassenanalyse ist mit zahlreichen Softwareprogrammen möglich, die Analyse eines Mixture-Survivalmodells bietet bislang jedoch nur das Softwareprogramm Mplus. Eine Anmerkung zu der Analyse von latenten Klassenmodellen mit Mplus soll daher noch an dieser Stelle gemacht werden. Bei einer Mixture-Analyse werden alle Variablen im Modell als Indikatoren für die latenten Klassen verwendet, sofern sie nicht als Kovariaten spezifiziert werden, auf die die latente Klassenvariable regressiert wird. Es gibt nicht die Möglichkeit, nur bestimmte Indikatoren zur Definition einer latenten Klassenvariablen auszuwählen. Das bedeutet aber, dass auch die Response oder Zielvariable ein Indikator der latenten Klassen ist. Das kann ein Vorteil sein, da die Zielvariable Informationen für die Bestimmung der latenten Klassenvariablen beisteuert. Andererseits ist es vorstellbar, dass die Fragestellung des Anwenders darin besteht, bestimmte Indikatoren und die daraus geschätzten Klassen als Vorhersageinstrument für eine Zielvariable zu beurteilen. Ein mögliches Vorgehen für diesen Fall beschreibt Muthen [Mut10].

7.1.2 Modellanpassung

Die Anpassung des Modells geschieht schrittweise wie bei [MAB+09] und [Lar04]. Es werden unabhängig voneinander das optimale Mess- und Survivalmodell gesucht. Diese beiden Modelle werden dann miteinander kombiniert. Auf diese Weise kann nachvollzogen werden, zu welchem Teil die latenten Klassen, neben den Indikatorvariablen, durch die Kovariaten und die Ereigniszeit bestimmt werden [Lar04].
Die einzelnen Schritte der Modellanpassung sind wie folgt: Zunächst wird für die beobachteten Items die latente Klassenanalyse angepasst. Dafür wird das Modell für eine unterschiedliche Anzahl von Klassen berechnet. Die Entscheidung für das beste Modell kann unter anderem mithilfe des BIC, des Vuong-Lo-Mendell-Rubinson-Tests und der Entropie getroffen werden [Gei10, Lar04]. Ein Entscheidungskriterium können auch inhaltliche Motive und die Prävalenz der einzelnen Klassen sein.
Nachdem die optimale Anzahl latenter Klassen ermittelt wurde, kann das Modell um den Einfluss von Kovariaten erweitert werden. Es wird untersucht, welche der beobachteten Variablen einen signifikanten Einfluss auf die Definition der latenten Klassen besitzen.
Für das Survivalmodell sollte überprüft werden, ob die Proportional-Hazards-Annahme für alle Kovariaten erfüllt ist. Variablen, die die Annahme nicht erfüllen, können zeitabhängig in der Cox-Regression berücksichtigt werden. Modellfitmaße können dabei behilflich sein, das optimale Survivalmodell zu ermitteln.
Die gefundenen Modelle werden schließlich miteinander kombiniert. Kovariaten, die keinen signifikanten Einfluss auf die Ereigniszeit besitzen, werden schrittweise aus dem Modell entfernt, bis das finale Modell gefunden ist. Wie beim Survivalmodell mit latenten kontinuierlichen Variablen ist es an dieser Stelle auch möglich, Korrelationen zwischen Kovariaten zu berücksichtigen. Im Allgemeinen empfiehlt sich Larsens Methode zur Modellschätzung. Es kann überprüft werden, ob die Proportional-Hazards-Annahme auch für die latenten Klassen gilt. In [Lar04] wird dafür eine simulationsbasierte Methode vorgestellt. Einen Anhaltspunkt dafür, ob die klassenspezifischen Hazards proportional sind, bieten auch die Kaplan-Meier-Kurven in den Subgruppen.

7.2 Ein Anwendungsbeispiel: Herzfrequenzvariabilität als Prädiktor für kardiale Mortalität (fortgesetzt)

Als Anwendungsbeispiel wird noch einmal der HRV-Datensatz betrachtet. Es werden weiterhin die standardisierten Werte der Variablen verwendet. Im Folgenden werden Subgruppen von Patienten gesucht, die sich in Bezug auf ihre Ausprägungen in den HRV-Parametern ähnlich sind.
Im ersten Schritt der Modellanpassung wird die optimale latente Klassen-Lösung gesucht. Tabelle 7.1 gibt eine Übersicht über den Fit der verschiedenen Modelle.

Modell	Log-Likelihood	# freie Parameter	BIC	VM-Test
2 C	-6611	31	13415	p<0.001
3 C	-6313	42	12888	p<0.001
4 C	-6197	53	12724	p=0.382

Tabelle 7.1: Modellanpassung

Modelle mit 5 oder mehr Klassen konvergieren nicht. Das BIC spricht für das 4-Klassen-Modell, der Vuong-Lo-Mendell-Rubinson (VM)-Test befürwortet das 3-Klassen-Modell. Die Entscheidung soll hier zugunsten des sparsameren Modells getroffen werden, das ist die Lösung mit 3 latenten Klassen.

Klasse	Anz.	Prozent
C1	90	17.7
C2	289	56,9
C3	129	25,4

Tabelle 7.2: Klassenmitgliedschaft

Tabelle 7.2 gibt eine Übersicht über die Zuordnung der Individuen zu den jeweiligen Klassen. Die Individuen sind den Klassen zugeordnet, in denen sie die höchste Mitgliedswahrscheinlichkeit aufweisen.

	C1	C2	C3
C1	0.943	0.057	0.000
C2	0.028	0.952	0.019
C3	0.000	0.045	0.955

Tabelle 7.3: Klassifikationstabelle

Anhand der Klassifikationstabelle 7.3 kann beurteilt werden, wie gut sich die geschätzten

Klassen voneinander unterscheiden. In den Zeilen stehen die Klassen, denen ein Individuum zugeordnete wurde, in den Spalten sind die mittleren posterioren Wahrscheinlichkeiten für die einzelnen Kategorien angegeben. Hohe Werte auf der Hauptdiagonale sprechen für eine gute Klassifizierung. Diese ist hier erreicht, die Entropie beträgt 0.89.
Der Itemmeanprofileplot gibt eine Vorstellung davon, wie die drei Klassen definiert sind. Es liegt eine geordnete Lösung vor, das heißt, dass die erste Klasse für jedes Item die niedrigsten, die zweite Klasse überall etwas höhere und die dritte Klasse jeweils die höchsten Mittelwerte besitzt.

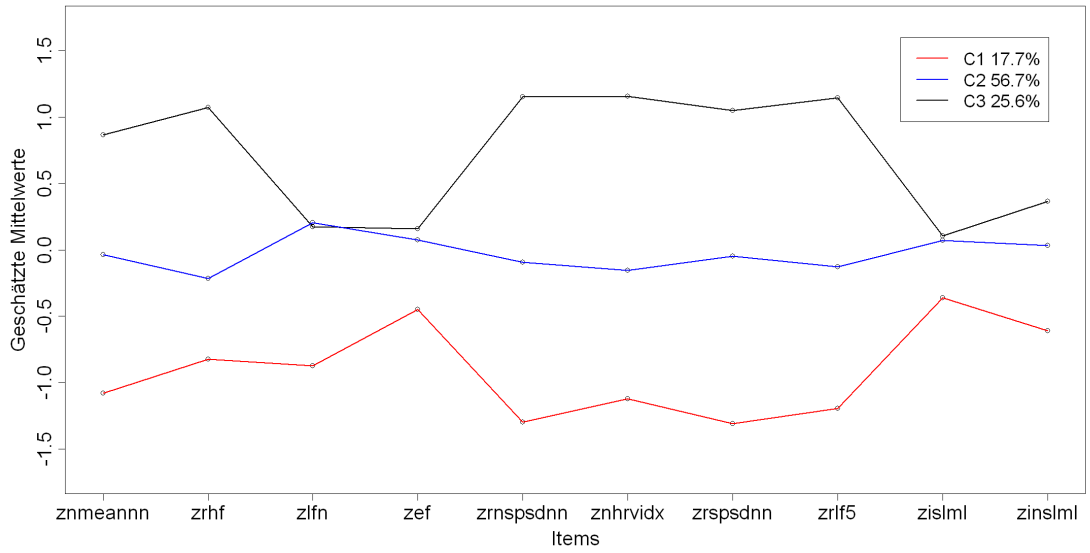

Abbildung 7.1: Itemmeanprofileplot

Im nächsten Schritt wird eine latente Klassenregression berechnet. Der Einfluss der Kovariate zlves auf die Definition der latenten Klassen ist nicht signifikant. Dieser Einfluss wird daher im Folgenden nicht weiter betrachtet. Als Survivalmodell wird das Proportional-Hazards-Modell mit einer Kovariate zlves verwendet.
Abschließend wird das gesamte Modell geschätzt. Es wird der Ansatz von Larsen verwendet. Um den Einfluss der Klassen auf die Überlebenszeit zu ermitteln, werden klassenspezifische Achsenabschnittsschätzer in der Cox-Regression berechnet. Die Ergebnisse finden sich in Tabelle 7.4.

Im ersten Abschnitt der Tabelle sind die mittleren erwarteten Werte der HRV-Parameter in den jeweiligen Klassen angegeben. Die Schätzer sind denen der latenten Klassenanalyse, die in Abbildung 7.2 dargestellt sind, sehr ähnlich. Die Patienten werden den Klassen zugeteilt, zu denen ihre Mitgliedswahrscheinlichkeit am größten ist. Die Anteile sind: C1:17.7%, C2:56.7%, C3:25.6%.
Die Patienten in Klasse 1 sind die mit der niedrigsten Herzfrequenzvariabilität, Mitglieder der Klasse 2 besitzen durchgehend mittlere Werte und die Patienten in Klasse 3 haben bei jeder Variable die höchsten Werte. Die meisten der Schätzer sind signifikant.

		Variable	Schätzer	SE	p-Wert	Haz. Ratio
C1	Means	zef	-0.46	0.15	<0.001	
	Means	znmeannn	-1.09	0.12	<0.001	
	Means	zrhf	-0.80	0.10	<0.001	
	Means	zlfn	-0.90	0.22	<0.001	
	Means	zrnspsdnn	-1.29	0.12	<0.001	
	Means	znhrvidx	-1.11	0.08	<0.001	
	Means	zrspsdnn	-1.31	0.14	<0.001	
	Means	zrlf5	-1.19	0.08	<0.001	
	Means	zinslml	-0.37	0.10	<0.001	
	Means	zislml	-0.61	0.12	<0.001	
C2	Means	zef	0.08	0.06	0.193	
	Means	znmeannn	-0.04	0.07	0.586	
	Means	zrhf	-0.22	0.06	<0.001	
	Means	zlfn	0.21	0.06	<0.001	
	Means	zrnspsdnn	-0.10	0.06	0.126	
	Means	znhrvidx	-0.16	0.06	0.008	
	Means	zrspsdnn	-0.05	0.06	0.455	
	Means	zrlf5	-0.13	0.07	0.053	
	Means	zinslml	0.07	0.07	0.299	
	Means	zislml	0.03	0.06	0.585	
C3	Means	zef	0.16	0.08	0.050	
	Means	znmeannn	0.86	0.09	<0.001	
	Means	zrhf	1.07	0.12	<0.001	
	Means	zlfn	0.18	0.06	0.006	
	Means	zrnspsdnn	1.15	0.08	<0.001	
	Means	znhrvidx	1.15	0.10	<0.001	
	Means	zrspsdnn	1.05	0.08	<0.001	
	Means	zrlf5	1.14	0.10	<0.001	
	Means	zinslml	0.11	0.09	0.209	
	Means	zislml	0.36	0.10	<0.001	
C1	Intercepts	Time	0.00	0.00	999.00	999.00
C2	Intercepts	Time	-1.16	0.56	0.038	0.31
C3	Intercepts	Time	-1.44	0.64	0.023	0.23
C1	Time on	zlves	0.86	0.28	0.002	2.36
C2	Time on	zlves	0.82	0.23	<0.001	2.27
C3	Time on	zlves	0.37	0.40	0.357	1.44

Tabelle 7.4: Survivalmodell mit latenten Klassen, standardisiert

Die nächsten Zeilen der Tabelle geben die Achsenabschnittsschätzer in der Hazard-Funktion für die einzelnen Klassen an. Zur Identifikation wurde der Parameter für die erste Klasse auf 0 gesetzt. Das Hazard Ratio für Klasse 2 beträgt 0.31, für Klasse 3 entsprechend 0.23. Dieser Wert ist so zu interpretieren, dass die Mortalität in Klasse 2 im Vergleich zu Klasse 1 um 70% verringert ist. Das Risiko zu versterben ist für Patienten der Klasse 3 im Vergleich zu

denen aus Klasse 1 um 84% verringert. Beide Schätzer gelten bedingt auf den Einfluss der Kovariaten und sind signifikant. Eine höhere Herzfrequenzvariabilität ist nach diesen Ergebnissen klar mit einer besseren Überlebenswahrscheinlichkeit assoziiert.
Im letzten Abschnitt der Tabelle sind die Parameterschätzer für zlves in der Cox-Regression aufgeführt. Die Koeffizienten wurden klassenabhängig geschätzt. In Klasse 1 und 2 ist der Einfluss von zlves besonders hoch, die Hazard Ratios betragen 2.36 und 2.27. Ein Anstieg von zlves um eine Standardabweichung erhöht das Risiko in C1 demnach um 136%, in C2 entsprechend um 127%. In Klasse 3 ist der Einfluss der Kovariate nicht signifikant.

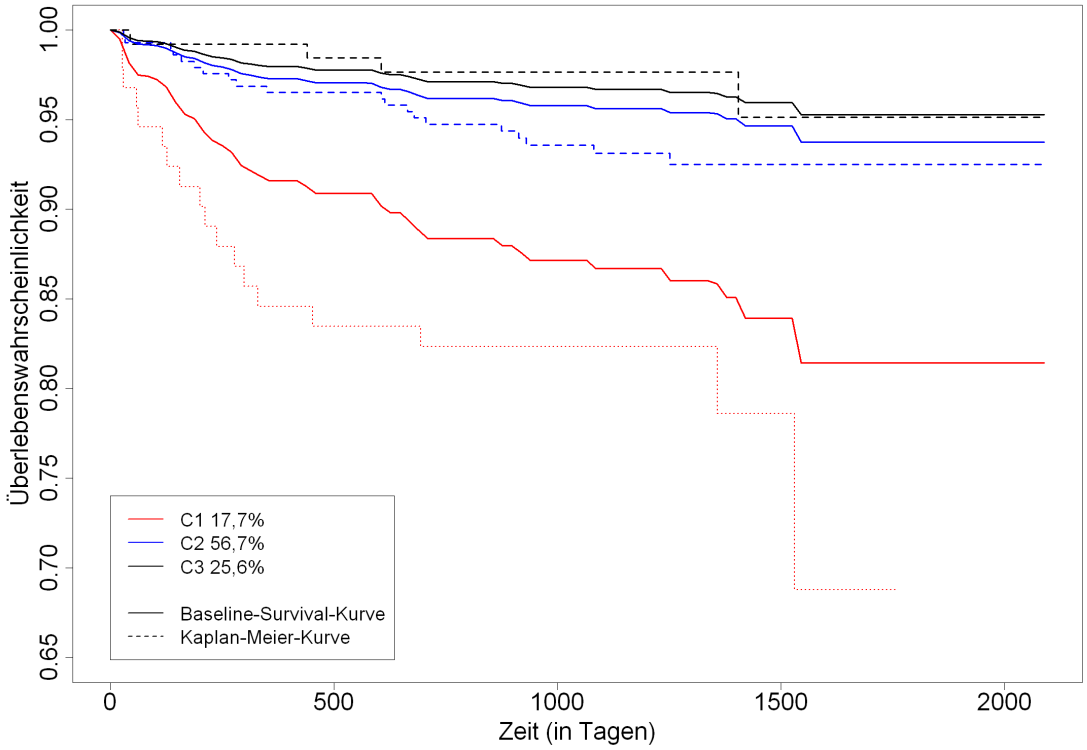

Abbildung 7.2: Kaplan-Meier-Kurven und geschätzte Baseline-Survival-Kurven der drei Klassen

In Abbildung 7.2 sind die Kaplan-Meier (KM)-Kurven und die geschätzten Baseline-Survival-Kurven der drei Klassen dargestellt. In der Graphik ist zu sehen, dass Patienten der Klassen 2 und 3 deutlich bessere Überlebenschancen im Vergleich zu denen aus Klasse 1 besitzen. Die KM-Kurven und die geschätzte Baseline-Survival-Kurve von Klasse 3 liegen nahe zusammen. Die Kaplan-Meier-Kurven zu Klasse 2 und 1 befinden sich konstant unter den jeweiligen Baseline-Survival-Kurven. Der Unterschied kommt durch den Einfluss von zlves zustande, der in diesen beiden Klassen besonders stark ist.
Bei der Modellbildung wurde nicht explizit untersucht, ob die Proportional-Hazards-Annahme zwischen den Klassen erfüllt ist. In Abbildung 7.2 sind die -Log-Log-Plots der geschätzten KM-Kurven abgebildet. Die Linien verlaufen weitestgehend parallel. Diese Graphik spricht dafür, dass die Proportional-Hazards-Annahme getroffen werden kann.

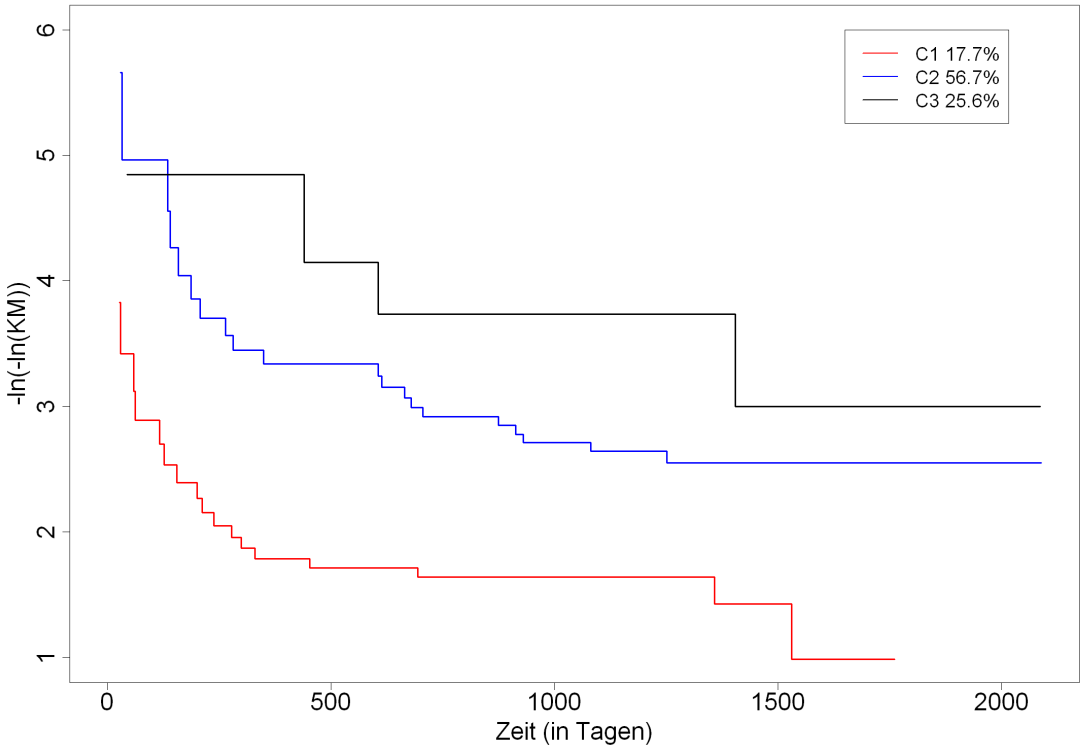

Abbildung 7.3: Evaluation der PH-Annahme zwischen den latenten Klassen mit -Log-Log-Plots

7.2.1 Vergleich der Modelle

Beim Vergleich mit der Cox-Regression besitzt die das Survivalmodell mit latenten Klassen ähnliche Vorteile wie das Survivalmodell mit latenten Faktoren. Das Modell bietet eine Lösung für das Problem der Kollinearität in der multiplen Cox-Regression mit stark korrelierten Items. Außerdem gibt es die Möglichkeit, den Einfluss von Kovariaten auf die Herzfrequenzvariabilitäts-Klassen zu untersuchen. Damit können neben den direkten auch indirekte Einflüsse auf die Ereigniszeit untersucht werden. Möglich gemacht wird das durch die simultane Schätzung der Parameter.
Durch den schrittweisen Prozess der Modellierung kann zur Theoriebildung beigetragen werden. Mithilfe von Modellfitmaßen können Annahmen eindeutig bestätigt oder abgelehnt werden. Dies ist einer der Vorzüge von Strukturgleichungsmodellen im Allgemeinen. Der Nachteil sind die zusätzlichen Modellannahmen, die im Gegenzug akzeptiert werden müssen.
Bei der Modellierung kann die Struktur der latenten Variablen untersucht werden. Das bedeutet bei der latenten Klassenanalyse, dass die optimale Anzahl von Subgruppen mit unterschiedlichen HRV-Profilen ermittelt werden kann. Die Hazard Ratios beziehen sich nicht wie bei der Standard-Cox-Regression auf einzelne Variablen, sondern auf die latenten Klassen. Itemmeanprofileplots bieten jedoch eine einfache Visualisierung der Zusammenhänge. Die

Beziehung zwischen den Klassen und der Überlebenszeit lässt sich außerdem gut durch die Darstellung der geschätzten Baseline-Survival-Kurven verdeutlichen.
Das erweiterte Modell bietet allerdings einige Limitationen, die im Zusammenhang mit der latenten Klassenanalyse stehen. Es ist empfehlenswert, wenn auch keine Voraussetzung, dass die betrachteten Items den gleichen Wertebereich besitzen. Im Blick auf die Interpretierbarkeit der latenten Klassen ist es sinnvoll, eine übersichtliche Anzahl von Indikatorvariablen zu verwenden. Die Anzahl der Klassen ist natürlich durch die Identifizierbarkeit des Modells begrenzt. Doch auch bei theoretisch identifizierten Modellen führt die Verwendung von vielen Items häufig zu Konvergenzproblemen.
Insgesamt kann man sagen, dass das Survivalmodell mit latenten Klassen im Vergleich zur Standard-Cox-Regression ein deutlich komplexeres Modell ist. Der Ansatz ermöglicht einen anderen Blickwinkel auf die Fragestellung. Es können zusätzliche Erkenntnisse über den Zusammenhang zwischen der Herzfrequenzvariabilität und dem Ereignis Koronartod gewonnen werden. Die Ergebnisse lassen sich leicht visualisieren und sind dadurch gut zu vermitteln.

Der Unterschied zwischen dem Survivalmodell mit latenten Faktoren und dem mit latenten Klassen liegt im Messmodell, das für die Indikatorvariablen verwendet wird. Anhand von Modellgütemaßen kann der Fit der unterschiedlichen Modelle verglichen werden. Beim HRV-Datensatz werden dazu die Ergebnisse aus den Tabellen 7.1 und 5.4 betrachtet. Das BIC ist bei allen latenten Klassenanalysen schlechter als bei beiden Faktorenanalysen. Für den HRV-Datensatz wäre also die Cox-Regression mit zwei latenten Faktoren das Modell der Wahl.

7.2.2 Evaluation der Modellstabilität mit Bootstrapping

Im Folgenden soll die Stabilität des Survivalmodells mit latenten Klassen mithilfe des Bootstrap-Ansatzes überprüft werden. Dazu wurden 250 Stichproben aus dem originalen Datensatz gezogen. Zur Unterscheidung von den beobachteten Koeffizienten und den Bootstrap-Schätzern und Standardfehlern wird die gleiche Notation wie in Abschnitt 5.2.4 verwendet.
Zunächst stellt sich die Frage, ob die geordnete latente Klassen-Lösung in Tabelle 7.4 replizierbar ist oder ob es mehrere Möglichkeiten gibt, wie die Patienten in drei Klassen mit unterschiedlichen HRV-Verläufen unterteilt werden können. Die geschätzten Mittelwerte der HRV-Parameter für alle 250 Replikationen sind in Abbildung 7.4 im Liniendiagramm dargestellt. Eine Parallele zur x-Achse kennzeichnet in jeder Graphik den erwarteten Mittelwert 0. Es ist zu sehen, dass es drei klar voneinander unterschiedene Profile gibt. Diese sind der geordneten Lösung aus 7.2 sehr ähnlich. Mitglieder der Klasse 3 zeigen bei allen Items die höchsten Werte, Patienten der Klassen 2 und 1 entsprechend mittlere und niedrige Ausprägungen. Dieses Ergebnis spricht dafür, dass die beobachtete Lösung in Tabelle 7.2 reliabel ist. Die Boxplots in Abbildung 7.4 sollen einen Eindruck von der Streuung der einzelnen Parameterschätzer vermitteln. Bis auf einzelne Ausreißer ist diese in allen drei Klassen verhältnismäßig gering.
Im Folgenden sollen die Ergebnisse für die Parameterschätzer in der Cox-Regression näher betrachtet werden. Das sind zum einen die Achsenabschnittschätzer und zum anderen die klassenspezifischen Koeffizienten von zlves. In Abbildung 7.5 sind zunächst die Bootstrap-Schätzer für die Achsenabschnitte dargestellt. Die beobachteten Parameter sind mit einer vertikalen Linie dargestellt. Um zu überprüfen, ob die Daten Normalverteilung aufweisen, ist in Abbildung 7.6 der zugehörige QQ-Plot dargestellt. Die Bootstrap-Schätzer zeigen eine

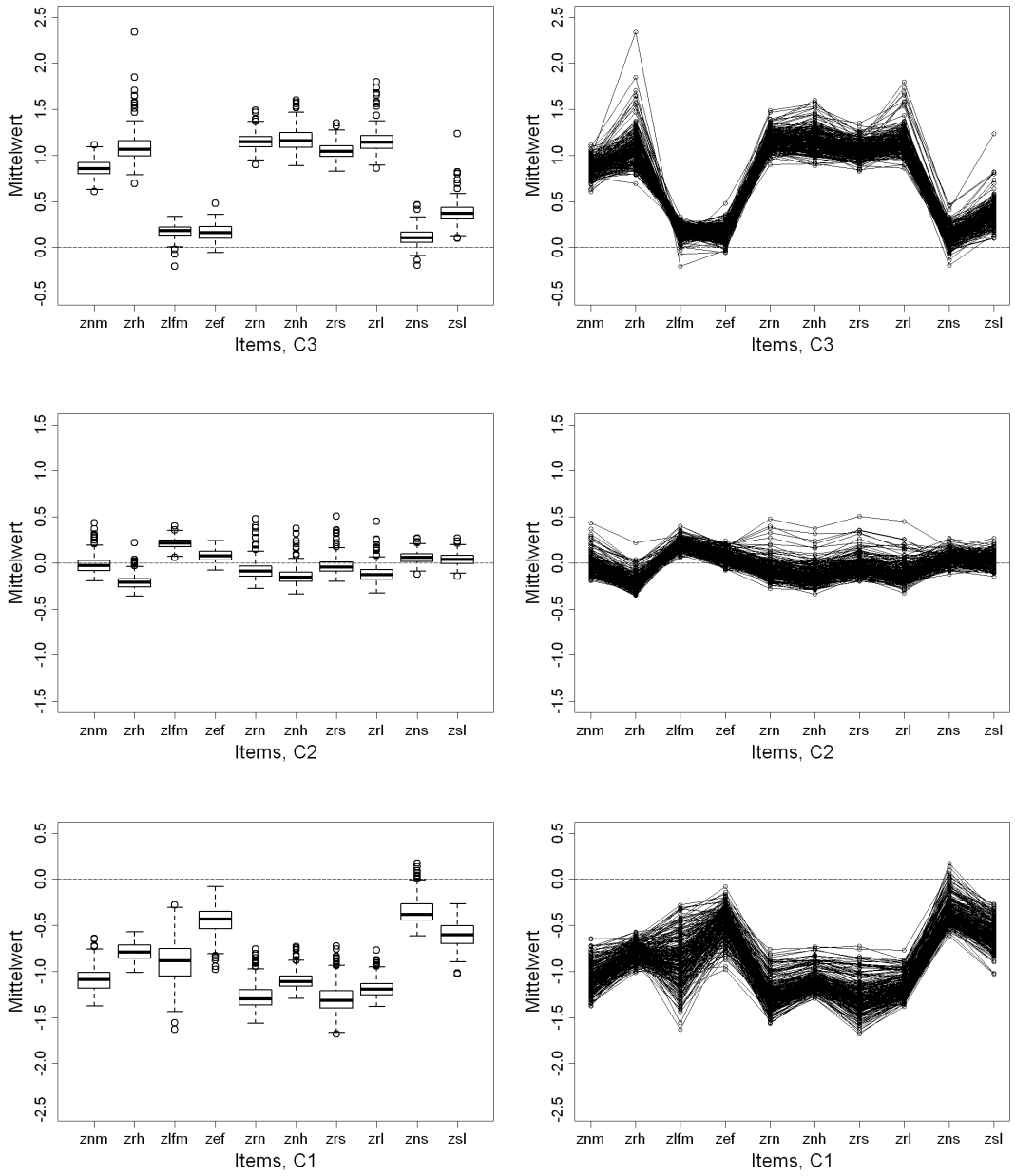

Abbildung 7.4: Bootstrap-Ergebnisse für die Mittelwerte der HRV-Parameter in den drei Klassen

approximative Normalverteilung. Histogramme für die anderen Parameter sind hier nicht abgebildet, da sie ähnliche Verteilungen zeigen.

Die Boxplots in Abbildung 7.7 geben einen Eindruck von der Streuung der Koeffizienten der Cox-Regression. Die beobachteten Parameter, Mittelwerte der Bootstrap-Replikationen, Bootstrap-Standardfehler und die zugehörigen approximativen Konfidenzintervalle sind in Tabelle 7.5 zusammengefasst.

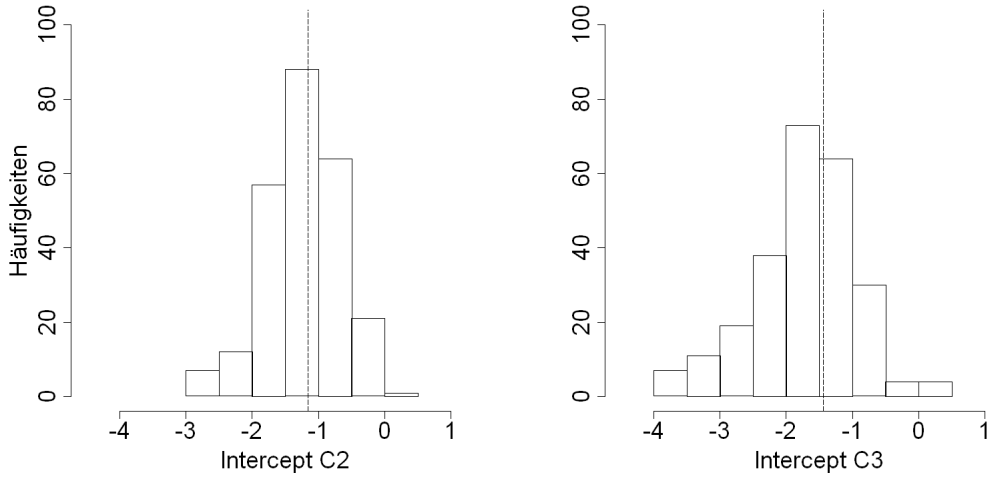

Abbildung 7.5: Bootstrap-Ergebnisse der Achsenabschnittschätzer von C2 und C3

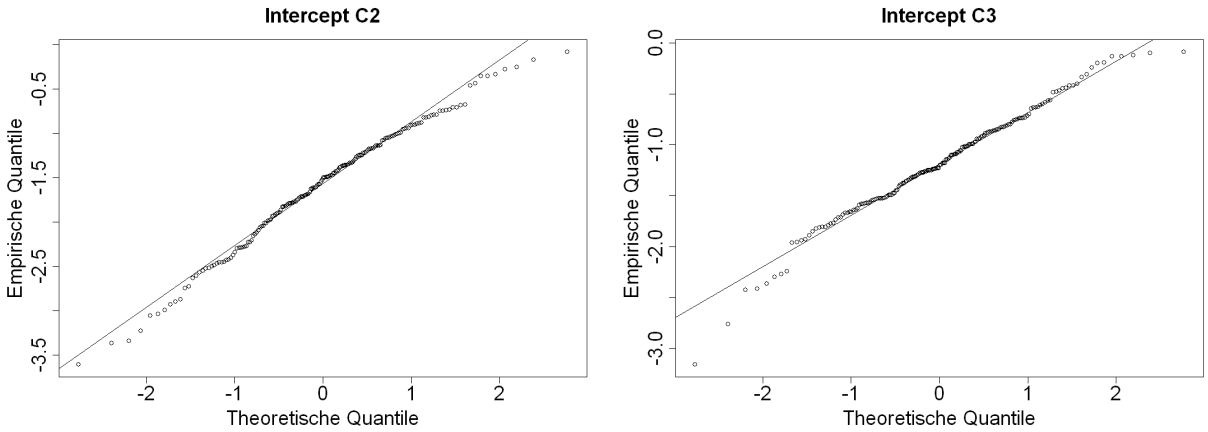

Abbildung 7.6: QQ-Plots für die Bootstrap-Replikationen

Der beobachtete Schätzer für den Einfluss von C2 ist nur geringfügig kleiner als der Bootstrap-Mittelwert. Die Streuung ist moderat. Das Konfidenzintervall lässt auf eine signifikant verbesserte Prognose von Patienten in Klasse 2 im Vergleich zu denen in Klasse 1 schließen. Der beobachtete Achsenabschnittsschätzer von C3 ist absolut etwas niedriger als der Bootstrap-Mittelwert. Die Streuung ist wesentlich größer. Das zugehörige Konfidenzintervall beinhaltet auch den Wert 0.
Die Koeffizienten von zlves $\hat{\theta}$ befinden sich für alle drei Klassen nah an den Bootstrap-Mittelwerten. Die Streuung für die Schätzer in C1 und C2 ist wesentlich geringer als die der Achsenabschnittsschätzer. Der Standardfehler in C3 ist bedeutend größer, das zugehörige Konfidenzintervall ist entsprechend breit.

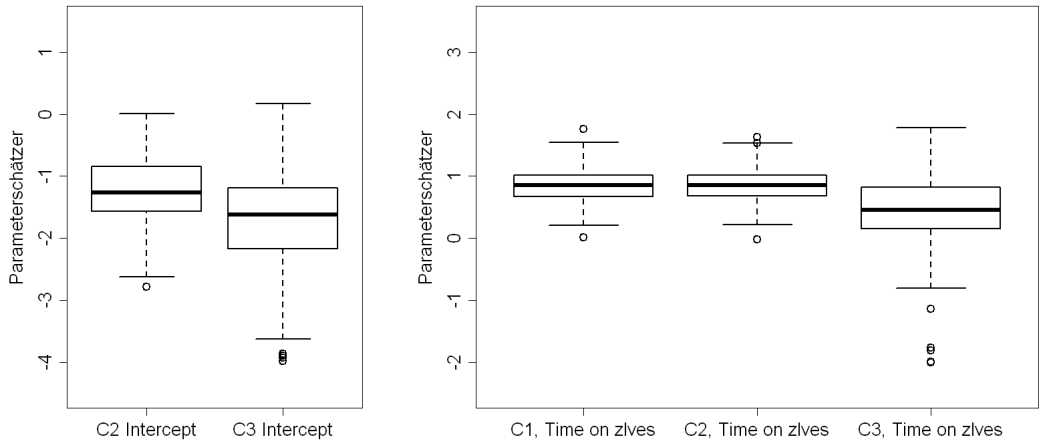

Abbildung 7.7: Bootstrap-Ergebnisse der Parameterschätzer in der Cox-Regression

Variable	B.str.-Mittelw. ($\hat{\theta}^*$)	B.str.-SE ($\hat{se}^*$)	Schätzer ($\hat{\theta}$)	95% B.str. - KI	
C2 Intercept	-1.25	0.55	-1.16	-2.23	-0.08
C3 Intercept	-1.72	0.78	-1.44	-2.96	0.08
C1 Time on zlves	0.85	0.28	0.86	0.31	1.41
C2 Time on zlves	0.84	0.26	0.82	0.30	1.34
C3 Time on zlves	0.43	0.58	0.37	-0.75	1.50

Tabelle 7.5: Übersicht der Bootstrap-Ergebnisse

Kapitel 8

Prognosegüte der Survivalmodelle mit latenten Klassen im Vergleich zur Cox-Regression

8.1 Ein Anwendungsbeispiel: Nutzungsdauer von Prepaidkarten für Mobiltelefone

Im Folgenden soll ein zweites Beispiel betrachtet werden, um die Anwendungsmöglichkeiten der Cox-Regression mit latenten Klassen zu verdeutlichen. Für diese Arbeit steht der Datensatz eines deutschen Mobilfunkanbieters zur Verfügung. Das Geschäftsmodell des Unternehmens besteht in dem Vertrieb von Prepaid-Karten mit einem einfachen Telefontarif. Die Prepaid-Karte wird beim Kauf aktiviert. Das Guthaben kann auf verschiedene Arten aufgeladen werden. Kunden, die über einen Zeitraum von 12 Monaten kein sogenanntes "Topup" vorgenommen haben, werden deaktiviert. Die Hauptfragestellung des Mobilfunkanbieters besteht darin, frühzeitig vorhersagen zu können, wie lange ein Kunde die Prepaid-Karte nutzen wird. Mithilfe einer solchen Prognose sollen Werbemaßnahmen optimal gesteuert werden. Es handelt sich dabei um ein typisches Time-to-Event Problem, bei dem das betrachtete Ereignis die Deaktivierung eines Kunden ist.
Es stehen die Daten aller Personen, die sich zwischen dem 01.07.2006 und dem 01.07.2007 registriert haben, zur Verfügung. Damit liegen die Informationen zu 240347 Klienten vor. Die Kunden wurden bis zum 01.07.2008 nachbeobachtet, folglich ist einen Beobachtungszeitraum von maximal 2 Jahren pro Person gegeben.
Die große Datenmenge ermöglicht es, die Modellierung an einer Stichprobe des Datensatzes durchzuführen und die Prognosegüte des gefundenen Modells anschließend an einer zweiten Stichprobe zu bewerten. Zunächst werden daher zwei Zufallsstichproben von je 10000 Individuen aus dem Datensatz gezogen: Ein Modelldatensatz, auf den unterschiedliche Modelle angepasst werden, und ein Testdatensatz, der später zur Evaluation dieser Modelle dient.

Im Modelldatensatz liegt die mittlere Follow-Up Zeit bei 16.79 Monaten. In dieser Zeit sind 709 Deaktivierungen dokumentiert. Die Einflussvariablen auf den Nutzungszeitraum teilen sich in die Angaben der Kunden bei Vertragsabschluss und die laufenden Informationen während der Nutzung auf. In Tabelle 8.1 ist eine Übersicht über die Baseline-Informationen gegeben.

Parameter		Anzahl	in Prozent
Tarif	Klassik	7409	74.1
	Ethno	2591	25.9
Bestellherkunft	online	4662	46.6
	Fachhandel/Cash&Carry	5338	53.4
Email	nein	4137	41.4
	ja	5863	58.6
Newsletter	nein	3531	35.3
	ja	6469	64.7
Geschlecht	m	6257	62.6
	w	3743	37.4
Alter (in Jahren)	MW=40.48 ; SD=13.02		

Tabelle 8.1: Baseline-Informationen

Die Variable Tarif gibt an, ob ein Kunde den "Klassik"-Tarif verwendet, den Standardtarif des Unternehmens, oder den "Ethno"-Tarif, mit höheren Preisen für Inlandsgespräche und besonders günstigen Auslandstarifen. Bei dem Parameter Bestellherkunft sind die möglichen Ausprägungen "Fachhandel" (3.5%) und "Cash and Carry" (49.8%) zusammengefasst, um Kategorien mit zu geringen Ausprägungen zu vermeiden. Eine explorative Analyse hat gezeigt, dass sich die beiden Kategorien kaum in ihrem Deaktivierungsrisiko unterscheiden. Die Variable E-Mail gibt Auskunft darüber, ob beim Kauf die E-Mail-Adresse angegeben wurde. Entsprechend gibt die Variable Newsletter an, ob ein Kunde beim Erwerb der Karte eingewilligt hat, den Newsletter des Unternehmens zu empfangen.
Neben den Baseline-Variablen wurden im Verlauf monatliche Nutzungsdaten der Kunden dokumentiert. Besonders wichtige Einflussgrößen auf die Dauer der Kartenverwendung sind die Telefongebühren eines Kunden. Tabelle 8.2 gibt eine Übersicht über die Umsätze in den ersten fünf Monaten nach Aktivierung der Prepaid-Karte.

Umsatz je Monat	Mittelwert	Minimum	Maximum	Median	SD	Anteil $\neq 0$
1	7.48	0.00	876.74	1.89	19.25	74.3
2	12.40	0.00	1297.59	5.05	28.34	83.1
3	11.41	0.00	843.87	3.99	24.38	79.2
4	10.62	0.00	884.60	3.47	23.23	76.6
5	10.19	0.00	591.36	3.05	23.06	74.3

Tabelle 8.2: Monatliche Umsätze in den ersten fünf Monaten

Die letzte Spalte der Tabelle gibt den Anteil der Kunden an, die irgendeinen Umsatz in dem jeweiligen Monat hatten. Insgesamt gab es 537 Kunden (5.4%), die in keinem der ersten drei Monate die Prepaid-Karte verwendet haben.
Die monatlichen Umsätze sind nicht symmetrisch verteilt. Da bei der latenten Klassenanalyse besonders schief verteilte Daten zu Schätzungsproblemen führen, werden die Werte transformiert. Zu den Umsatzwerten wird eine Konstante von 1 addiert und von dem Ergebnis wird

der natürliche Logarithmus gebildet. Die Verteilung der Variable vor und nach der Transformation ist in Abbildung C.1 im Anhang dargestellt. In den folgenden Modellen werden die transformierten Parameter verwendet.
Zu den Nutzungsdaten, die im Zeitverlauf dokumentiert wurden, gehört auch der verwendete Tarif. Innerhalb der ersten fünf Monate findet bei keinem Kunden ein Wechsel statt, sodass es genügt, den Tarif zum Baseline-Zeitpunkt zu betrachten. Die Kunden des Mobilfunkanbieters haben die Möglichkeit, verschiedene Einstellungen für die automatische Aufladung (ATU) der Telefonkarte zu machen. Die automatische Aufladung kann ausgestellt sein, es ist möglich, ein monatliches zeitgesteuertes Topup einzurichten oder ein guthabenbasiertes Topup zu verwenden. Im ersten Monat sind 99.5% der Kunden der ersten, 0.1 % der zweiten und 0.4 % der dritten Kategorie zuzuordnen. In einer explorativen Analyse zeigt sich, dass weniger der gewählte Tarif von prognostischem Wert für die Deaktivierungszeit ist, als die Information, ob ein Kunde in den betrachteten Monaten seine ATU-Einstellung wechselt. Der Wechsel von einer aktivierten zu einer deaktivierten ATU ist so selten, dass alle Kunden, die in irgendeine Richtung gewechselt sind, zusammengefasst werden. Im Folgenden wird eine Variable ATU-Wechsel als Prädiktor in den Modellen verwendet. Tabelle 8.3 gibt eine Übersicht über den Parameter zu unterschiedlichen Zeitpunkten.

ATU-Wechsel bis Monat	Anzahl	in Prozent
3	704	7.0
4	804	8.0
5	816	8.2

Tabelle 8.3: Änderung der automatischen Aufladeeinstellung (ATU)

Außerdem wurde dokumentiert, ob die Besitzer der Prepaid-Karte Neukunden werben. Tabelle 8.4 gibt einen Überblick über den Anteil der Kunden, die bis zu den Monaten drei bis fünf mindestens einen Neukunden geworben haben.

Neukunden geworben bis Monat	Anzahl	in Prozent
3	809	8.1
4	911	9.1
5	997	10.0

Tabelle 8.4: Werbung von Neukunden

Im Testdatensatz beträgt die mittlere Follow-Up Zeit ebenfalls 16.79 Monate, es traten 720 Ereignisse auf. Die Baseline-Informationen unterscheiden sich nicht signifikant von den Werten im Modelldatensatz. Dies gilt für alle betrachteten Parameter.

Es soll eine multiple Cox-Regression angepasst werden, die den Einfluss aller beschriebenen Einflussvariablen auf die Nutzungsdauer untersucht. Bei Betrachten der Einflussgrößen ist zu erkennen, dass hier wieder ein Fall vorliegt, bei dem die Kovariaten im Cox-Modell stark

miteinander korreliert sind. Offensichtlich hängen die Telefonkosten eines Kunden in einem Monat mit denen im darauffolgenden Monat zusammen. Ob die Nutzer immer hohe, mittlere oder niedrige Umsätze haben oder ob es verschiedene Verläufe der Kosten gibt, bleibt zu ermitteln. Die Latente-Klassen-(LCA)-Survivalanalyse scheint ein sinnvolles Modell für die vorliegende Fragestellung zu sein. Es werden latente Klassen von Nutzern modelliert, die sich in Bezug auf ihre monatlichen Telefonkosten ähnlich sind. Dieses Modell berücksichtigt jedoch nicht die Reihenfolge der Umsätze. Aus diesem Grund wird außerdem ein Latente-Klassen-Growth- (LCGA)-Survivalmodell angepasst. Dieses Modell untersucht ebenfalls, ob es Subgruppen von Nutzern mit ähnlichen Umsatzverläufen gibt, berücksichtigt dabei aber die longitudinale Datenstruktur der Indikatorvariablen.
Dem Mobilfunkanbieter ist daran gelegen, möglichst früh vorhersagen zu können, welche Kunden deaktiviert werden. Daher wird zunächst betrachtet, wie gut die Prognose der Modelle ist, wenn nur die Umsätze der ersten drei Monate berücksichtigt werden. Anschließend werden auch Modelle mit allen Informationen bis zu den Monaten 4 und 5 angepasst. Es wird untersucht, ob und in welchem Maße sich die Vorhersagekraft dadurch verbessern lässt.

8.1.1 Cox-Regression

Die Überprüfung der proportionalen Hazardsannahme der Einflussgrößen wird mithilfe von -Log-Log-Survival-Plots durchgeführt. Dazu werden die kontinuierlichen Variablen auf verschiedene Weise kategorisiert. Alle Graphiken zeigen vollkommen parallele Verläufe (hier nicht dargestellt). Demzufolge kann die Annahme getroffen werden.
Es werden univariate und eine multiple Cox-Regression berechnet. Es werden die logarithmierten Umsätze (lUmsatz) der Monate 1 bis 3 verwendet, außerdem werden sogenannte Zustandsvariablen (ZUmsatz) betrachtet, die angeben, ob der Umsatz des jeweiligen Monats ungleich 0 war. In die Modelle gehen alle zuvor beschriebenen Baseline-Parameter ein sowie die Information zur Änderung der ATU-Einstellung und Werbung von Neukunden. In den Tabellen 8.5 und 8.6 sind die Ergebnisse aufgeführt.

In der multiplen Regression ist der Einfluss der Variable Geschlecht nicht signifikant und wurde daher entfernt. Die Ergebnisse zeigen, dass alle anderen Einflussgrößen einen hoch signifikanten Einfluss auf die Nutzungsdauer haben.
In den univariaten Cox-Regressionen ist zu sehen, dass ein höheres Alter, die Angabe der E-Mail-Adresse, eine Anforderung des Newsletters und die Werbung von Neukunden mit einer verringerten Deaktivierungschance assoziiert sind. Eine Bestellung, die im Fachhandel/Cash&Carry getätigt wurde, eine Nutzung des Tarifs Ethno und ein Wechsel der ATU-Einstellung erhöhen hingegen die Chance einer Deaktivierung. Höhere Umsätze in den ersten zwei Monaten der Nutzungsdauer steigern die Wahrscheinlichkeit eines Ereignisses, häufiges Telefonieren im dritten Monat verringert die Chance der Deaktivierung. Ein Umsatz ungleich 0 im ersten Monat erhöht die Chance einer Deaktivierung, in den Monaten 2 und 3 haben Kunden, die ihre Prepaid-Karte genutzt haben, ein deutlich geringeres Risiko.
In der multiplen Cox-Regression sehen die Effekte ähnlich aus, allerdings kehrt sich der Einfluss der Variable Newsletter unter Berücksichtigung der anderen Einflussgrößen um. Das Einverständnis, das Rundschreiben zu empfangen, steigert hier die Wahrscheinlichkeit für ein Ereignis. Außerdem kehrt sich der Einfluss der Zustandsvariablen für den Umsatz im ersten Monat um. Unter Beachtung der anderen Einflussgrößen führt auch ein Umsatz von 0 in

Parameter	Schätzer	SE	p-Wert	Haz. Ratio
Geschlecht	-0.31	0.080	<0.001	0.73
Alter	-0.01	0.003	<0.001	0.99
E-Mail	-0.92	0.076	<0.001	0.40
Newsletter	-0.63	0.078	<0.001	0.54
Bestellherkunft	0.86	0.080	<0.001	2.36
Tarif	0.93	0.087	<0.001	2.54
ATU-Wechsel M3	0.65	0.111	<0.001	1.91
Neukunden M3	-1.54	0.262	<0.001	0.22
lUmsatz M1	0.41	0.029	<0.001	1.51
lUmsatz M2	0.11	0.031	<0.001	1.11
lUmsatz M3	-0.25	0.033	<0.001	0.79
ZUmsatz M1	0.37	0.097	<0.001	1.45
ZUmsatz M2	-0.55	0.095	<0.001	0.58
ZUmsatz M3	-1.42	0.078	<0.001	0.24

Tabelle 8.5: Univariate Cox-Regressionen

Parameter	Schätzer	SE	p-Wert	Haz. Ratio
Alter	-0.01	0.003	0.015	0.99
E-Mail	-0.64	0.143	<0.001	0.53
Newsletter	0.43	0.132	0.001	1.54
Bestellherkunft	0.38	0.115	0.001	1.46
Tarif	0.41	0.111	<0.001	1.52
ATU-Wechsel M3	1.17	0.121	<0.001	3.22
Neukunden M3	-1.21	0.262	<0.001	0.30
lUmsatz M1	0.38	0.047	<0.001	1.46
lUmsatz M2	0.19	0.055	0.001	1.21
lUmsatz M3	-0.21	0.055	<0.001	0.81
ZUmsatz M1	-0.09	0.149	0.560	0.91
ZUmsatz M2	-0.35	0.161	0.029	0.70
ZUmsatz M3	-1.03	0.149	<0.001	0.36

Tabelle 8.6: Multiple Cox-Regression

Monat 1 zu einem erhöhten Deaktivierungsrisiko.
Die Regressionen geben einen klaren Eindruck von den Effekten der einzelnen Einflussfaktoren. Dennoch scheint die Berücksichtigung der monatlichen Umsätze als separate Parameter nicht sinnvoll. Eine hohe Korrelation zwischen diesen Variablen ist zu erwarten und sollte modelliert werden. Es wäre interessant zu sehen, ob es tatsächlich unterschiedliche Klassen von Nutzern gibt und wie diese mit der Ereigniszeit zusammenhängen. Für den Mobilfunkanbieter wäre damit eine Einordnung der Kunden in unterschiedliche Risikoklassen möglich.
Im Folgenden wird die Latente-Klassen-Growth-Survivalanalyse vorgestellt. Um die vorliegenden Fragestellungen beantworten zu können, wird anschließend das Modell auf den Datensatz angepasst.

8.1.2 Latente-Klassen-Growth-Analyse

Die latente Klassenanalyse bietet eine Möglichkeit, den Zusammenhang zwischen mehreren ähnlichen Items zu modellieren. Der Faktor Zeit geht hier, wie bereits erwähnt, nicht in die Analyse mit ein, die Reihenfolge der Messpunkte bleibt unberücksichtigt.
Es kann vorkommen, dass die multiplen Indikatoren zu wiederholten Messungen eines Parameters zu unterschiedlichen Zeitpunkten gehören. Die individuellen Unterschiede im Verlauf werden üblicherweise mit linearen gemischten Modellen bzw. im Kontext von latenten Variablen mit sogenannten Random-Effect-Growth-Modellen analysiert. Diese Modelle können erweitert werden, wenn man annehmen muss, dass die untersuchte Population aus verschiedenen Subgruppen besteht, die sich wesentlich in ihrem Verhalten unterscheiden. Es können latente Klassen modelliert werden, für die separate Verlaufskurven der Items über die Zeit geschätzt werden. Solche Modelle heißen Growth-Mixture-Modelle und sollen im Folgenden betrachtet werden. Latente-Klassen-Growth-Modelle sind ein Spezialfall dieser Modellklasse.

Modell

Es wird ein quadratisches Growth-Mixture-Modell betrachtet, dass durch drei zufällige Effekte η_{0i}, η_{1i} und η_{2i} und einen zeitabhängigen Residualterm ϵ beschrieben wird [Mut01].

$$y_{it} = \eta_{0i} + \eta_{1i} x_{it} + \eta_{2i} x_{it}^2 + \epsilon_{it} \tag{8.1}$$

Hier ist y_i ein $(1 \times p)$ Vektor stetiger Items, die zu den Zeitpunkten $t = 1, 2 \ldots$ gemessen werden. Es wäre zusätzlich möglich, den Einfluss von Kovariaten auf die Responsevariable zu berücksichtigen. Im Folgenden wird angenommen, dass die Messungen der y_i für alle Individuen zu den gleichen diskreten Zeitpunkten geschehen, sodass $x_{it} = x_t$ gilt. Die zeitabhängigen Residuen ϵ_t besitzen einen Mittelwert von 0 und eine Kovarianzmatrix Θ mit unterschiedlichen Varianzen und Elementen auf der Nebendiagonale zur Darstellung der residualen Korrelation über die Zeit.
Im Kontext der latenten Variablenmodelle werden die zufälligen Effekte als Growth-Faktoren bezeichnet. Es wird nun angenommen, dass sich die Verläufe in k latenten Klassen unterscheiden. Demnach gilt für die drei Growth-Faktoren:

$$\eta_{0i} = \alpha_{0k} + \zeta_{0i} \tag{8.2}$$
$$\eta_{1i} = \alpha_{1k} + \zeta_{1i} \tag{8.3}$$
$$\eta_{2i} = \alpha_{2k} + \zeta_{2i}. \tag{8.4}$$

Diese drei Variablen werden als Intercept-, Slope- und Quadratischer-Growth-Faktor bezeichnet. Dabei geben die α´s Mittelwerte des jeweiligen Faktors an. Es wird angenommen, dass die Residuen ζ einen Mittelwert von 0 und eine Kovarianzmatrix Ψ besitzen. Beide Residualmatrizen, Θ und Ψ, können klassenspezifisch modelliert werden.
Ein spezieller Typ von Growth-Mixture-Modellen ist die Latente-Klassen-Growth-Analyse (LCGA). Dabei wird angenommen, dass die Verlaufskurven der Individuen innerhalb einer Klasse homogen sind. Dafür werden die Varianz- und Kovarianzparameter in den einzelnen Klassen auf 0 gesetzt. Variation zwischen den Individuen ist in diesem Modell durch die zeitabhängigen Varianzterme weiterhin erlaubt. Das Gebiet der Latenten-Klassen-Growth-Modelle wurde insbesondere von Nagin, siehe [Nag05], entwickelt und ist auch in SAS in der Prozedur "Proc Traj" implementiert.
Es gibt kein klares Entscheidungskriterium dafür, ob ein Growth-Mixture-Modell oder die Latente-Klassen-Growth-Analyse verwendet werden soll. Die Schätzung des Modells wird durch die Restriktion der Varianzen deutlich vereinfacht und kann damit eine Lösung bei Konvergenzproblemen sein. Das restringierte Modell hat außerdem den Vorteil, dass die Klassen klarer geschätzt werden. Auf der anderen Seite werden dabei im Allgemeinen mehr Klassen benötigt. Die Latente-Klassen-Growth-Modelle sind der Latenten-Klassenanalyse durch ihre Annahme von Homogenität in den Subgruppen besonders ähnlich [Mut04a]. Insbesondere aus diesem Grund wird im Folgenden das LCGA-Modell auf den Datensatz angewendet. Bei der Modellanpassung werden die geschätzten Mittelwerte mit den beobachteten Werten verglichen. So kann ein Eindruck davon gewonnen werden, ob tatsächlich auf klassenspezifische Varianzterme verzichtet werden kann [JW08]. Einen praktischen Grund für die Modellwahl gibt es außerdem: Das Growth-Mixture-Modell benötigt mindestens vier Zeitpunkte, um auch einen quadratischen Growth-Faktor schätzen zu können, mit dem Latente-Klassen-Growth-Modell ist das bereits bei der Betrachtung von drei Zeitpunkten möglich.
Analog zum LCA-Modell kann das LCGA-Modell mit der Cox-Regression kombiniert werden. Der Einfluss der latenten Kategorien, die hier durch unterschiedliche Mittelwerte der Growth-Faktoren charakterisiert sind, wird in der Hazard-Funktion durch einen klassenspezifischen Achsenabschnittsschätzer berücksichtigt:

$$h_i(t|x_i, c_i = k) \quad = \quad h_{0k}(t) * exp(\iota_k + \beta_k x_i). \tag{8.5}$$

Die Schätzung des Modells geschieht mit der Maximum-Likelihood-Methode. Für eine genauere Beschreibung des Verfahrens sei an dieser Stelle auf [AMM06] und [MA08] verwiesen. Der Einfluss der latenten Klassen auf die Überlebenszeit kann analog zum LCA-Survivalmodell auf zwei Wege geschätzt werden. Hier wird der Ansatz von Larsen (2004) verwendet, siehe Abschnitt 7.1.1. Das heißt, es wird angenommen, dass sich die Baseline-Hazard-Funktion der einzelnen Klassen nur durch einen einzigen multiplikativen Faktor unterscheiden. Alternativ wäre es möglich, komplett unrestringierte Baseline-Hazard-Funktionen für jede latente Klasse zu schätzen. Abbildung 8.1 zeigt das schematische Pfaddiagramm für das Latente-Klassen-Growth-Survivalmodell. Die Anwendung von Growth-Mixture-Survivalmodellen sind in der Literatur bislang kaum zu finden. Ein Beispiel ist bei [MAB$^+$09] gegeben. Im Folgenden soll das Modell auf den vorliegenden Datensatz angepasst werden.

Anwendung auf den Datensatz

Die Modellanpassung geschieht schrittweise. Zunächst werden LCGA-Modelle für die logarithmierten Umsatzparameter der ersten drei Monate für eine unterschiedliche Anzahl von Klas-

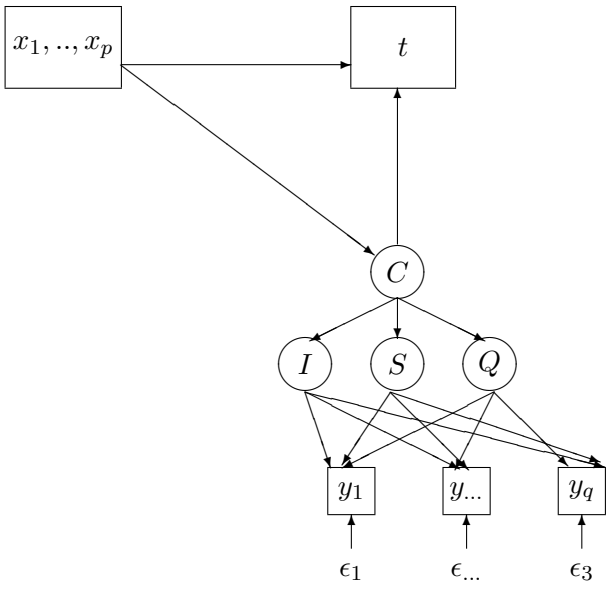

Abbildung 8.1: Schematisches Pfaddiagramm für das LCGA-Survivalmodell

sen berechnet. Diese Modelle verwenden jeweils einen Intercept-, Slope- und Quadratischen-Growth-Faktor. Die Faktorladungen, x_t, werden passend zu den äquidistanten Messzeitpunkten auf 0, 1, und 2 gesetzt. Die Entscheidungskriterien zur Wahl eines Modells sind die gleichen wie bei der Latenten-Klassenanalyse. Dazu gehören das BIC und der Vuong-Lo-Mendell-Rubinson-Test [JW08]. Im nächsten Schritt wird das gefundene LCGA-Modell mit der Cox-Regression kombiniert. Das gesamte Modell berücksichtigt den Einfluss der Kovariaten auf die latenten Klassen sowie den Einfluss der Kovariaten und der latenten kategoriellen Variablen auf die Überlebenszeit. Es werden die gleichen Kovariaten wie in der multiplen Cox-Regression verwendet.

Bei der Modellanpassung zeigt sich, dass sich der Modellfit gemessen am BIC immer weiter verbessert, desto mehr Ausprägungen die latente kategorielle Variable hat. Dem Vuong-Lo-Mendell-Rubinson-Test zufolge sollte das Modell mit 12 latenten Klassen gewählt werden. Abgesehen von den statistischen Kriterien gibt es hier jedoch Limitationen bei der praktischen Modellierung. Bei einem Modell mit zu vielen Kategorien werden einzelnen Klassen nur noch sehr wenige Kunden zugeordnet. Wird die Latente-Klassen-Growth-Analyse dann mit der Cox-Regression kombiniert, kann das zu Konvergenzproblemen führen. Ohne Probleme lässt sich bei dem betrachteten Datensatz ein Modell mit 9 latenten Klassen berechnen. Dabei werden jeder Klasse mehr als 3% der Kunden zugeordnet.
Im vorliegenden Fall ist der Mobilfunkanbieter an einer Lösung mit einer übersichtlichen Anzahl von Klassen interessiert. Betrachtet werden daher Modelle mit nur 2, 3 oder 4 Kategorien für die Umsatzverlaufskurven. Es zeigt sich, dass das Modell mit 4 Klassen schon sehr gut in der Lage ist, die Kunden zu charakterisieren. Dieses Modell soll im Folgenden vorgestellt und diskutiert werden. Die detaillierten Ergebnisse für das Modell mit 9 latenten Klassen sind im Anhang gezeigt.

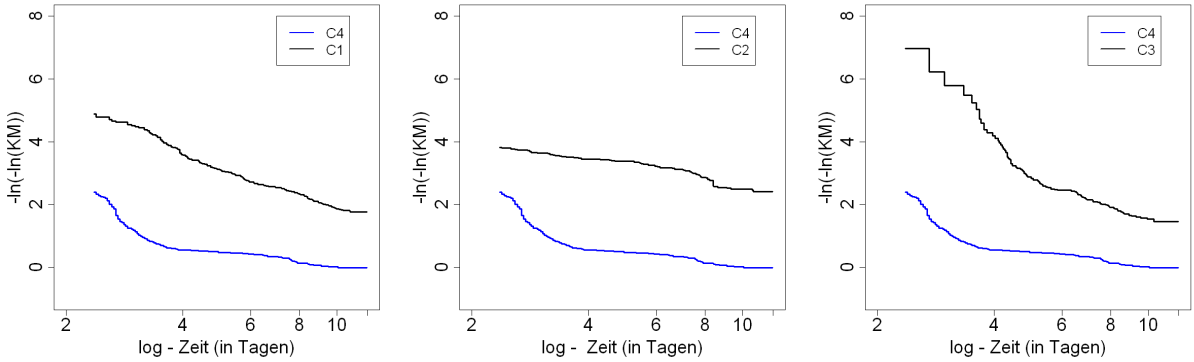

Abbildung 8.2: -Log-Log-Plots für das LCGA-Survialmodell mit 4 Klassen

4-Klassen-LCGA-Modell

Tabelle 8.7 gibt die Ergebnisse des Latenten-Klassen-Growth-Survivalmodells an. Die Faktorladungen sind $x_t = 0, 1, 2$. Es sind nur die Parameterschätzer der latenten Klassenregression aufgeführt, die im Vergleich zur Referenzkategorie signifikant sind. Die Klasse mit der höchsten Deaktivierungswahrscheinlichkeit ist C4, der zugehörige Achsenabschnittsschätzer in der Cox-Regression wurde auf den Wert 0 gesetzt, damit dient diese Klasse als Referenzkategorie. Die Parameterschätzer der multinomialen logistischen Regression sind ebenfalls in Bezug auf C4 zu interpretieren.
Es wird mithilfe von -Log-Log-Survival-Plots untersucht, ob die Annahme proportionaler Hazards für die ermittelte latente kategorielle Variable erfüllt ist. Das Ergebnis ist in Abbildung 8.2 zu sehen. Es zeigt sich, dass die Kurven aller Kategorien im Vergleich zur Referenzkategorie weitestgehend parallel verlaufen.

Im ersten Teil von Tabelle 8.7 finden sich klassenspezifische Intercept-, Slope- und Quadratische-Growth-Parameter, diese werden im Folgenden auch mit I, S und Q bezeichnet. Diese Werte geben die klassenspezifischen Umsatzverläufe an. Die Interpretation der Parameter lässt sich an Klasse 1 verdeutlichen. Kunden in dieser Kategorie weisen zu Beginn niedrige Umsätze auf, diese steigen im Folgenden an und bleiben danach konstant. Zu so einem Verlauf gehören ein niedriger I-Wert, ein positiver S-Wert und ein negativer Q-Wert. Die Umsatzkurven für die einzelnen Klassen sind in Abbildung 8.3 dargestellt. Hier ist zu sehen, dass sich die Klassen C1, C2 und C3 im Wesentlichen in ihrem Niveau unterscheiden. Einen vollkommen anderen Verlauf der erwarteten monatlichen Telefonumsätze zeigt jedoch die Klasse C4.
Im zweiten Teil von Tabelle 8.7 sind die Kovariaten aufgeführt, die einen signifikanten Einfluss in der multinomialen logistischen Regression besitzen. Mit Hilfe dieser Ergebnisse können die Eigenschaften der einzelnen Klassen, zusätzlich zu den erwarteten Umsatzverlaufskurven, näher beschrieben werden. In der letzten Spalte der Tabelle sind hier die zugehörigen Odds Ratios angegeben. Es zeigt sich zum Beispiel, dass für Kunden, die beim Kauf ihre E-Mail-Adresse angeben, die Chance C1 und nicht C4 zugeordnet zu werden, um den Faktor 1.55

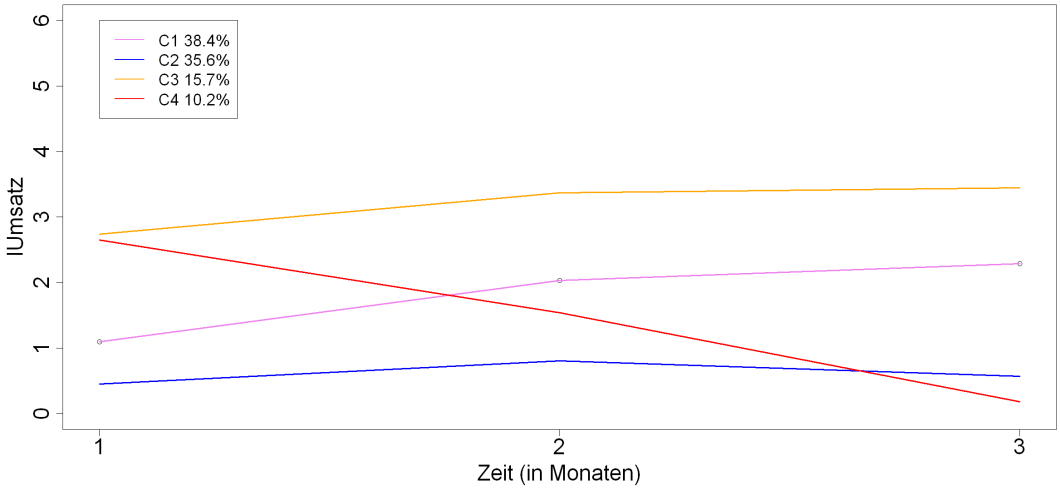

Abbildung 8.3: Geschätzte Umsatzverläufe in den latenten Klassen

erhöht ist. Des Weiteren ist zu sehen, dass die Chance, einer der ersten drei Klassen im Vergleich zu C4 anzugehören, für Kunden mit einem Ethno-Tarif verringert ist. Eine mögliche Erklärung ist, dass Nutzer des Ethno-Tarifes nur zeitlich begrenzt in Deutschland sind und danach aufhören, die Prepaid- Karte zu nutzen. Dieser Effekt findet sich nun in den Klassenschätzern von C4 wieder, deren Mitglieder eben diese fallenden Umsatzverlaufskurven aufweisen.
Im dritten Teil von Tabelle 8.7 sind die Parameterschätzer für die Einflüsse auf die Zeit bis zur Deaktivierung aufgeführt. In der letzten Spalte der Tabelle stehen dazu die resultierenden Hazard Ratios. Es sind zunächst die klassenspezifischen Achsenabschnittsschätzer in der Cox-Regression angegeben. Die Klasse mit der höchsten Deaktivierungschance ist C4 mit 10.2% der Kunden. Diese Kategorie ist durch relativ hohe Umsätze im ersten Monat charakterisiert, die in den beiden darauffolgenden Monaten linear fallen, vgl. Abbildung 8.3. Die beste Vorhersage haben die Kunden in Klasse 2. Dabei handelt es sich nicht um die Klienten mit den höchsten Telefonkosten. Stattdessen ist diese Klasse durch verhältnismäßig niedrige, aber konstante Umsätze charakterisiert. Im Vergleich zu C4 ist die Chance einer Deaktivierung für Kunden in C2 um 91% verringert. Der Unterschied ist hoch signifikant. Ähnlich geringe Deaktivierungsrisiken besitzen die Kunden in Klasse 1 und 3. C1 ist durch einen mittleren Umsatz zu Beginn charakterisiert, der in den beiden folgenden Monaten ansteigt. Im Vergleich zu C4 ist die Chance für ein Ereignis in dieser Klasse um 87% verringert. Die erwarteten Umsätze sehen in Klasse 3 ähnlich aus, nur der Achsenabschnitt ist hier nach oben verschobenen. Die Deaktivierungschance in dieser Klasse ist im Vergleich zu C4 ebenfalls um 87% verringert.
Im letzten Teil von Tabelle 8.7 sind außerdem Parameterschätzer für die direkten Einflüsse der Kovariaten auf das Ereignis aufgeführt. Diese sind nicht klassenspezifisch geschätzt. Die Hazard Ratios sind denen aus der multiplen Cox-Regression relativ ähnlich, die Schätzer der Variablen Newsletter, Tarif und ATU Wechsel sind ist in diesem Modell etwas höher. Alle Schätzer gelten selbstverständlich adjustiert für den Einfluss der anderen Kovariaten.

Klasse		Parameter	Schätzer	SE	p-Wert	0R/HR
C1	Means	Intercept	1.09	0.05	<0.001	
C1	Means	Slope	1.29	0.12	<0.001	
C1	Means	Quadratic	-0.34	0.04	<0.001	
C2	Means	Intercept	0.45	0.03	<0.001	
C2	Means	Slope	0.66	0.06	<0.001	
C2	Means	Quadratic	-0.30	0.02	<0.001	
C3	Means	Intercept	2.74	0.12	<0.001	
C3	Means	Slope	0.91	0.21	<0.001	
C3	Means	Quadratic	-0.28	0.06	<0.001	
C4	Means	Intercept	2.65	0.04	<0.001	
C4	Means	Slope	-0.99	0.18	<0.001	
C4	Means	Quadratic	-0.12	0.09	0.149	
C1	On	Alter	0.01	0.00	0.010	1.01
C1	On	E-Mail	0.44	0.17	0.011	1.55
C1	On	Bestellherkunft	-0.99	0.15	<0.001	0.37
C1	On	Tarif	-0.83	0.14	<0.001	0.44
C1	On	Neukunden (ja/nein) M3	0.60	0.23	0.010	1.82
C2	On	Alter	0.04	0.00	<0.001	1.04
C2	On	E-Mail	0.52	0.18	0.004	1.68
C2	On	Bestellherkunft	-1.07	0.16	<0.001	0.34
C2	On	Tarif	-0.42	0.11	<0.001	0.66
C3	On	Alter	-0.01	0.00	0.011	0.99
C3	On	Bestellherkunft	-0.71	0.18	<0.001	0.49
C3	On	ATU Wechsel M3	0.89	0.26	0.001	2.44
C3	On	Neukunden (ja/nein) M3	0.84	0.25	0.001	2.32
C1	Intercepts	Time	-2.04	0.14	<0.001	0.13
C2	Intercepts	Time	-2.42	0.18	<0.001	0.09
C3	Intercepts	Time	-2.04	0.15	<0.001	0.13
C4	Intercepts	Time	0.00	0.00	999.00	999.00
C1-C4	Time on	Alter	-0.01	0.00	0.092	0.99
C1-C4	Time on	E-Mail	-0.70	0.15	<0.001	0.50
C1-C4	Time on	Newsletter	0.53	0.14	<0.001	1.70
C1-C4	Time on	Bestellherkunft	0.32	0.12	0.007	1.38
C1-C4	Time on	Tarif M3	0.64	0.11	<0.001	1.90
C1-C4	Time on	ATU Wechsel M3	1.31	0.13	<0.001	3.71
C1-C4	Time on	Neukunden (ja/nein) M3	-1.29	0.26	<0.001	0.28

Tabelle 8.7: Ergebnis des LCGA-Survivalmodells mit 4 Klassen

Abbildung 8.4 zeigt für die vier Klassen die geschätzten und beobachteten individuellen Umsätze. Zur Übersichtlichkeit sind hier nur die Verlaufskurven für eine zufällige Stichprobe von 20% der Personen abgebildet. Diese Graphik gibt einen Eindruck von der Varianz innerhalb der einzelnen Klassen.
In Abbildung 8.5 sind die geschätzten Baseline-Hazard-Survival-Kurven für die vier latenten Klassen dargestellt. Die Kurven verlaufen vollkommen parallel, was an der Spezifikation des Modells liegt. Die Abstände entsprechend den zugehörigen Achsenabschnittsschätzern aus der Cox-Regression. Ist der Einfluss aller Kovariaten gleich 0, so liegt die Wahrscheinlichkeit, dass Kunden der Klasse C1 am Ende der Beobachtungszeit noch die Prepaid-Karte nutzen, bei 85.7%. Unter der gleichen Annahme ist zu erwarten, dass nach 24 Monaten in C2 noch 90.0% der Kunden die Karte nutzen und in C3 noch 85.8%. Deutlich schlechter sieht die Prognose für Klasse 4 aus. Hier liegt die geschätzte Baselinehazard-Survival-Kurve am Ende des Beobachtungszeitraums bei nur 30.7%.

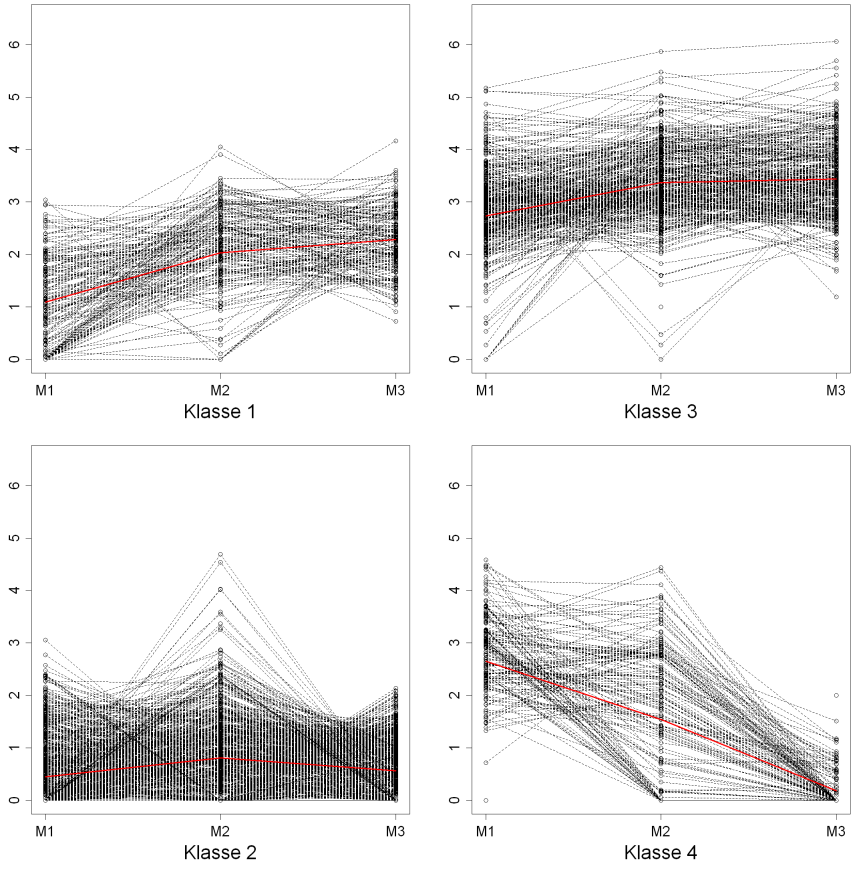

Abbildung 8.4: Geschätzte und beobachtete individuelle Umsatzverläufe

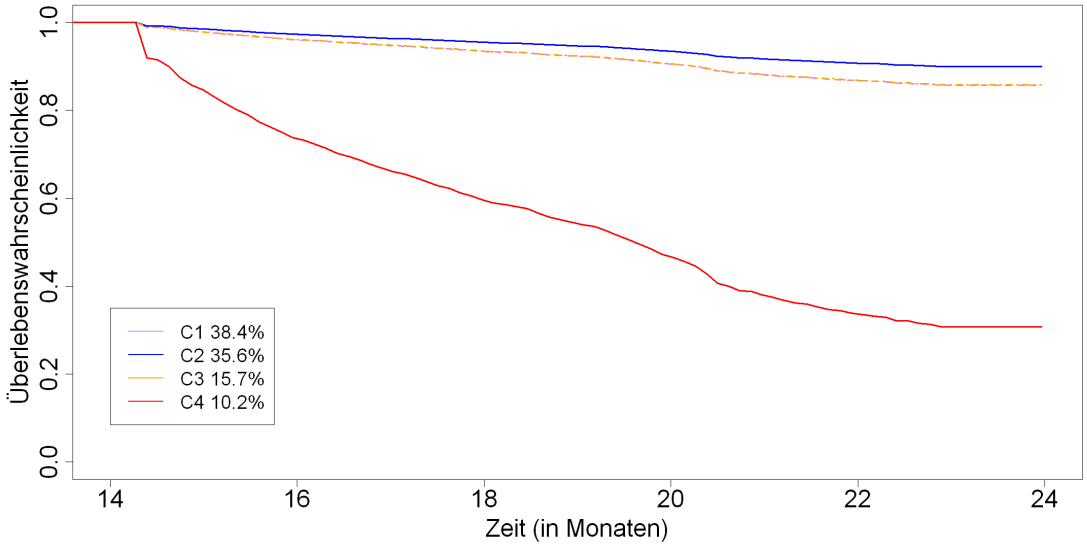

Abbildung 8.5: Geschätzte Baseline-Survival-Kurven der latenten Klassen

8.1.3 Survivalanalyse mit latenten Klassen

Wenn nur drei Indikatorvariablen für die latente Klassenvariable verwendet werden, führen das LCGA- und das LCA-Survivalmodell auf die gleichen Ergebnisse. Daher erübrigt sich die Präsentation der Ergebnisse des Latenten-Klassen-Survivalmodells an dieser Stelle. Anstelle von Intercept-, Slope- und Quadratischer-Growth-Parametern schätzt das Modell die Mittelwerte in den vier latenten Klassen. Diese lassen sich auch aus den Parametern in 8.7 berechnen und korrespondieren dann genau mit den Verlaufskurven in Abbildung 8.3.
Bei der Verwendung von vier oder mehr Indikatorvariablen für die latente kategorielle Variable führen das LCA- und das LCGA-Survivalmodell auf unterschiedliche Ergebnisse. Die Modellanpassung der Survivalanalyse mit latenten Klassen geschieht dabei weiterhin schrittweise, wie in Abschnitt 7.1.2 beschrieben.

8.1.4 Vergleich der Modelle

Nachdem die Modelle berechnet wurden, sollen sie im Folgenden auf ihren prognostischen Wert hin untersucht werden. Es gibt verschiedene Maße, um die prädiktive Genauigkeit eines Survivalmodells zu beurteilen. Das bekannteste Maß ist die Fläche unter der Receiver-Operating-Characteristic-Kurve, kurz ROC-AUC. Die traditionelle Anwendung für einen binären Endpunkt wurde für Survival Endpunkte erweitert. Mit dieser Methode kann für jeden beliebigen Ereigniszeitpunkt eine ROC-Kurve berechnet werden. Es gibt verschiedene Möglichkeiten, zeitabhängige Sensitivitäten und Spezifitäten zu definieren. Im Folgenden wird die sogenannte inzident/dynamische Definition verwendet, siehe [HZ05]. Die Fläche unter der ROC-Kurve kann als die Wahrscheinlichkeit interpretiert werden, dass bei der Betrachtung von zwei zufällig ausgewählten Personen diejenige mit einem Ereignis einen höheren Marker aufweist, als die Person ohne Ereignis. Bei einem Cox-PH-Modell ist der Marker eines Individuums der lineare Prädiktor, $\sum \beta X_i$.

In vielen Fällen ist für die Beurteilung der Prognosegüte kein spezifischer Zeitpunkt von Interesse. Ein globales Maß für die Genauigkeit eines Modells bietet der Concordance (C)-Index [HZ05]. Diese Größe ist als die Wahrscheinlichkeit definiert, dass von zwei Personen diejenige mit dem früheren Ereigniszeitpunkt einen höheren Marker besitzt. Es wird die Annahme getroffen, dass die Beobachtungen (Marker und Ereigniszeit) der Individuen voneinander unabhängig sind und die Ereigniszeiten stetig gemessen werden. Der C-Index ist dann ein gewichteter Durchschnitt der Flächen unter den zeitabhängigen inzident/dynamischen ROC-Kurven. Die Gewichte hängen hierbei von der studienspezifischen Zensierungsverteilung ab [HZ05].
Eine neuere Größe, um die prädiktive Genauigkeit eines Modells zu bewerten, ist die Net Reclassification Improvement (NRI) [Pea08]. Es werden Risikoprognosen auf Basis eines alten und eines neuen Survivalmodells betrachtet. Für Personen mit und ohne ein Ereignis wird getrennt beurteilt, wie sich die Risikobewertung verändert. Die NRI gibt an, ob die Klassifikation insgesamt besser oder schlechter geworden ist. Diese Methode ist in der Lage, auch geringe Veränderungen der Prognosegüte zu zeigen, wenn ein Modell bereits eine hohe Trennschärfe besitzt [Pea08].

Die Survival-ROC gilt als Standardmaß [Pea08]. Im Folgenden sollen daher zunächst die ROC-Kurven für einen bestimmten Zeitpunkt und der C-Index für den gesamten Beobachtungszeitraum berechnet werden. Die Berechnung wird mit dem R-Paket "risksetROC" von Heagerty und Zheng durchgeführt. Es gibt dabei nicht die Möglichkeit, Standardfehler oder Konfidenzintervalle für die Schätzer zu berechnen. Folglich kann keine Aussage darüber getroffen werden, ob signifikante Unterschiede zwischen den Modellen vorliegen. Die Autoren des Paketes verweisen darauf, dass die Varianzschätzung mithilfe von Bootstrapping-Verfahren durchgeführt werden kann [HZ05]. Zusätzlich werden die Modelle anhand der kontinuierlichen NRI verglichen. Die Berechnung erfolgt mit der Funktion "improveProb" aus dem R-Paket "Hmisc" von Frank E. Harrell Jr. [HJ12].

Bei der Berechnung der Gütemaße wird bei allen drei Modellen schrittweise vorgegangen. Bei der multiplen Cox-Regression werden nur die Parameterschätzer des angepassten Modells benötigt. Mit diesen ist es möglich, den linearen Prädiktor für jedes Individuum des Testdatensatzes zu berechnen.
Die Parameterschätzer des LCGA-Survivalmodells werden gespeichert. Die Personen im Testdatensatz werden mit einer Latenten-Klassen-Growth-Analyse, unter Verwendung der Parameterschätzer des kombinierten Modells, den latenten Kategorien zugeordnet. Aufgrund der Klassenzugehörigkeit und Kovariatenausprägungen wird für jedes Individuum der lineare Prädiktor berechnet.
Ähnlich wird bei der Survivalanalyse mit latenten Klassen vorgegangen. Parameterschätzer aus dem kombinierten Modell werden benutzt, um mit einer latenten Klassenregression die Individuen den einzelnen Kategorien zuzuordnen. Die gespeicherten Parameterschätzer für den Einfluss der Klassen und Kovariaten werden dann verwendet, um den linearen Prädiktor zu errechnen.

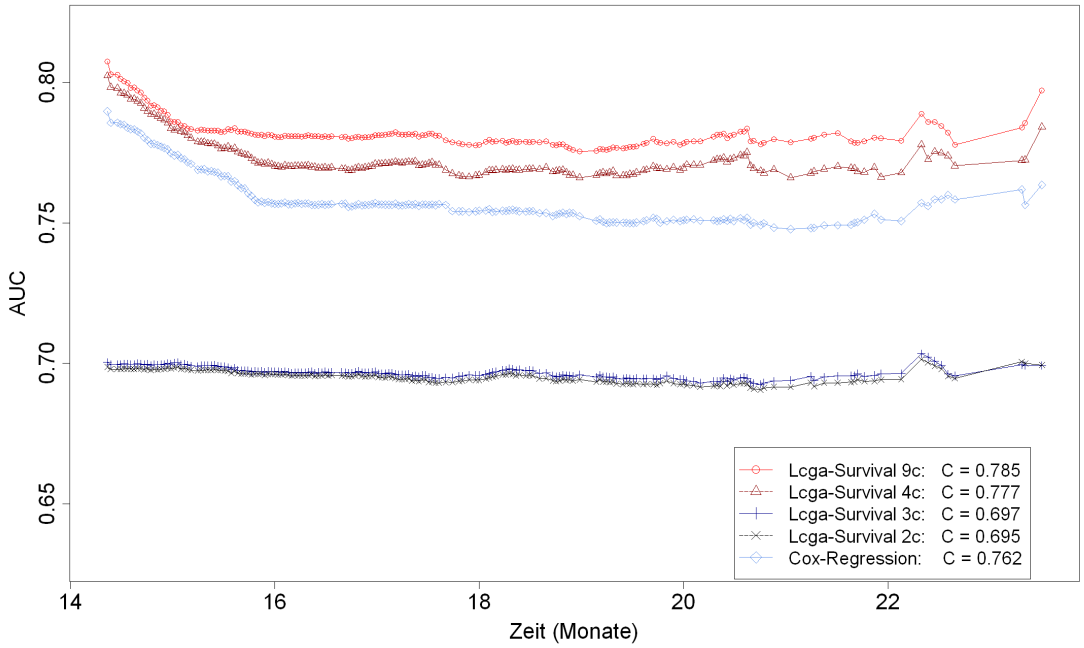

Abbildung 8.6: AUC der Modelle im Zeitverlauf

Vergleich anhand der AUC

In Abbildung 8.6 ist die Fläche unter der ROC-Kurve für die Cox-Regression und verschiedene LCGA-Survivalmodelle im Zeitverlauf abgebildet. Des Weiteren ist der C-Index für den gesamten betrachteten Beobachtungszeitraum angegeben. Die Abbildung beginnt bei Monat 14 nach der Aktivierung. Kunden können frühestens nach einem Jahr deaktiviert werden. Das erste Ereignis wird jedoch erst nach etwa 14 Monaten beobachtet, bis dahin sind nur Zensierungen dokumentiert.
In der Graphik ist zu sehen, dass zwei Kurven eine besonders niedrige AUC besitzen, dass sind die Modelle mit zwei und drei latenten Klassen. Offensichtlich genügt diese Anzahl von Kategorien nicht für eine gute Prognose. Die AUC der beiden Modelle liegt konstant bei nur 0.70. Die beste Prognosegüte besitzt das latente Variablenmodell mit 9 Klassen. Die AUC des LCGA-Survivalmodells mit 4 Klassen liegt jedoch nur wenig niedriger. Dieses Modell ist das Modell der Wahl, das in Abschnitt 8.1.2 vorgestellt wurde. Es ist interessant zu sehen, dass dieses viel sparsamere Modell, beim Vergleich der Prognosegüte anhand der AUCs, bereits sehr gut abschneidet.
Das LCGA-Survivalmodell mit 4 Klassen soll nun im direkten Vergleich zur Cox-Regression betrachtet werden. Es ist zu sehen, dass die prognostische Genauigkeit des latenten Variablenmodells zu jedem Zeitpunkt über der des Cox-Modells liegt. Der AUC(t) des LCGA-Survivalmodells mit 4 Klassen befindet sich zu Beginn ungefähr bei 0.80 und fällt danach deutlich ab, auf einen Wert von etwa 0.77 für $16 < t < 20$. Demnach ist in dieser Periode die Wahrscheinlichkeit, dass ein Kunde, der zu einem Zeitpunkt t deaktiviert wird, einen höheren Marker besitzt als ein Kunde, der zum gleichen Zeitpunkt noch aktiviert ist, immer größer als 77%. Der C-Index für den gesamten Zeitraum beträgt 0.777. Auch bei der Cox-

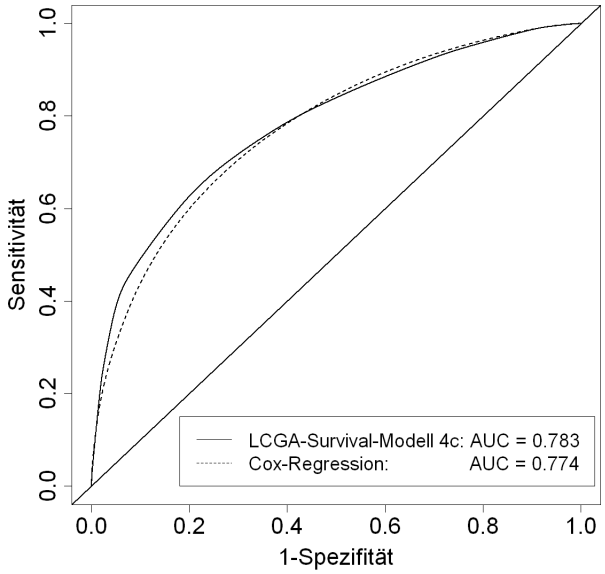

Abbildung 8.7: ROC-Kurven für das LCGA-Survivalmodell mit 4 Klassen und die Cox-Regression zum Zeitpunkt 15 Monate

Regression ist die kurzfristige Prognosefähigkeit, mit etwa 0.78, besser als die langfristige. Diese liegt für $16 < t < 20$ etwa bei 0.75. Unter Betrachtung des gesamten Zeitraums kommt die Cox-Regression auf einen Wert von 0.762 für den Concordance-Index. Demnach ist die Wahrscheinlichkeit, dass die Vorhersagen von einem zufälligen Paar von Kunden mit dem Outcome übereinstimmen, 76.2%.
Im Folgenden soll die prognostische Genauigkeit der Modelle zu einem bestimmten Zeitpunkt genauer betrachtet werden. Es wurde der Zeitpunkte t=15 gewählt, repräsentativ für die kurzfristige prognostische Genauigkeit. In Abbildung 8.7 sind die zugehörigen ROC-Kurven für das LCGA-Survivalmodell und die Cox-Regression abgebildet. Aus der Graphik kann abgelesen werden, welche Sensitivität für eine bestimmte Rate von Falsch-Positiven erwartet werden kann. So ist zu sehen, dass bei dem latenten Variablenmodell eine Sensitivität von 70% bedeutet, dass 27% der Fälle falsch-positiv klassifiziert werden. Bei der Cox-Regression ist bei einer Sensitivität von 70% eine Spezifität von 71% zu erwarten. Beide Kurven unterscheiden sich deutlich von der Diagonale, bei der die Fläche unter der Kurve 0.5 beträgt.
Tabelle 8.8 gibt noch einmal einen Überblick über die berechneten Ergebnisse. Im Vergleich zu dem latenten Variablenmodell besitzt die Cox-Regression nach 15 Monaten eine etwa 1% schlechtere Wahrscheinlichkeit, das Outcome richtig zu prognostizieren.

	AUC Monat 15	C-Index
LCGA-Survivalmodell 4c	0.783	0.777
Cox-Regression	0.774	0.762

Tabelle 8.8: Vergleich des LCGA-Survivalmodells mit 4 Klassen und der Cox-Regression bezüglich der prognostischen Genauigkeit

Ereignis	Reklassifikation nach unten		nach oben			
nein	2979	(50.0%)	2983	(50.0%)	5962	(100%)
ja	112	(47.7%)	123	(52.3%)	235	(100%)
	3091	(49.9%)	3106	(50.1%)	6197	(100%)

Tabelle 8.9: Vergleich des LCGA-Survivalmodells mit 4 Klassen und der Cox-Regression anhand der NRI

Vergleich anhand der NRI

Der Unterschied zwischen den beiden Modellen zu t=15 ist nicht besonders groß. Als weiteres Vergleichsmaß wird daher die Net Reclassification Improvement berechnet. Es werden alle Kunden betrachtet, die eine Nachverfolgungszeit von mindestens 15 Monaten besitzen. Verglichen wird die Risikobewertung des LCGA-Survivalmodells mit der der Cox-Regression. Dazu wird die geschätzte Ereigniswahrscheinlichkeit für jedes Individuum zum Zeitpunkt 15 Monate betrachtet. Liegt die geschätzte Wahrscheinlichkeit des latenten Variablenmodells über der der Cox Regression, so gilt eine Person als "nach oben" klassifiziert, liegt sie niedriger, so zählt der Kunde als "nach unten" klassifiziert. Die Ergebnisse sind in Tabelle 8.9 gegeben.
Insgesamt haben 6197 Personen eine Nachverfolgungszeit von mindestens 15 Monaten. Bis zu diesem Zeitpunkt sind 235 Ereignisse aufgetreten. Bei den Personen mit Ereignis liegt die Netto-Reklassifikation bei 52.3%-42.2%=4.7%. Für die Kunden ohne ein Ereignis liegt dieser Wert bei -0.07%. Die gesamte NRI beträgt demnach 4.6% mit dem Konfidenzintervall (-8.4%;17.7%). Dieser Unterschied ist nicht signifikant (p=0.488).
Vergleicht man für den gleichen Zeitpunkt die Vorhersagegüte des LCGA-Survivalmodells mit 9 Klassen mit der der Cox-Regression, so zeigt sich ein signifikanter Unterschied (p=0.002). Die NRI beträgt 20.3% mit dem Konfidenzintervall (7.3%;33.3%).

Ergebnisse

Beim Vergleich der AUC im Zeitverlauf liegt die Kurve des LCGA-Survivalmodells mit 4 Klassen zu jedem Zeitpunkt klar über der der Cox-Regression. Betrachtet man die ROC-Kurve zu einem bestimmten Zeitpunkt, hier wurden 15 Monate gewählt, zeigt sich weiterhin der Vorteil des latenten Variablenmodells. Dieser ist jedoch relativ gering. Die Berechnung der NRI für t= 15 Monate bestätigt diesen Eindruck. Das LCGA-Survivalmodell mit 4 Klassen besitzt eine bessere Prognosegüte als die Cox-Regression, der Unterschied ist jedoch nicht signifikant.
Ein 4-Klassenmodell wurde hier gewählt, um dem Mobilfunkanbieter eine relativ einfache Klassifizierungsregel für die Kunden anzubieten. Während in der Cox-Regression für jeden Kunden anhand seines genauen Umsatzes im ersten, zweiten und dritten Monat eine Prognose gemacht wird, ordnet das latente Variablenmodell die Kunden in vier Klassen mit unterschiedlichen Basisrisiken ein. Bei der Berechnung der Ereigniswahrscheinlichkeit wird dann nur der Einfluss der Klassenzugehörigkeit berücksichtigt. Die Ergebnisse weisen darauf hin, dass dieses Modell eine mindestens genauso gute Prognose liefert wie die Cox-Regression.
Um eine signifikant bessere Risikoklassifikation als die der Cox-Regression zu erzielen, genügt das sparsame latente Variablenmodell aber nicht. Eine überlegene Prognose kann jedoch mit

dem LCGA-Modell mit 9 Klassen gemacht werden. Das zeigt die signifikant bessere NRI für den Vergleich der beiden Modelle. Dieses Ergebnis wird beim Blick auf die AUC-Kurven in Abbildung 8.6 bestätigt.

Prognosegüte im Zeitverlauf

Bislang wurden nur die Telefonumsätze der Kunden innerhalb der ersten drei Monate nach Aktivierung der Prepaid-Karte als Prädiktoren in den Modellen verwendet. Im Folgenden soll untersucht werden, wie sich die Prognosegüte verbessern lässt, wenn zusätzlich die Umsätze von weiteren Monaten betrachtet werden. Dazu werden die Cox-Regression, das LCA- und das LCGA-Survivalmodell neu und mit allen Informationen bis zu den Monaten 4 beziehungsweise 5 berechnet. In die Modelle gehen die gleichen Baseline-Variablen ein wie zuvor, nur die Information über den ATU-Wechsel und über die geworbenen Neukunden ändert sich je nach betrachtetem Zeitpunkt.
Die latenten Variablenmodelle werden wieder schrittweise angepasst. Dabei zeigt sich auch hier, dass der BIC mit zunehmender Anzahl von latenten Klassen immer weiter sinkt. Eine natürliche Grenze ist gesetzt, wenn die einzelnen Kategorien nicht mehr repräsentativ sind, weil sie kaum noch Individuen enthalten. Bei der Modellanpassung der LCA- und LCGA-Survivalmodelle wurde jeweils die höchste Anzahl von Klassen gewählt, sodass jede Kategorie noch mindestens 1.5% der Individuen enthält.
Tabelle 8.10 gibt eine Übersicht über die Ergebnisse. Die Modelle werden anhand des C-Index verglichen. Dieser Wert bezieht sich auf den gesamten betrachten Follow-Up Zeitraum. Zunächst ist zu sehen, dass mit zunehmender Anzahl von Indikatoren in den latenten Variablenmodellen immer mehr latente Klassen, charakterisiert durch individuelle Umsatzverläufe, geschätzt werden. Die Diskriminationsfähigkeit des Cox-Modells ist zu jedem Zeitpunkt deutlich schlechter als die der LCA- und LCGA-Survivalmodelle. Mit der Information der ersten 5 Monate ist beim Cox-PH-Modell die Wahrscheinlichkeit, dass ein Kunde der früher deaktiviert wird, einen höheren Marker aufweist, 81.7%.
Unter Verwendung der Kundeninformationen aus den ersten drei Monaten stimmen die beiden latenten Variablenmodelle überein und damit auch der zugehörige C-Index. Der Vorhersagegüte unter Verwendung der Informationen bis zu Monat 4 und 5 ist bei dem LCGA-Modell etwas besser. Das LCA-Survivalmodell ist mit der Information der ersten 5 Monate in der Lage, einen Kunden mit einer Wahrscheinlichkeit von 84.3% richtig zu klassifizieren. Das LCGA-Survivalmodell kommt hier auf einen Wert von 84.4%.
Die Überlegenheit der latenten Variablenmodelle zur einfachen Cox-Regression ist deutlich zu sehen. Obwohl die Survivalanalyse mit latenten Klassen die longitudinale Datenstruktur der Umsatzmessungen ignoriert, ist die prädiktive Genauigkeit des Modells aber nahezu genauso gut wie die des LCGA-Survivalmodells.

Monat	Cox-Regression	LCA-Survival	Klassen	LCGA-Survival	Klassen
3	0.762	0.785	9	0.785	9
4	0.795	0.815	10	0.826	10
5	0.817	0.843	11	0.844	11

Tabelle 8.10: Vergleich der Modelle anhand des C-Index

Insgesamt überzeugen die Ergebnisse der Modelle als Vorhersageinstrumente für die Deaktivierungswahrscheinlichkeit eines Kunden. Zu welchem Zeitpunkt die Prognose gemacht werden sollte, ist eine Entscheidung, die der Mobilfunkanbieter anhand einer Kosten-Nutzen-Abwägung treffen muss. Die Erwägung ist abhängig von den Ausgaben für Personen, die in die eine oder andere Richtung falsch klassifiziert werden.

Kapitel 9

Zusammenfassung

In dieser Dissertation sollten die Anwendungsmöglichkeiten der Ereigniszeitanalyse mit latenten Variablen betrachtet werden. Von besonderem Interesse war dabei, welche Fragestellungen sich mit dieser Modellklasse beantworten lassen. Erörtert werden sollten weiterhin die Vor- und Nachteile der neuen Modelle im Vergleich zur Standard-Lösung mit der Cox-PH-Regression. Von besonderer Bedeutung war dabei die Prognosegüte. Außerdem stellte sich die Frage nach der Reliabilität der neuen Analysemethode.

Am Beginn der Arbeit steht eine Übersicht über die Literatur zur Ereigniszeitanalyse mit latenten Variablen. Die Artikel stammen aus den letzten Jahren, der älteste ist von 2002. Sie stellen Modelle vor, mit denen verschiedene gemessene Items als latente Faktoren oder Klassen in einem Cox-PH-Modell berücksichtigt werden können. Kommerzielle Software, um die entsprechenden Analysen zu berechnen, ist jedoch erst seit Kurzem vorhanden. Bislang wurden die Modelle daher wenig untersucht und selten angewendet. Diese Lücke konnte zu einem Teil in der vorliegenden Arbeit geschlossen werden. Bei der Anwendung der Modelle und deren Evaluation konnten zusätzliche Erkenntnisse über ihre Charakteristika gewonnen werden. Diese Ergebnisse können nachfolgenden Anwendern der Modellklasse den Weg ebenen.

Nach diesem Überblick werden die Grundlagen der Überlebenszeitanalyse formal spezifiziert, insbesondere wird auf das Cox-PH-Modell eingegangen. Es folgt eine Einführung in die Strukturgleichungsmodelle. Mit diesen Modellen können ähnliche, im Allgemeinen hoch korrelierte Items zu latenten Faktoren zusammengefasst werden. Es können Zusammenhänge zwischen den gemessenen, aber auch zwischen den unbeobachteten Variablen untersucht werden. An dieser Stelle werden verschiedene Modellfitmaße vorgestellt, mithilfe derer eine Entscheidung für das optimale latente Variablenmodell getroffen werden kann.

Nachdem die Grundlagen gelegt sind, wird die Cox-Regression mit einem latenten Faktor nach der Formulierung von Muthen erläutert und auf einen Datensatz angepasst. In dieser Arbeit werden zwei sehr unterschiedliche Anwendungsbeispiele betrachtet. Daran lässt sich erkennen, wie vielfältig die Bereiche sind, in denen sich die vorgestellten Methoden verwenden lassen.
Der erste Datensatz stammt aus einer kardiologischen Klinik. Mehrere Parameter, die unterschiedliche Aspekte der Herzfrequenzvariabilität (HRV) messen, werden zusammengefasst. Es

wird das Risiko der Patienten betrachtet, an einem koronaren Tod zu versterben.
Mit dem neuen Modell kann die Struktur der Herzfrequenzvariabilität untersucht werden. Es zeigt sich, dass diese Größe am besten durch zwei kontinuierliche latente Variablen beschrieben werden kann. Die Cox-Regression mit den beiden latenten Faktoren bietet die Möglichkeit, die Beziehungen zwischen Kovariaten, latenten Variablen und der Ereigniszeit simultan zu untersuchen. Das Modell hat den Vorteil, dass die komplexen Zusammenhänge zwischen den Parametern angemessen abgebildet werden können. Außerdem ist es damit möglich, zwischen direkten und indirekten Effekten auf die Ereigniszeit zu unterscheiden. Wie bei Strukturgleichungen üblich, können Annahmen mithilfe von Modellfitmaßen überprüft werden. So kann ein Beitrag zur Theoriebildung geleistet werden. Auf der Gegenseite stehen die höhere Modellkomplexität und eine Reihe von zusätzlichen Modellierungsannahmen.
Schließlich wird die Reliabilität des Survivalmodells mit den zwei latenten Faktoren mit dem Bootstrapping-Verfahren untersucht. Die Unsicherheit bei der Schätzung der meisten Koeffizienten des Modells ist relativ gering. Nur die Parameterschätzer der latenten Faktoren zeigen eine etwas größere Streubreite. Diese Schwankung kommt dadurch zustande, dass die zwei latenten Faktoren fast durch das gleiche Set von Parametern definiert sind. Es gibt mehr als eine Möglichkeit, wie sich das Gesamtrisiko der Herzfrequenzvariabilität auf die beiden Variablen aufteilen kann.

Latente Variablen können nicht nur kontinuierliches, sondern auch kategorielles Skalenniveau besitzen. Zunächst wird in Kapitel 6 eine Einführung in die latente Strukturanalyse gegeben. Hier können ähnliche Modellgütemaße wie in Kapitel 4 verwendet werden, um eine Entscheidung für das optimale Modell zu treffen. Anschließend wird das formale Modell der Cox-Regression mit latenten Klassen vorgestellt und auf den kardiologischen Datensatz angewendet. Die Herzfrequenzvariabilität wird am besten durch drei latente Klassen erklärt, die durch ihre mittleren Itemausprägungen charakterisiert sind. Der Einfluss dieser Kategorien auf die Überlebenszeit wird mithilfe von klassenspezifischen Achsenabschnitten in der Cox-Regression modelliert. Gleichzeitig wird der Einfluss einer Kovariate auf die Überlebenszeit für jede Klasse separat geschätzt. Das Modell bietet, im Vergleich zur einfachen Cox-Regression, ähnliche Vor-und Nachteile wie die Survivalanalyse mit einem latenten Faktor. Mit dem Modell wird ein anderer Blickwinkel auf die Fragestellung geworfen. Es können zusätzliche Erkenntnisse über die Zusammenhänge zwischen der Herzfrequenzvariabilität und dem Ereignis Koronartod gewonnen werden. Die Evaluation der Modellstabilität mit dem Bootstrap-Ansatz zeigt, dass die Lösung reliabel ist. Die mittleren Itemausprägungen in den latenten Klassen unterliegen keinen großen Schwankungen. Die Parameterschätzer des angepassten Modells unterscheiden sich nur gering von den Mittelwerten der Bootstrap-Replikationen.

Das zweite Beispiel betrachtet eine Fragestellung aus der Ökonomie. Es geht darum vorherzusagen, welche Kunden eine besonders hohe Wahrscheinlichkeit für die Deaktivierung der Prepaid-Karte ihres Mobiltelefons haben. Dazu werden, neben einer Reihe von Baseline-Parametern, die Telefonumsätze der Kunden in den ersten Monaten nach Vertragsabschluss als Prädiktoren in den Analysen verwendet.
Es werden drei Modelle zur Beantwortung der Fragestellung betrachtet und auf eine Stichprobe des Datensatzes angepasst. Zunächst wird das Standard-Cox-Modell berechnet. Alternativ kann der Einfluss der Umsätze mithilfe einer latenten Variablen in der Cox-Regression berücksichtigt werden. Als Messmodell wird, wie in Kapitel 7, zum einen eine latente Klassenregression verwendet. Da es sich bei den Items um longitudinale Messungen der gleichen Variablen

handelt, wird für das Messmodell zum anderen eine Latente-Klassen-Growth-Analyse benutzt. Die Modelle werden an einer zweiten Stichprobe aus dem Datensatz evaluiert. Es wird der prognostische Wert mithilfe des Concordance-Index und der Net Reklassification Improvement evaluiert. Die Prognosegüte der berechneten Survivalmodelle mit latenten Klassen ist insgesamt recht gut. Mit der Information eines Kunden drei Monate nach Vertragsabschluss beträgt die Wahrscheinlichkeit 78%, dass die Vorhersage mit dem Outcome Deaktivierung übereinstimmt. Dieser Wert kann unter Verwendung von mehr latenten Klassen und mehr Zeitpunkten bei dem vorliegenden Datensatz auf bis zu 84% gesteigert werden. Mit der NRI kann gezeigt werden, dass die Prognosen der Survivalmodelle mit latenten Klassen, bei der Verwendung von ausreichend Kategorien, signifikant besser sind als die der Standard-Cox-Regression.

Die Überlegenheit der Modelle konnte gezeigt werden, diese sind reliabel und in der Lage, einen tieferen Einblick in komplexe Zusammenhänge zwischen Kovariaten in der Ereigniszeitanalyse zu geben. Diese Vorzüge gibt es nicht umsonst, das Modell ist deutlich komplexer. Dennoch spricht alles für die vermehrte Verwendung dieser Modellklassen. Die Berechnung ist mit Mplus relativ einfach und es wurde gezeigt, wie vielfältig die Anwendungsmöglichkeiten sind.

Anhang A

Survivalanalyse mit einem latenten Faktor in Mplus

Es folgt die Syntax für die Analysen aus Kapitel 5. Alle Modelle in dieser Arbeit werden mit dem Programm Mplus, Version 6.11 berechnet. Anschließend werden einige Erläuterungen zu den wichtigsten Befehlen gegeben. Weitere Erklärungen zur Syntax und dem Programm sind im Mplus-User´s Guide [MM10] zu finden.

```
TITLE       Explorative Faktorenanalyse 1-3 F
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml ;
ANALYSIS:   Estimator is Mlr;
            Type = EFA 1 3;
```

```
TITLE       Faktorenanalyse 1 F
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml ;
ANALYSIS:   Estimator is Mlr;
MODEL:      f1 by Znmeannn - Zislml* ;
            f1@1;
```

```
TITLE       Faktorenanalyse 2 F
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml
ANALYSIS:   Estimator is Mlr;
MODEL:      f1 by Znmeannn - Zinslml *;
            f2 by Znmeannn - Zslml2* ;
            f1-f2@1;
            f1 with f2@0;
```

```
TITLE       Faktorenanalyse 2 F mit einer Kovariate
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=futage tod Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml zlves;
ANALYSIS:   Estimator is Mlr;
MODEL:      f1 by Znmeannn - Zinslml*;
            f2 by Znmeannn - Zslml2* ;
            f1-f2@1;
            f1 with f2@0;
            f2 on zlves;

TITLE:      Cox-Regression
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=time tod Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml zlves;
            SURVIVAL = time (All);
            TIMECENSORED = tod (0 = NOT 1 = RIGHT);
ANALYSIS:   Algorithm = Integration;
MODEL:      time on Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml Zlves;

TITLE:      Survivalanalyse mit zwei latenten Faktoren
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=time tod Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml Zlves;
            SURVIVAL = time (All);
            TIMECENSORED = tod (0 = NOT 1 = RIGHT);
ANALYSIS:   Algorithm = Integration;
            Bazehazard = Off;
MODEL:      time on f1 f2 zlves;
            f1 by Znmeannn - Zinslml*;
            f2 by Znmeannn - Zslml2* ;
            f1 with f2@0;
            f2 on zlves;
```

Erläuterungen:

- **Data:** Es werden die Rohdaten eingelesen. Alternativ kann die Korrelationsmatrix oder die Kovarianzmatrix verwendet werden.
- **Variable:** Survival: Identifiziert die Ereigniszeitvariable. Timecensored: Identifiziert den Zensierungsindikator.
- **Analysis:** Algorithm= Integration: ML-Schätzer, Berechnung mit EM-Algorithmus. Der Befehl "Bazehazard" gibt an, ob eine parametrische ("ON") oder eine nichtparametrische ("OFF") Baseline-Hazard-Funktion benutzt wird.
- **Modell:** Der Befehl "BY" definiert die Messung eines latenten Faktors durch manifeste

Indikatorvariablen. Der Befehl “ON” beschreibt eine Regression. Der Typ der Regression hängt vom Skalenniveau der Variablen ab. Standardmäßig wird für metrische Variablen eine lineare Regression berechnet, für kategorielle eine Probit bzw. Logistische Regression (bei ML-Schätzung). Für den Zusammenhang zwischen einer Ereigniszeitvariable und einer Kovariate wird eine Loglineare Regression berechnet.
Der Befehl “WITH” adressiert eine Korrelation. Das Zeichen “@” fixiert einen Parameter auf einen bestimmten Wert. Mit dem Zeichen “*” kann ein Parameter von einer Voreinstellung befreit werden.

Anhang B

Survivalanalyse mit latenten Klassen in Mplus

Es folgt die Syntax für die Analysen aus Kapitel 7.

```
TITLE       Latente-Klassenanalyse
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml;
            CLASSES = c(3);
ANALYSIS:   Type = mixture;
            Algorithm = Integration;
            Starts = 50 10;
MODEL:      %overall%
PLOT:       type is plot3;
            series = Znmeannn(1) Zrhf(2) Zlfn(3) Zef(4) Zrnspsdnn(5) Znhrvidx(6) Zrspsdnn(7)
            Zrlf5(8) Znislml(9) Zislml(10);
OUTPUT:     Tech11 Tech14;
```

```
TITLE       Latente-Klassenanalyse mit einer Kovariate
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml;
            CLASSES = c(3);
ANALYSIS:   Type = mixture;
            Algorithm = Integration;
            Starts = 50 10;
MODEL:      %overall%
            c#1-c#2 on zlves;
OUTPUT:     Tech11 Tech14;
```

```
TITLE       Survivalanalyse mit latenten Klassen
DATA:       File is "mm_mplus_ful_stand.csv";
VARIABLE:   NAMES=Znmeannn Zrhf Zlfn Zef Zrnspsdnn Znhrvidx Zrspsdnn
            Zrlf5 Zinslml Zislml;
            SURVIVAL = time (All);
            TIMECENSORED = tod (0 = NOT 1 = RIGHT);
            CLASSES = c(3);
ANALYSIS:   Type = mixture;
            Algorithm = Integration;
            Bazehazard = Off(equal);
            Starts = 50 10;
MODEL:      %overall%
            time on zlves;
            %c#1%
            time on zlves;
            [time@0];
            %c#2%
            time on zlves;
            [time*0];
            %c#3%
            time on zlves;
            [time*0];
PLOT:       type is plot3;
            series = Znmeannn(1) Zrhf(2) Zlfn(3) Zef(4) Zrnspsdnn(5) Znhrvidx(6) Zrspsdnn(7)
            Zrlf5(8) Zinslml(9) Zislml(10);
OUTPUT:     Tech11 Tech14;
```

Erläuterungen:

- **Variable:** Der Befehl "Classes(n)" spezifiziert die Anzahl der Kategorien der latenten Variable.
- **Analysis:** Mit "Type=mixture" wird eine latente Klassenanalyse angefordert. Das Programm verwendet zur Definition der latenten Klassen alle Variablen, die nicht als Kovariaten spezifiziert sind.
  Starts: Anzahl der Startwertesets im ersten und zweiten Schritt der Optimierung.
  Basehazard=off(equal): Eine nicht parametrische Baseline-Hazard-Funktion wird geschätzt, die sich zwischen den einzelnen Klassen nur um einen konstanten Faktor unterscheidet. Ab Version 6.1 ist das die Voreinstellung. Um eine vollkommen unrestringierte Baseline-Hazard-Funktionen für die einzelnen Klassen zu schätzen, muss "off(unequal)" spezifiziert werden.
- **Model:** Bei der Modellspezifikation kann bei einer Mixture-Analyse zwischen Berechnungen unterschieden werden, die über alle Klassen hinweg oder nur in einzelnen Klassen gemacht werden sollen.
- **Output:** Tech 11: Vuong-Lo-Mendell-Rubin-LR-Test und Lo-Mendell-Rubin-Adjusted-LR-Test, zum Vergleich der Lösung mit k vs. k-1 Klassen.

Anhang C

Latente-Klassen-Growth-Survivalanalyse in Mplus

Es folgt die Syntax für die Analysen aus Kapitel 8.

```
TITLE       Cox-Regression
DATA:       File is “G_Modell”;
VARIABLE:   NAMES=lum1 lum2 lum3 zum1 zum2 zum3 Time Event
            Age Email Newsl Besth Tarif Atu_3 Neuk_3;
            CATEGORICAL=Email Newsl Besth Tarif Atu_3 Neuk_3;
            SURVIVAL = Time (All);
            TIMECENSORED = Event (1 = NOT 0 = RIGHT);
ANALYSIS:   Algorithm = Integration;
MODEL:      Time on lum1 lum2 lum3 zum1 zum2 zum3
            Age Email Newsl Besth Tarif Atu_3 Neuk_3;
```

```
TITLE       Latente-Klassen-Growth-Analyse
DATA:       File is "G_Modell.csv";
VARIABLE:   NAMES=lum1 lum2 lum3;
            CLASSES = c(3);
ANALYSIS:   Type = mixture;
            Algorithm = Integration;
            Starts = 50 10;
MODEL:      %overall%
            i s q |lum1@0 lum2@1 lum3@2;
            i-q@0;
PLOT:       type is plot3;
            Series=lum1-lum3 (s)
OUTPUT:     Tech1 Tech8 Tech10;
```

```
TITLE       Latente-Klassen-Growth-Survivalanalyse
DATA:       File is "G_Modell.csv";
VARIABLE:   NAMES=lum1 lum2 lum3 Time Event
            Age Email Newsl Besth Tarif_3 Atu_3 Neuk_3;
            CATEGORICAL=Email Newsl Besth Tarif Atu_3 Neuk_3;
            CLASSES = c(4);
            SURVIVAL = Time (All);
            TIMECENSORED = Event (1 = NOT 0 = RIGHT);
ANALYSIS:   Type = mixture;
            Algorithm = Integration;
            Starts = 1000 100;
            Process = 4(starts);
MODEL:      %overall%
            Time on Age Email Newsl Besth Tarif_3 Atu_3 Neuk_3;
            c on Age Email Newsl Besth Tarif_3 Atu_3 Neuk_3;
            i s q |lum1@0 lum2@1 lum3@3;
            i-q@0;
            %c#1%
            [Time@0];
            [i] ;
            [s] ;
            [q];
            %c#2%
            [Time*0];
            [i] ;
            [s] ;
            [q];
            ⋮
PLOT:       type is plot3;
            Series=lum1-lum3 (s)
OUTPUT:     SValues;
            basehazard;
```

Erläuterungen:

- **Variable:** Categorical: Spezifiziert die kategoriellen Variablen.

- **Model:** Das Zeichen "|" wird dazu verwendet, Growth-Modelle zu spezifizieren. Die Faktorladungen sind hier passend zu den äquidistanten Zeitabständen zwischen den Messungen festgesetzt. Da der Score für den ersten Zeitpunkt auf 0 gesetzt ist, wird der Intercept-Growth-Faktor als anfänglicher Status-Faktor definiert.

- Mit dem Befehl "i-q@0" werden die Varianzen der Growth-Faktoren innerhalb einer Klasse auf 0 gesetzt. Wird diese Zeile entfernt, so werden die Varianzen der Growth-Parameter über alle Klassen hinweg gleich geschätzt. Die Varianzen können auch klassenspezifisch geschätzt werden.

- **Plot**: Es wird die graphische Darstellung der beobachteten Daten und Ergebnisse der Analysen angefordert.
- **Output:** Der Befehl “SValues” gibt die Modell-Syntax mit den Parameterschätzern der Analyse als Startwerte aus. Mit dem Befehl “basehazard” werden der klassenspezifische Baseline-Hazard und die geschätzte Baseline-Survivalwahrscheinlichkeit für jeden Ereigniszeitpunkt angegeben.

C.1 Ein Anwendungsbeispiel: Nutzungsdauer von Prepaid-Karten: Explorative Datenanalyse

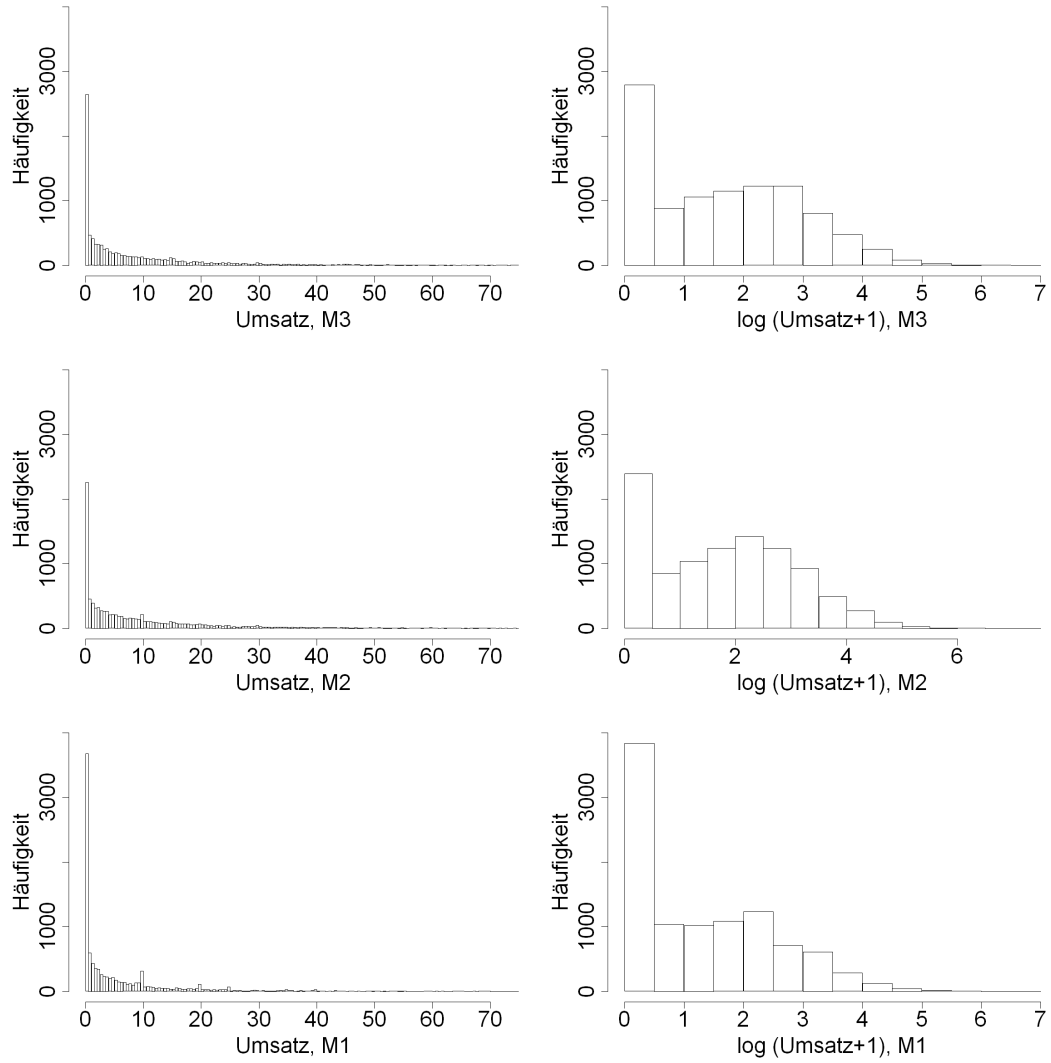

Abbildung C.1: Verteilungen der Umsätze für die Monate 1 bis 3

Der linke Teil von Abbildung C.1 zeigt die Verteilung der Umsätze für die Monate 1 bis 3. In der Graphik sind aus Platzgründen nur die unteren 95% der Werte aufgeführt. Die rechte Seite der Abbildung zeigt die Verteilung der Umsätze nach der logarithmischen Transformation.

C.2 Ein Anwendungsbeispiel: Nutzungsdauer von Prepaid-Karten: 9-Klassen-LCGA-Modell

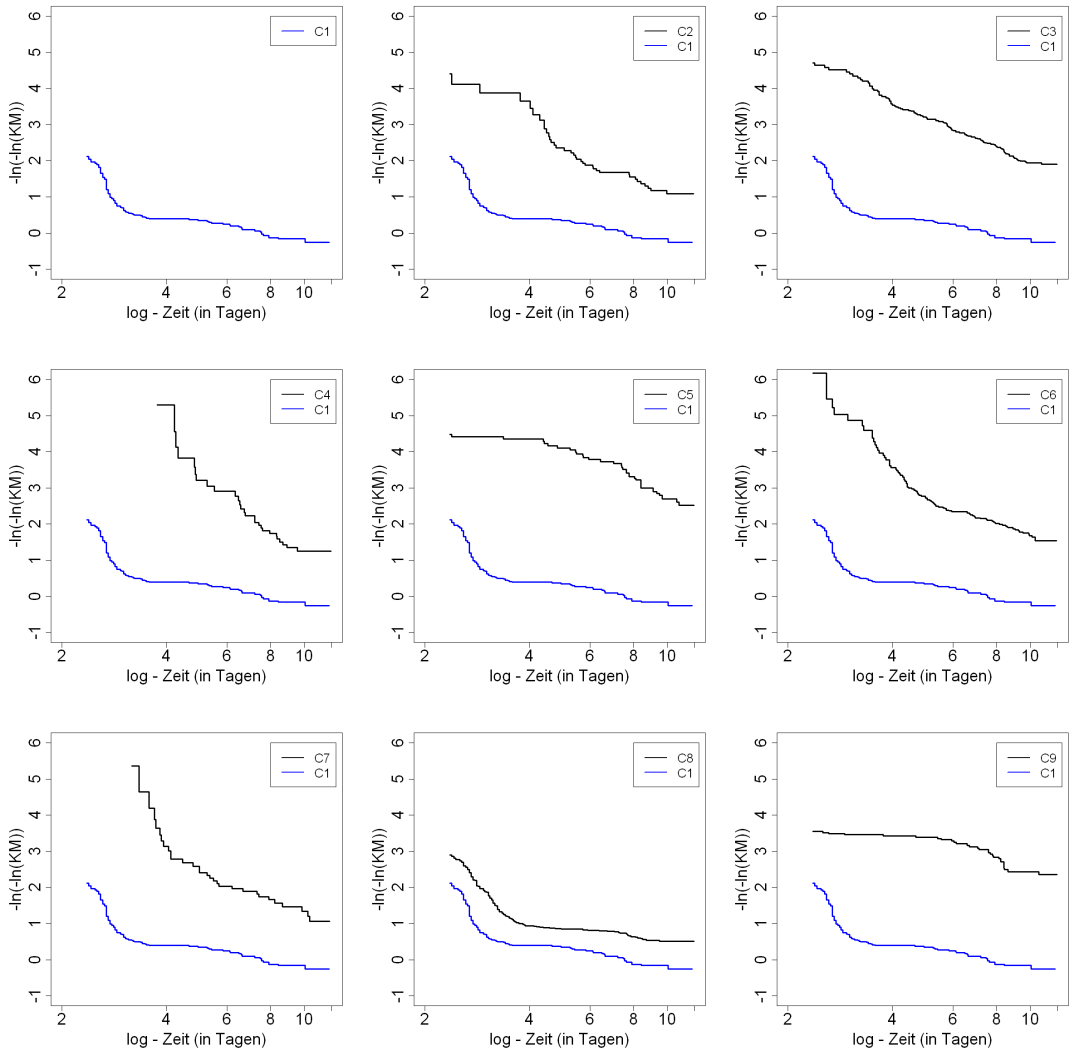

Abbildung C.2: -Log-Log-Plots für das LCGA-Survialmodell mit 9 Klassen

Klasse		Parameter	Schätzer	SE	p-Wert	OR/HR
C1	Means	Intercept	2.78	0.04	<0.001	
C1	Means	Slope	-3.66	0.08	<0.001	
C1	Means	Quadratic	1.14	0.03	<0.001	
C2	Means	Intercept	0.55	0.07	<0.001	
C2	Means	Slope	-1.74	0.12	<0.001	
C2	Means	Quadratic	1.43	0.06	<0.001	
C3	Means	Intercept	1.25	0.07	<0.001	
C3	Means	Slope	1.47	0.06	<0.001	
C3	Means	Quadratic	-0.50	0.02	<0.001	
C4	Means	Intercept	3.37	0.10	<0.001	
C4	Means	Slope	1.20	0.09	<0.001	
C4	Means	Quadratic	-0.41	0.04	<0.001	
C5	Means	Intercept	0.76	0.03	<0.001	
C5	Means	Slope	0.69	0.07	<0.001	
C5	Means	Quadratic	-0.22	0.03	<0.001	
C6	Means	Intercept	2.52	0.06	<0.001	
C6	Means	Slope	0.77	0.06	<0.001	
C6	Means	Quadratic	-0.26	0.02	<0.001	
C7	Means	Intercept	0.36	0.03	<0.001	
C7	Means	Slope	4.23	0.14	<0.001	
C7	Means	Quadratic	-1.33	0.05	<0.001	
C8	Means	Intercept	1.71	0.05	<0.001	
C8	Means	Slope	2.69	0.09	<0.001	
C8	Means	Quadratic	-1.72	0.04	<0.001	
C9	Means	Intercept	0.30	0.02	<0.001	
C9	Means	Slope	-0.00	0.03	0.948	
C9	Means	Quadratic	-0.01	0.02	0.508	
C1	On	Alter	-0.05	0.00	<0.001	0.95
C1	On	Email	-0.51	0.26	0.047	0.60
C1	On	Bestellherkunft	1.08	0.24	<0.001	2.94
C1	On	Tarif	0.82	0.15	<0.001	2.27
C2	On	Alter	-0.04	0.01	<0.001	0.96
C2	On	E-Mail	-0.89	0.26	<0.001	0.41
C2	On	Tarif	1.05	0.15	<0.001	2.86
C3	On	Alter	-0.03	0.00	<0.001	0.97
C3	On	Tarif	-0.93	0.14	<0.001	0.39
C3	On	Neukunden M3	0.76	0.14	<0.001	2.14
C4	On	Alter	-0.05	0.00	<0.001	0.95
C4	On	E-Mail	-0.76	0.28	0.007	0.47
C4	On	Tarif	0.96	0.19	<0.001	2.61
C4	On	ATU Wechsel M3	0.94	0.26	<0.001	2.56
C4	On	Neukunden M3	0.88	0.26	0.001	2.41
C5	On	Tarif	-2.58	0.48	<0.001	0.08
C5	On	ATU Wechsel M3	-0.38	0.17	0.023	0.68
C5	On	Neukunden M3	0.51	0.17	0.003	1.67

Klasse		Parameter	Schätzer	SE	p-Wert	OR/HR
C6	On	Alter	-0.06	0.00	<0.001	0.94
C6	On	E-Mail	-0.51	0.17	0.003	0.60
C6	On	Bestellherkunft	0.44	0.13	<0.001	1.55
C6	On	ATU Wechsel M3	0.54	0.19	0.005	1.72
C6	On	Neukunden M3	0.96	0.17	<0.001	2.61
C7	On	Alter	-0.05	0.01	<0.001	0.95
C7	On	E-Mail	-0.71	0.29	0.016	0.49
C7	On	Bestellherkunft	-0.55	0.24	0.020	0.58
C7	On	Tarif	1.12	0.19	<0.001	1.21
C7	On	ATU Wechsel M3	0.77	0.24	0.001	2.16
C8	On	Alter	-0.04	0.00	<0.001	0.96
C8	On	E-Mail	-0.52	0.19	0.006	0.59
C8	On	Bestellherkunft	0.75	0.16	<0.001	2.12
C8	On	Tarif	0.58	0.12	<0.001	1.79
C1	Intercepts	Time	0.00	0.00	999.00	999.00
C2	Intercepts	Time	-1.90	0.23	<0.001	0.15
C3	Intercepts	Time	-2.37	0.20	<0.001	0.09
C4	Intercepts	Time	-2.24	0.24	<0.001	0.11
C5	Intercepts	Time	-3.28	0.26	<0.001	0.04
C6	Intercepts	Time	-2.35	0.20	<0.001	0.10
C7	Intercepts	Time	-2.09	0.26	<0.001	0.12
C8	Intercepts	Time	-0.85	0.18	<0.001	0.43
C9	Intercepts	Time	-2.71	0.21	<0.001	0.07
C1-C9	Time on	Alter	-0.01	0.00	0.111	0.99
C1-C9	Time on	E-Mail	-0.68	0.15	<0.001	0.51
C1-C9	Time on	Newsletter	0.52	0.14	<0.001	1.68
C1-C9	Time on	Bestellherkunft	0.36	0.12	0.002	1.43
C1-C9	Time on	Tarif	0.39	0.11	0.001	1.12
C1-C9	Time on	ATU Wechsel M3	1.24	0.13	<0.001	3.46
C1-C9	Time on	Neukunden M3	-1.25	0.26	<0.001	0.29

Tabelle C.1: Ergebnis des LCGA-Survivalmodells mit 9 Klassen

Symbolverzeichnis

Symbol	Bedeutung
α	Parameterschätzer in multinomialer Logit-Regression
α'	Parameterschätzer in Latent-Response-Variablen-Regression
$\alpha_0, \alpha_1, \alpha_2$	Mittelwerte der Growth-Faktoren
B	$(m \times m)$ Koeffizientenmatrix für die Beziehungen zwischen den η
$\beta = (\beta_1 \ldots, \beta_q)$	Parameterschätzer für die Kovariaten in der-Cox Regression
γ	Parameterschätzer in multinomialer Logit-Regression
γ'	Parameterschätzer in Latent-Response-Variablen-Regression
c	Kategorielle latente Variable, mit $k = 1, \ldots, K$ Kategorien
C_i	Zensierungszeit von Individuum i
D_i	Zensierungsindikator von Individuum i
δ	$(q \times 1)$ Vektor der Residuen, Kovarianzmatrix Θ_δ
ϵ	$(p \times 1)$ Vektor der Residuen, Kovarianzmatrix Θ_ϵ
Γ	$(m \times n)$ Koeffizientenmatrix für den Einfluss von ξ auf η
$h(t)$	Hazard-Funktion zur Zeit t
$h_0(t)$	Baseline-Hazard-Funktion zur Zeit t
H_t	Kumulierte Hazard-Funktion zur Zeit t
$H_0(t)$	Kumulierte Baseline-Hazard-Funktion zur Zeit t
H	$(m \times q)$ Koeffizientenmatrix für den Einfluss von x auf η
$\eta = (\eta_1, \ldots, \eta_m)$	Vektor latenter abhängiger Variablen, Kovarianzmatrix Φ
η_0, η_1, η_2	Intercept-, Slope- und Quadratischer-Growth-Faktor
Λ_y	$(p \times m)$ Matrix von Faktorladungen
Λ_x	$(q \times n)$ Matrix von Faktorladungen
ι_k	Klassenspezifischer Achsenabschnitt in der Cox-Regression
$f(t)$	Dichte der Verteilung der Ereigniszeiten U
$F(t)$	Verteilung der Ereigniszeiten U
ν	Parameterschätzer in logistischer Regression
μ	Mittelwert einer Latent-Response-Variablen u^*
$\hat{p}_{ik}$	Geschätzte Wahrscheinlichkeit von Individuum i der Klasse k anzugehören
π_v	Wahrscheinlichkeit für eine Kategorie v der abhängigen Variable u_j
$R(t)$	Risiko Set zum Zeitpunkt t
$\hat{se}^*$	Bootstrap-Standardfehler
S	Stichprobenmatrix
$S(t)$	Survival-Funktion zur Zeit t
Σ	Populationsmatrix
T_i	Ereigniszeit von Individuum i

Symbol	Bedeutung
τ	Schwellenwerte der Latent-Response-Variablen u^*
$\hat{\theta}$	Beobachteter Schätzer
$\hat{\theta^*}$	Bootstrap-Schätzer
U_i	Überlebenszeit von Individuum i
$u = (u_1, \ldots, u_r)$	Vektor manifester Indikatorvariablen $j = 1, \ldots, r$
$u^* = (u_1^*, \ldots, u_r^*)$	Vektor der Latent-Response-Variablen $j = 1, \ldots, r$
$x = (x_1 \ldots, x_q)$	Vektor unabhängiger Variablen, Kovarianzmatrix Σ_{xx}
$\xi = (\xi_1, \ldots, \xi_n)$	Vektor latenter unabhängiger Variablen, Kovarianzmatrix Φ
$y = (y_1, \ldots, y_p)$	Vektor abhängiger Variablen, Kovarianzmatrix Σ_{yy}
$y^* = (y_1^*, \ldots, y_p^*)$	Vektor der Latent-Response-Variablen
z	Fehlerterm in Latent-Response-Variablen-Regression, mit Varianz ς^2
ζ	$(m \times 1)$ Vektor der Residuen, Kovarianzmatrix Ψ
$\zeta_0, \zeta_1, \zeta_2$	Residuen der Growth-Faktoren, mit Kovarianzmatrix Ψ

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Aka87] H. Akaike. Factoranalysis and aic. *Psychometrica*, 52:317–332, 1987.

[AMM06] T. Asparouhov, K. Masyn, and B. Muthen. Continuous time survival in latent variable models. In *Proceedings of the Joint Statistical Meeting in Seattle*, 2006.

[Büh11] Markus Bühner. *Einführung in die Test- und Fragebogenkonstruktion.* Pearson Studium, 2011.

[Bir68] A. Birnbaum. *Statistical Theories of Mental Test Scores*, chapter Some latent trait models and their use in inferring an examinees ability, pages 392–479. Reading, Massachusetts: Addison-Wesley, 1968.

[Bre74] N. E. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

[Cox72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[Efr77] B. Efron. The efficiency of cox´s likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

[ESC96] ESC. Task force of the european society of cardiology and the north american society of pacing and electrophysiology. heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation*, 93:1043–65., 1996.

[ET93] B. Efron and R.J. Tibshirani. *An Introduction to the bootstrap.* Chapman and Hall, CRC, 1993.

[Gei10] C. Geiser. *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung.* VS Verlag für Sozialwissenschaften, 2010.

[Goo74] L.A. Goodman. The analysis of systems of qualitative variables when some of the variables are unobservable. part i: A modified latent structure approach. *American Journal of Sociology*, 79:1179–1259, 1974.

[HJ12] F.E. Harrell Jr. *Package 'Hmisc'.* f.harrell@vanderbilt.edu, 3.9-3 edition, March 29, 2012.

[HZ05] P. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105, 2005.

[Joh83] S. Johansen. An extension of cox´s regression model. *International Statistical Review*, 51:165–174, 1983.

[Jör02] K.G. Jöreskog. Structural equation modeling with ordinal variables using lisrel, 2002.

[JS89] K. G. Jöreskog and D. Sörbom. *LISREL 7: A guide to the program and application. Chicago: Spss Inc.* 1989.

[JW08] T. Jung and K.A.S. Wickrama. An introduction to latent class growth analysis and growth mixture modeling. *Social and Personaility Psychology Compass*, 2/1:302–317, 2008.

[KK05] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text.* Springer, New York, second edition, 2005.

[KM03] J. Klein and M. Moeschberger. *Survival Analysis: Techniques for censored and truncated data.* Springer, New York, 2003.

[KP80] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data.* Wiley, 1980.

[Lar04] K. Larsen. Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, 60(1):85–92, 2004.

[Lar05] K. Larsen. The cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics*, 61(4):1049–1055, 2005.

[LH68] P.F. Lazarsfeld and N.W. Henry. *Latent Structure Analysis.* Boston: Houghton Mill., 1968.

[Lon83] J.S. Long. Covariance structure models - an introduction to lisrel. Sage University Paper Series: on Quantitative Applications in the Social Siences Beverly Hills: Sage Publications, 1983.

[LTMS02] H. Lin, B.W. Turnbull, C.E. McCulloch, and E.H. Slate. Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65, 2002.

[MA02] B. Muthen and T. Asparahouv. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. Mplus Web Note 4, 2002.

[MA06] B. Muthen and T. Asparahouv. Continuous-time survival analysis in mplus. Technical report, 2006.

[MA08] B. Muthen and T. Asparahouv. chapter Growth mixture modeling: Analysis with non-Gaussian random effects. 2008.

[MAB+09] B. Muthen, T. Asparouhov, T.M. Boye, M. Hackshaw, and A. Naegeli. Applications of continuous-time survival in latent variable models for the analysis of oncology randomized clinical trial data using mplus. 2009.

[Mas03] K.E. Masyn. *Discrete-Time Survival Mixture Analysis for Single and Recurrent Events Using Latent Variables*. PhD thesis, University of California, Los Angeles, 2003.

[MLST02] C.E. McCulloch, H. Lin, E.H. Slate, and B.W. Turnbull. Discovering subpopulation structure with latent class models. *Statistics in Medicine*, 21:417–429, 2002.

[MM10] L. Muthen and B. Muthen. Mplus user´s guide, version 6. Technical report, Los Angeles, CA: Muthen & Muthen, 2010.

[MP00] G.J. McLachlan and D. Peel. *Finite Mixture models*. Wiley, 2000.

[Mut83] B. Muthen. Latent variable structural equation modeling. *Journal of Econometrics*, 22:43–65, 1983.

[Mut01] B. Muthen. *New Developments and Techniques in Structural Equation Modeling*, chapter Latent Variable Mixture Modeling, pages 1–33. Lawrence Erlbaum Associates, 2001.

[Mut04a] B. Muthen. *Handbook of quantitative methodology for the social sciences*, chapter Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data., pages 345–368. Newbury Park, CA: Sage Publications., 2004.

[Mut04b] B. Muthen. Mplus technical appendices. Technical report, Los Angeles, CA: Muthen & Muthen, 2004.

[Mut10] B. Muthen. Specifying latent class indicators in mixture regression. Mplus Discussion: http://www.statmodel.com/discussion/messages/13/5853.html, 2010.

[Nag05] D.S. Nagin. Group-based modeling of development. *Boston, MA: Harvard University Press.*, 2005.

[NAM07] K. L. Nylund, T. Asparouhov, and B. Muthen. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, 14(4):535–569, 2007.

[Pea08] M.J. Pencina et al. Evaluation the added predicitve ability of a new marker: From area under the roc curve to reclassification and beyond. *Statis*, 27:157–172, 2008.

[PLT09] C. Proust-Lima and J.M.G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment psa: a joint modeling approach. *Biostatistics*, 10:535–549, 2009.

[Rei05] J. Reinecke. *Strukturgleichungsmodelle in den Sozialwissenschaften*. Oldenbourg Verlag, München, Wien, 2005.

[SB01] A. Satorra and P.B. Bentler. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66:507–514, 2001.

[Sch78] G. Schwartz. Estimation the dimension in a model. *The Annal of Statistics*, 6:461–464, 1978.

[SL04] R.E. Schumacker and R.G. Lomax. *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates, second edition edition, 2004.

[SNL07] P. Stein and M.-A. Nehrkorn-Ludwig. Einfueuhrung in die analyse mit kovarianz-strukturmodellen, 2007.

[Ueb12] J. Uebersax. http://www.john-uebersax.com/stat/, 01.05.2012.

[VM03] J. Vermunt and J. Magidson. *Encyclopedia of Social Science Research Methods*, chapter Latent Class Analysis. Sage Publications, 2003.

[Wol70] J.H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–50, 1970.

Hiermit erkläre ich, Lena Herich, an Eides statt, dass ich die Dissertation mit dem Titel:

“Erweiterung des Cox-Proportional-Hazards-Modells um latente Faktoren und latente Klassen”

selbständig und ohne fremde Hilfe verfasst habe.

Andere als die von mir angegebenen Quellen und Hilfsmittel habe ich nicht benutzt. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Hamburg, 16.08.2012

Lena Herich