Michael Schomaker

Selektieren und Kombinieren von Modellen unter Berücksichtigung der Problematik fehlender Daten

Cuvillier Verlag

Selektieren und Kombinieren von Modellen unter Berücksichtigung der Problematik fehlender Daten

Michael Schomaker

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

München, im Dezember 2009

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Aufl. - Göttingen: Cuvillier, 2010
Zugl.: München (Univ.), Diss., 2009

978-3-86955-330-6

Berichterstatter:	PD Dr. Christian Heumann
Auswärtiger Gutachter:	Prof. Dr. Susanne Rässler
Rigorosum:	23. Februar 2010

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2010
Gedruckt auf säurefreiem Papier

978-3-86955-330-6

Do not fear to be eccentric in opinion,
for every opinion now accepted was once eccentric.

Bertrand Russel

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen Menschen bedanken, die in unterschiedlicher Form zum Gelingen dieser Arbeit beigetragen haben. Mein besonderer Dank gilt Christian Heumann für die Betreuung der Dissertation, viele hilfreiche Vorschläge und Denkanstöße, seine Zeit und sein Interesse sowie Gespräche, die auch über den Tellerrand der Statistik hinausgingen. Helge Toutenburg danke ich für seine große Unterstützung während der letzten Jahre, eine immer produktive Zusammenarbeit und vor allem natürlich für seine humorvolle Art und eine abwechslungsreiche Zeit. Bei Alan Wan bedanke ich mich für sein außerordentliches Engagement, viel konstruktive und fachkundige Kritik und dass er mir einen in jeder Hinsicht lohnenden Aufenthalt in Hongkong ermöglicht hat. Mein Dank geht auch an Shalabh für viele wertvolle Hinweise in puncto konsequenten, effektiven wissenschaftlichen Arbeitens, viele unübertroffene Lebensweisheiten und eine äußerst herzliche Zusammenarbeit. Susanne Rässler danke ich für ihre Zeit und ihr Interesse und der Übernahme des auswärtigen Gutachtens.

Bedanken möchte ich mich auch bei Ingrid Kreuzmair für die schöne und ereignisreiche Zeit in und um Raum 340, Michael Obermeier für ausführliches Korrekturlesen und sachdienliche Hinweise zu einigen Absurditäten deutscher Rechtschreibung, Birgit Schrödle für ein detailliertes inhaltliches Feedback und wertvolle Anmerkungen zur Modellmittelung, Benjamin Hofner für viele Vorschläge zur Präsentation meiner Inhalte, Stefan Pilz für hilfreiche Anmerkungen und Korrekturen zu dieser Arbeit, Jan Ulbricht für angeregte Diskussionen zu Akaike & Co. und Gero Walter und Andrea Wiencierz für Gespräche rund um die Statistik und das tägliche Dolce Vita.

Großer Dank gilt auch meinen Eltern, die mich in meinen Plänen immer unterstützt und ermutigt haben. Auch hätte ich ohne die zahlreichen Nachmittage in Iffezheim wohl nie Statistik studiert...

Zusammenfassung

In den letzten Jahren haben sich Modellmittelungsverfahren als Alternative zur Modellselektion etabliert. Anstatt sich auf ein einziges Siegermodell zu beschränken, werden hierbei mehrere konkurrierende Modelle betrachtet und ihre Parameterschätzer gewichtet miteinander kombiniert. Das Hauptaugenmerk liegt dabei meist auf der Konstruktion der Gewichte, wie auch der Optimalität der daraus resultierenden gewichteten Parameterschätzung. In der vorliegenden Arbeit werden verschiedene Konzepte bayesianischer, vor allem aber auch frequentistischer Modellmittelung (Frequentist Model Averaging, FMA) erläutert und ihre Stärken und Schwächen gegenüber einer Vielzahl an traditionellen Modellselektionsmethoden herausgestellt.
Schwerpunkt ist dabei die Konstruktion und Diskussion verschiedener Strategien zur Verwendung von FMA-Methoden unter Berücksichtigung der Problematik fehlender Daten. Hierfür werden zwei Kernkonzepte vorgeschlagen: Der erste Ansatz konstruiert Gewichte für einen FMA-Schätzer auf Basis eines für fehlende Daten adjustierten Kriteriums, welches der aktuellen Literatur aus dem Bereich der Modellselektion entstammt und das das im Kontext fehlender Werte bekannte Prinzip des inverse probability weighting verwendet; der zweite Ansatz ersetzt die fehlenden Werte durch einfache und multiple Imputationen, um darauf aufbauend geeignete Punktschätzungen und deren Varianz mit Hilfe bekannter Modellmittelungsschätzer zu konstruieren. Zu diesem Zweck wird auch ein rekursiver Imputationsalgorithmus präsentiert, der die geläufige Idee einer Regressionsimputation unter Verwendung generalisierter additiver Modelle verallgemeinert.
Die Arbeit zeigt die Eigenheiten, Stärken und Schwächen der vorgestellten Ansätze im Kontext von linearen und logistischen Regressionsanalysen anhand weitreichender Monte-Carlo-Simulationen auf und diskutiert am Beispiel der Faktorenanalyse mögliche Erweiterungen und Verallgemeinerungen der angeführten Schätzer für weitere multivariate, statistische Analysemethoden. Alle Verfahren werden an realen Datensätzen illustriert.
Es zeigt sich, dass in vielen Situationen beide vorgestellten Konzepte einem Verwerfen der nicht-vollständigen Beobachtungen vorzuziehen sind, die Strategie einer Modellmittelung nach Imputation in der Regel bessere Resultate erzielt als die Verwendung eines FMA-Schätzers, der Gewichte auf Basis eines für fehlende Daten adjustierten Kriteriums verwendet, und insbesondere die technisch weniger aufwändigen Modellmittelungsverfahren zu besseren Schätzungen führen als diejenigen, die aus einer klassischen Modellselektion resultieren.

Abstract

Model averaging or combining is often considered as an alternative to model selection. Rather than attaching to a single „winning“ model, a model average estimator weights across the estimators of many potential models. The construction of appropriate weights and the properties of the resulting estimator are issues of high concern.
In this work, Frequentist Model Averaging (FMA) is considered extensively and strategies for the application of FMA methods in the presence of missing data based on two distinct approaches are presented and compared to concepts of traditional model selection: The first approach combines estimates from a set of appropriate models which are weighted by scores of a missing data adjusted criterion developed in the recent literature of model selection. The second approach averages over the estimates of a set of models with weights based on conventional approaches but with the missing data replaced by imputed values prior to estimating the models. For this purpose several imputation methods that have been programmed in currently available statistical software are considered, and a recursive algorithm is proposed to implement a extended version of regression imputation that relies on generalized additive models.
Focusing on the linear and binary logistic regression model, the properties of the FMA estimators resulting from these strategies are explored by means of Monte-Carlo studies and extensions of the presented methodologies to other areas of multivariate statistical modeling are discussed briefly in the context of factor analysis by way of example. As an illustration, the proposed methods are applied to real data.
The results show that in many situations both approaches are superior to a Complete Case Analysis, averaging after imputation is normally preferred to averaging using weights that adjust for the missing data and model average estimators, especially those of technically easy conception, often provide better estimates than those resulting from any single model.

Inhaltsverzeichnis

1. Einleitung

Die Konstruktion, die Wahl und das Verständnis von Modellen sind von zentraler Bedeutung innerhalb vieler wissenschaftlicher Erkenntnisprozesse: Modelle helfen, die wesentlichen Strukturen wahrnehmbarer und nicht wahrnehmbarer Gegenstandsbereiche aufzudecken und damit Phänomene darzustellen und zu erklären; um es mit den Worten von Frigg und Hartmann (2006) zu formulieren:

> *„Models are vehicles for learning about the world“*

Um Phänomene erfassen und damit auch deren relevante Effekte beschreiben und verstehen zu können, werden innerhalb der Statistik eine Vielzahl an multivariaten Verfahren wie etwa parametrische und nichtparametrische Regressionsmodelle, autoregressive Prozesse, Bild-, Kontur-, Faktor- oder auch Clusteranalysen verwendet. Die konkrete Entscheidung zugunsten einer adäquaten Modellierung, also beispielsweise die Wahl geeigneter Kovariablen bei Regressionsmodellen oder die Bestimmung der Anzahl der Faktoren in der Faktorenanalyse, ist ein klassisches Modellselektionsproblem, für das zahlreiche Methoden unterschiedlicher Konzeption existieren. Typischerweise wird für die interessierende/n Größe/n eine im Kontext passende Verteilungsfamilie gewählt und es werden für die Wahl der zur Modellierung geeigneten Variablen sukzessive Hypothesentests, Shrinkageschätzungen, risikobasierte Entscheidungen auf Basis von Modellwahlkriterien, statistische Lernverfahren oder ad-hoc Vermutungen herangezogen. Einige aus diesem weitreichenden Feld repräsentative Ansätze werden im Verlauf der Arbeit noch näher diskutiert; eine ausführliche Übersicht findet sich auch bei Rao und Wu (2001).

Modellmittelung

Ungeachtet der Schwierigkeit, ein passendes Verfahren für eine konkrete Fragestellung zu wählen, wird insbesondere seit Mitte der 1990er Jahre die Problematik der Modellselektionsunsicherheit in der Literatur diskutiert: Die Eigenschaften von Punktschätzungen statistischer Modelle, wie auch deren Varianz, hängen sowohl davon ab, auf welche Art und Weise das Modell gewählt wurde als auch von dessen stochastischen Begebenheiten.

In der Regel wird aber nur der letztgenannte Punkt beachtet und jede Form von Inferenz wird so durchgeführt, als wäre das betrachtete Modell *a priori* gewählt worden und die Schätzungen unabhängig von der Modellwahl. Tatsächlich ist dies nicht korrekt, da der Selektionsschritt datenbasiert ist und die Unsicherheit bezüglich der Modellselektion in den entsprechenden Schätzungen reflektiert werden sollte. Von Interesse sind damit also nicht die Eigenschaften der Schätzer bedingt auf *ein* gewähltes Modell, sondern die unbedingten Parameterschätzungen, die den durch die Modellselektion hervorgerufenen Inferenzschritt berücksichtigen, vergleiche auch Leeb und Pötscher (2003, 2005, 2006a, 2008b). Wird dieser Sachverhalt dennoch ignoriert, können als Konsequenz Parameterschätzungen verzerrt sein und die Varianz wird ob der nicht berücksichtigten Unsicherheit systematisch unterschätzt, was zu durchweg überoptimistischen Konfidenzintervallen führt.

Inzwischen ist weitgehend akzeptiert, dass als eine Möglichkeit, diesem Problem entgegenzutreten, Parameterschätzungen mehrerer Modelle kombiniert werden können. Kerngedanke ist dabei, dass nicht nur ein, sondern verschiedene Modellierungsansätze eine gute Beschreibung der Datenstruktur bieten können und eine gewichtete Mittelung vieler plausibler Schätzungen die Modellselektionsunsicherheit adäquat erfassen kann, also die Berechnung eines sinnvollen Schätzers *post model selection* ermöglicht. Dieser Gedanke wird in der Literatur erstmals sauber von Leamer (1978) formuliert, der einen bayesianischen Ansatz von *model averaging*, also Modellmittelung, propagiert: Hierbei werden für alle infrage kommenden Modelle posteriori-Wahrscheinlichkeiten berechnet und die entsprechenden Parameterschätzungen so kombiniert, dass die Schätzungen der Modelle mit höherer posteriori-Wahrscheinlicheit auch ein insgesamt höheres Gewicht erhalten. Die zum damaligen Zeitpunkt noch sehr limitierten computergestützten Ressourcen ermöglichten jedoch keine konkrete Umsetzung der vorgeschlagenen Konzepte, so dass sich nur vereinzelte Veröffentlichungen in den darauffolgenden Jahren der Thematik annahmen. Erst die wegweisenden Artikel von Draper (1995) und Chatfield (1995) und eine mittlerweile deutlich erhöhte Computerpower führten zu neuen Denkanstößen und einer Vielzahl an kohärenten Ideen und Konzepten bayesianischer Modellmittelung innerhalb kürzester Zeit. Ein detaillierter Überblick findet sich unter anderem bei Hoeting et al. (1999).

Eine korrekte Umsetzung der Methodik verlangt jedoch weiterhin einen hohen computationalen Aufwand; insbesondere deswegen und auch aufgrund einiger strittiger philosophischer Positionen – so etwa die Frage was priori- und posteriori-Wahrscheinlichkeiten von Modellen überhaupt bedeuten – haben sich über die letzten 15 Jahre einige nicht-

bayesianische, in gewisser Weise frequentistische, Alternativen ergeben, die insgesamt einfacher und schneller umzusetzen sind, bisher jedoch nicht wirklich einem einheitlichen Konzept entsprechen. Aufgrund ihres großen, jedoch noch nicht vollständig erschlossenen Potentials bilden sie auch den Schwerpunkt der vorliegenden Arbeit. Einige Kerngedanken zur frequentistischen Methodik finden sich unter anderem bei Buckland, Burnham und Anderson (1997), Hjort und Claeskens (2003), Yuan und Yang (2005), Hansen (2007) und Schomaker, Wan und Heumann (2010).

Grundsätzlich stellen sich ungeachtet des zugrundeliegenden Paradigmas die Fragen: Welche Modelle sollen kombiniert werden? Wie sollen diese Modelle kombiniert werden? Was bedeutet dies für die kombinierten Parameterschätzungen? Prinzipiell lässt sich festhalten, dass die Punkt- und deren Varianzschätzungen für alle Modelle von Interesse gewichtet miteinander kombiniert werden sollten, wobei die Gewichte der einzelnen Modelle in der Regel so bestimmt werden, dass entweder über ein Selektionskriterium bzw. die posteriori-Wahrscheinlichkeit jedem Modell eine gewisse Plausibilität zugesprochen wird oder der finale, gemittelte Schätzer auf irgendeine Art und Weise optimal ist. Exemplarisch für diese beiden, grundsätzlich unterschiedlichen Vorgehensweisen werden in dieser Arbeit insbesondere dem Schätzer von Buckland, Burnham und Anderson (1997), der zur Gewichtung der Kandidatenmodelle exponentielle AIC-Gewichte verwendet sowie dem Schätzer von Hansen (2007), der in gewissem Sinne einen optimalen Modellmittelungsschätzer garantiert, besondere Beachtung geschenkt.

Fehlende Daten

Die vorliegende Arbeit beschäftigt sich mit Modellselektion und Modellmittelung unter Berücksichtigung der Problematik fehlender Daten. In der Praxis sind fehlende Werte ein häufig auftretendes Problem, eine weitere Komponente der Unsicherheit, die verschiedenste Ursachen besitzen kann: etwa wenn in Meinungsumfragen die befragten Personen Antworten verweigern, bei klinischen Studien Patienten nicht über die volle Zeitspanne beobachtet werden können oder in naturwissenschaftlichen Experimenten Ergebnisse aufgrund fehlerhafter Messungen verworfen werden müssen. Um diese Problematik im Rahmen statistischer Inferenz explizit zu berücksichtigen, werden typischerweise zwei konzeptionell unterschiedliche Ansätze verfolgt:

- Zum einen können die fehlenden Werte durch andere, „plausible“ Werte ersetzt werden; so etwa indem konkrete Ausprägungen ähnlicher Beobachtungen impu-

tiert werden (Chen und Shao (2000)) oder die Abhängigkeitsstruktur der Variablen über ein passendes Regressionsmodell innerhalb der vollständigen Fälle erfasst wird und dessen Vorhersagen als Imputationen verwendet werden (Little und Rubin (2002)). Bei einem solchen oder ähnlichen Vorgehen wird jedoch vernachlässigt, dass die neuen, imputierten Werte nur datenbasierte Schätzungen für die wahren, fehlenden Daten sind und somit eine Unsicherheit bezüglich der Imputation vorliegt, die nicht vernachlässigt werden sollte. Dieser Sachverhalt kann durch die Verwendung multipler Imputationen (Rubin (1978), Rubin (1996)) berücksichtigt werden: Hierbei wird jeder fehlende Wert einer Datenmatrix durch $M > 1$ zufällig gezogene Werte aus der prädiktiven a-posteriori-Verteilung der fehlenden Daten gegeben die beobachteten Daten (oder einer Approximation davon) ersetzt, wodurch M neue Datensätze zur Verfügung stehen, deren Schätzungen kombiniert werden können, um so die Unsicherheitskomponente bezüglich der Imputation zu erfassen.

- Zum anderen kann der Umstand fehlender Daten auch direkt im Inferenzprozess mitberücksichtigt werden, etwa über den EM-Algorithmus, bei dem die Likelihood des interessierenden Parameters gegeben die beobachteten Daten approximiert wird (Dempster, Laird und Rubin (1977)) oder auch unter Verwendung von Gewichtungsansätzen, bei denen typischerweise die Inferenz auf den gewichteten, vollständigen Fällen durchgeführt wird, um konsistente Schätzungen zu ermöglichen (vergleiche auch Molenberghs und Kenward (2007, Kapitel 10)).

Es stellt sich die Frage, wie die Konzepte von Modellselektion, Modellmittelung und fehlenden Daten zusammenpassen und miteinander kombiniert werden können.

Übergreifende Ansätze

Derzeit existieren im Wesentlichen vier wissenschaftliche Arbeiten, die die Auswirkungen fehlender Werte auf die Modellselektion untersuchen, nämlich die Artikel von Shimodaira (1994), Cavanaugh und Shumway (1998), Hens, Aerts und Molenberghs (2006) und Claeskens und Consentino (2008). Alle angeführten Arbeiten unterstreichen, dass die Verwendung gängiger Verfahren der Modellselektion für den Fall fehlender Daten zur Wahl unpassender Modelle führen kann und schlagen allesamt eine Modifizierung von Akaikes Informationskriterium (AIC, Akaike (1973)) vor. Der in dieser Arbeit zentrale Artikel von Hens, Aerts und Molenberghs (2006) greift dabei die in vielen Teilgebieten

der Statistik bekannte und bereits angedeutete Idee des *inverse probability weighting* (Horvitz und Thompson (1952)) auf, bei der nur die vollständigen Beobachtungen eines Datensatzes verwendet werden – gewichtet mit ihrer inversen geschätzten Auswahlwahrscheinlichkeit. Die Autoren schlagen darauf aufbauend ein adjustiertes Selektionskriterium (AIC_W) vor und konstatieren diesem in einer Auswahl an Simulationsstudien ein durchweg gutes Verhalten im Kontext fehlender Werte.

Es existiert bis dato jedoch keine einzige Veröffentlichung, die die Modellselektionsunsicherheit und die Problematik fehlender Werte gemeinsam diskutiert. Es stellt sich die Frage, ob und wie sich die Konzepte von Modellmittelung und fehlenden Daten in Einklang bringen lassen und inwiefern sich eine solche Kombination auch praktisch umsetzen lässt. Die vorliegende Arbeit zeigt die Vielfältigkeit beider Themengebiete auf und versucht sowohl für unterschiedlichste Ansätze aus dem Bereich fehlender Werte als auch aus dem Bereich der Modellmittelung einige repräsentative Ideen aufzugreifen und entsprechend zu erweitern.

Hierfür werden exemplarisch für ein mögliches Vorgehen drei konträre Ansätze erläutert: Es werden verschiedene Modellselektions- und Modellmittelungsschätzer unterschiedlicher Konzeption betrachtet, so etwa die oben angeführten Schätzer von Buckland, Burnham und Anderson (1997) bzw. Hansen (2007), wobei für deren Berechnung

(i) die fehlenden Werte verworfen und nur die vollständigen Fälle eines Datensatzes verwendet werden,

(ii) die fehlenden Werte durch einfache oder multiple Imputationen ersetzt werden und die Inferenz auf den entsprechend aufgefüllten Datensätzen durchgeführt wird und

(iii) das für fehlende Werte adjustierte Kriterium von Hens, Aerts und Molenberghs (2006) zur Konstruktion von Gewichten für einen neuen, kriteriumsbasierten Modellmittelungsschätzer verwendet wird.

Für den zweiten der genannten Punkte werden dabei drei unterschiedliche Imputationsmethoden betrachtet. Zum einen die k-Nächste-Nachbarn-Methode (Chen und Shao (2000), Gottardo (2008)), bei der jeder fehlende Wert durch das arithmetische Mittel seiner k nächsten, vollständig beobachteten, Nachbarn ersetzt wird; ferner wird die Idee der Regressionsimputation (Little und Rubin (2002)) aufgegriffen, bei der die fehlenden Werte durch Vorhersagen einer linearen Regression ersetzt werden. Dieses Konzept wird

verallgemeinert, indem zur Vorhersage flexiblere, generalisierte additive Modelle verwendet und in einen rekursiven Algorithmus integriert werden. Darüber hinaus werden einfache und multiple Imputationen über das Paket „Amelia II“ (Honaker, King und Blackwell (2008)) der statistischen Software *R* generiert, das einen schnellen und robusten Bootstrap-basierten Ansatz zur Modellierung der prädiktiven a-posteriori-Verteilung der fehlenden Daten gegeben die beobachteten Daten verwendet.

Es zeigt sich im Verlauf der Arbeit, dass das Verhalten der vorgestellten Modellselektions- und Modellmittelungsstrategien höchst unterschiedlich ist und sich für die entsprechenden Schätzer abhängig von der konkreten Behandlung fehlender Werte interessante und vielfältige Eigenschaften nachweisen lassen.

Die Ansatzpunkte und Konzepte im Bereich der Modellwahl und Modellmittelung sind außerordentlich vielfältig: Je nach Zweck, Ziel und Fragestellung ergeben sich unterschiedliche Ausgangspositionen, die zu Methoden, Kriterien und Verfahren führen, die einerseits unterschiedlicher nicht sein könnten, andererseits am Ende dann doch zu ähnlichen oder sogar identischen Resultaten führen können. Die vorliegende Arbeit versucht, im Gegensatz zu vielen anderen Veröffentlichungen, die Kernkonzepte zur Wahl und Kombination von Modellen relativ weitläufig zu diskutieren, um so die Chancen und Risiken für Modellselektions- und Modellmittelungsschätzer im Kontext fehlender Daten unter den verschiedensten Gesichtspunkten für viele Themenbereiche zu öffnen. Dabei sollen ausgewählte Verfahren und Ideen in ein einheitliches Spektrum eingeordnet und ihre Eigenheiten, Stärken und Schwächen herausgestellt werden.

Ziel der Arbeit ist es, interessante und repräsentative Verfahren der Modellselektion und Modellkombination in den Kontext fehlender Werte zu übertragen und die Eigenschaften und Erfolgsaussichten der vorgestellten Ideen mit Hilfe weitreichender Überlegungen, Monte-Carlo-Simulationen und Anwendungsbeispielen zu diskutieren, bewerten, analysieren und illustrieren.

Ausblick

Kapitel 2 versucht zunächst zu klären, was ein Modell überhaupt ist, was man darunter generell in der Wissenschaft verstehen kann, inwiefern sich dies im Rahmen statistischer Modellierung wiederfindet, welche Ziele mit Modellselektion verfolgt werden und auf welche Art und Weise die vielfältigen Konzepte zu rechtfertigen sind. Einen Überblick ausgewählter Modellwahlverfahren, insbesondere aus dem Anwendungsbereich linearer und

generalisierter linearer Regressionsmodelle, werden in Kapitel 3 vorgestellt. Wie Modelle kombiniert werden können und in welche Richtung aktuelle Entwicklungen führen, wird in Kapitel 4 diskutiert. Der Schwerpunkt liegt dabei in Anbetracht der darauffolgenden Analysen vor allem auf nicht-bayesianischen Konzepten. Kapitel 5 vermittelt grundlegende Ideen aus dem Bereich fehlender Daten und versucht repräsentative Methoden frequentistischer Modellselektion und Modellmittelung in diesem Kontext zu erweitern, ergänzen und verbessern. Die vorgeschlagenen Schätzer werden anschließend mit Hilfe ausführlicher Simulationsstudien (Kapitel 6) und mehrerer Anwendungsbeispiele (Kapitel 7) bewertet und illustriert. Schwerpunkt sind dabei das lineare und das logistische Regressionsmodell mit fehlenden Werten in den Kovariablen sowie mit Einschränkungen auch ausgewählte Aspekte der Faktorenanalyse. Kapitel 8 fasst die wichtigsten Resultate noch einmal zusammen und diskutiert die Chancen, Erfolgsaussichten und Probleme der betrachteten Verfahren unter Beachtung aller gewonnen Erkenntnisse. Ein Ausblick auf mögliche Erweiterungen der vorgestellten Strategien wird ebenfalls gegeben.

2. Modelle in Wissenschaft und Statistik

Das Ziel der Wissenschaft wird in der allgemeinen Wissenschaftstheorie mit Erkenntnisgewinn, dem Erwerb von neuem Wissen[1], gleichgesetzt. Insbesondere will die Wissenschaft nicht nur Tatsachen feststellen, sondern auch Ursachen und Erklärungen für diese finden. Die Suche allgemeiner Strukturen und Beziehungen beschränkt sich dabei nicht nur auf wahrnehmbare, sondern auch auf nicht-wahrnehmbare Gegenstandsbereiche. Die Systematisierung dieser Bereiche, also die Reduzierung ihrer Vielfältigkeit auf einige wenige elementare Faktoren, ist dabei von größtem Interesse. Diese Form der Abstraktion wird in der Wissenschaft unter anderem durch die Konstruktion von *Modellen* erreicht.

Seien G_1 und G_2 Gegenstände[2], S_1, S_2 Sätze[3] und bedeute $M(A, B)$, dass A Modell für B ist; dann lassen sich folgende Formen von Modellen unterscheiden (vgl. auch Detel (2007, Seite 94)):

(i)	$M(G_1, G_2)$	G_1 ist ein *strukturelles Modell* für G_2 ,
(ii)	$M(S_1, G_1)$	S_1 ist ein *abstraktes* bzw. *idealisiertes Modell* für G_1 ,
(iii)	$M(G_1, S_1)$	G_1 ist ein *semantisches Modell* für S_1 ,
(iv)	$M(S_1, S_2)$	S_1 ist ein *theoretisches Modell* für S_2 .

Im Falle struktureller Modelle sind also Gegenstände Modelle anderer Gegenstände; so zum Beispiel das maßstabsgetreue Modell einer Brücke oder das Doppelhelix-Modell der DNA. Bei abstrakten und idealisierten Modellen werden hingegen Sätze als Modelle für Gegenstände verwendet. Sie versuchen die Komplexität von Phänomenen einzuschränken, nur ihre Kernelemente zu erfassen und sie unter dem Ziel der Verständlichkeit

[1] Die traditionelle Definition nach Plato bezeichnet „Wissen“ als *wahre, gerechtfertigte Meinung*. Auch wenn diese Definition nicht unumstritten ist (vgl. Gettier (1963)), so ist sie im Kontext von Modellbildung und Modellselektion völlig ausreichend.

[2] In der Philosophie bezeichnet ein Gegenstand eine Sache oder eine Entität, die Eigenschaften besitzen und Beziehungen zu anderen Gegenständen haben kann.

[3] Ein Satz ist eine im Sinne der Logik widerspruchsfreie Aussage (z.B. über einen oder mehrere Gegenstände), die mittels eines Beweises, das heißt aus Axiomen und bereits vorhandenen Sätzen, hergeleitet werden kann.

vereinfachend darzustellen. Damit bilden sie das Rückgrat aller empirischen Wissenschaften. Ein Beispiel dafür ist etwa das Modell idealer Gase, mit dem sich unter gewöhnlichen Bedingungen viele thermodynamische Prozesse von Gasen verstehen und beschreiben lassen, das für tiefe Temperaturen und hohen Druck jedoch keine adäquate Modellierung mehr bietet; auch die vereinfachende Annahme unabhängiger und identisch verteilter Beobachtungen einer Stichprobe, um Verfahren der statistischen Inferenz besser verwenden zu können, kann als idealisiertes Modell verstanden werden; oder das Gesetz von Henry Darcy, das die Wasserströmung in porösen Flüssigkeiten modelliert und ursprünglich durch Versuche in einem Sandbett entstanden ist, heutzutage jedoch vor allem als spezielle Lösung der Navier-Stokes-Gleichungen motiviert werden kann. Bei semantischen Modellen sind es Gegenstände, die Sätze wahr machen; beispielsweise die rationalen Zahlen $\mathbb{Q}$ ohne Null, mit der Multiplikation als Verknüpfung und der Eins als neutralem Element, für die abelschen Gruppen im Bereich der Gruppentheorie der Mathematik. Sind Sätze Modelle für andere Sätze, so spricht man von theoretischen Modellen - insbesondere dann, wenn wissenschaftliche Theorien vorteilhaft für andere Theorien verwendet werden können. In diesem Sinne ist die klassische Mechanik ein theoretisches Modell für die Quantenmechanik.

Auch in der Statistik spielen Modelle eine herausragende Rolle. Das Sammeln, das Aufbereiten, die Analyse und die Interpretation von Daten eröffnet eine Fülle an Möglichkeiten statistischer Modellierung. Als weitgehend empirische Wissenschaft sind es insbesondere abstrakte und idealisierte Modelle, denen eine große Bedeutung zuzuschreiben ist. Aufgrund von Beobachtungen in der Form von Daten, sollen statistische Modelle Phänomene abstrahieren und beschreiben. Ein statistisches Modell in seiner allgemeinsten Form bezeichnet dabei eine parametrisierte Familie von Wahrscheinlichkeitsverteilungen

$$\mathcal{F} \;=\; \{f(y;\theta),\, \theta \in \Theta\}\,, \tag{2.1}$$

bei der die Beziehung einzelner Elemente eines Phänomens anhand einer Parametrisierung der Dichte von y durch θ in $\mathcal{F}$ beschrieben wird. Der Parameterraum Θ muss dabei nicht endlich-dimensional sein. Ein Beispiel ist die Menge aller Normalverteilungen

$$\mathcal{F} \;=\; \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2 \right) ; \theta = (\mu,\sigma) \in \mathbb{R}^p \times (0,\infty) \right\},$$

wobei μ den Erwartungswert und σ^2 die Varianz von y beschreibt. Häufig meint ein statistisches Modell auch die funktionale Beziehung zwischen einer Zielgröße y und potentiellen Einflussgrößen $X_1, \ldots, X_p$,

$$y = f(X_1, \ldots, X_p; \theta) + \epsilon\,, \tag{2.2}$$

wobei $f(\cdot)$ eine noch unbestimmte Funktion und ϵ eine Zufallskomponente bezeichnet. Ein typisches Beispiel ist das lineare Regressionsmodell

$$y \;=\; \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)\,,$$

bei dem die Zielgröße y über eine Linearkombination der $X_1, \ldots, X_p$ modelliert wird. In der Regel betrachtet man zur Modellierung dieser Abhängigkeitsbeziehung die bedingte, parametrisierte Dichte $f(y|X_1, \ldots, X_p; \theta)$, weswegen (2.2) als Spezialfall von (2.1) aufgefasst werden kann.

Ziel der statistischen Modellselektion ist es, aus einer Menge von Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\} \subset \mathcal{F}$ ein Modell M_{κ^*} auszuwählen, das die Daten – unter noch zu definierenden Gesichtspunkten – gut beschreibt. Dies betrifft insbesondere die Wahl geeigneter Regressoren bei Regressionsmodellen, die Anzahl von Kontrollpunkten in der Kontur- und Bildanalyse, die Ordnung autoregressiver Prozesse, die Anzahl von Faktoren in der Faktorenanalyse, die Wahl eines geeigneten Kerndichteschätzers und andere Problemstellungen der Statistik und verwandter Gebiete, vergleiche auch Linhart und Zucchini (1986) sowie Rao und Wu (2001).

Sparsamkeit

Um Phänomene beschreiben und verstehen zu können, sollte ein gewähltes Modell einerseits ein möglichst genaues Abbild der Realität liefern, andererseits nur die Komplexität in Anspruch nehmen, die nötig ist, um die wichtigsten Kausalitäten und Merkmale der Daten abzubilden. Da zudem insbesondere im statistischen Kontext die Varianz, etwa der Parameterschätzung von θ, mit der Komplexität steigt, der Bias in der Regel dagegen fällt, stellt sich die Frage nach einem geeigneten Kompromiss. Ist ein statistisches Modell zu komplex, so nennt man es überangepasst, ist es zu simpel, so nennt man es unterangepasst. In der Literatur wird häufig die Auffassung vertreten, dass bei gleicher Erklärungskraft das weniger komplexe Modell gewählt werden sollte. Dieses Prinzip ist

auch als *Prinzip der Sparsamkeit* bzw. als *Occam's Razor*[4] bekannt. Die Wahl eines solchen Modells ist jedoch keineswegs trivial, Burnham und Anderson (2002) bemerken hierzu:

> *„Parsimony lies between the evil of under- and overfitting"*

und Forster (1998) fragt:

> *„How much better must the complex model fit before we say that the extra parameter is necessary? Or, when should the better fit of complex models be 'explained away' as arising from the greater tendency of complex models to fit noise? How do we trade off fit with simplicity?"*

Um sich in der Statistik, auch unter Beachtung der Sparsamkeit, zwischen mehreren konkurrierenden Modellen für ein bestes entscheiden zu können, sind Verfahren und Kriterien notwendig. Statistische Modellselektion umfasst dabei häufig

(i) eine risikobasierte Entscheidung durch Optimierung eines Selektionskriteriums,

(ii) das sukzessive Testen von Hypothesen,

(iii) oder ein ad-hoc Vorgehen.

Punkt (i) beinhaltet insbesondere Selektionskriterien auf Basis von Vorhersagefehlern, im Rahmen der Informationstheorie und bayesianischer Natur. Ausgewählte Verfahren und Methoden zu diesen drei, wie auch einigen anderen Punkten sollen in den Abschnitten 3.1–3.5 vorgestellt und motiviert werden. Dies geschieht, wenn möglich, in allgemeinster Form; stets jedoch im Hinblick auf die Wahl geeigneter Regressoren in linearen und generalisierten linearen Regressionsmodellen.

[4] William Ockham (1285–1349) formulierte als erster ein Prinzip der Sparsamkeit (häufig wiedergegeben als „*entia non sunt multiplicanda praeter necessitatem*"), das weit über die Statistik hinaus in Biologie, Medizin und Philosophie bekannt ist. Heutzutage existieren im Detail viele verschiedene Fassungen und Versionen dieses Prinzips; prinzipiell lässt es sich jedoch so verstehen, dass für zwei wissenschaftliche Theorien bzw. Erklärungen unter festen Bedingungen diejenige zu bevorzugen ist, die einfacher ist. Die Rechtfertigung dieses Prinzips wird in der Wissenschaftstheorie diskutiert, vergleiche etwa Sober (1981) und Forster und Sober (1994). Die Begründungen sind dabei sowohl pragmatisch motiviert, wie etwa der oben angedeutete Punkt, dass Phänomene auf diese Weise besser verstanden werden können, als auch wissenschaftstheoretisch; so das Argument, dass das Ziel der Wissenschaft in der Approximation der Wahrheit besteht (vgl. auch die untenstehende Diskussion) und dieses Ziel nicht ohne die Berücksichtigung von Sparsamkeit erreicht werden kann. Kritische Stimmen sprechen dem Prinzip keine allgemeine Gültigkeit zu. So meint selbst Sober (2002): „*It may turn out, that simplicity has no global justification – that its justification varies from problem to problem*".

Die meisten der vorgestellten Methoden beinhalten dabei das Sparsamkeitsprinzip mehr oder weniger explizit. Es stellen sich in diesem Zusammenhang jedoch die grundlegende Fragen: Was ist Sparsamkeit? Wie lässt sich Sparsamkeit messen und konstruieren? Ist Sparsamkeit eindeutig? Diesen Fragen entspringt ein natürlicher Diskurs, ob Sparsamkeit Teil eines empirischen Vorgehens sein kann oder ob es ein künstliches, insofern rationalistisches, Konzept ist. Prinzipiell setzt Empirismus voraus, dass jede Form von Wissen über Erfahrung, beispielsweise über Daten, gewonnen wird, während unter einer rationalistischen Denkweise Erkenntnis in erster Linie unabhängig von unseren Sinneseindrücken entsteht. Deswegen wird in der Literatur teilweise argumentiert, dass das Sparsamkeitsprinzip ein extraempirisches Element ist, das primär in der Statistik, aus pragmatischen Gründen, herangezogen wird und somit als rationalistisch angesehen werden muss; so schreiben Forster und Sober (1994):

> *„Giving weight to simplicity thus seems to embody a kind of rationalism“* .

Die folgenden Kapitel werden zeigen, dass dieses Argument nicht korrekt ist. Obgleich Bestandteil fast jedes statistischen Verfahrens bzw. Kriteriums, kann die Interpretation der Aufsplittung in einen Anpassungs- und einen Sparsamkeitsterm als Kompromiss zwischen Bias und Varianz in der Regel nur a posteriori erfolgen; a priori liegt den Methoden meist ein grundlegend anderes Prinzip zugrunde, etwa die Approximation von Wahrheit, die Verringerung von Vorhersagefehlern oder die Maximierung von posteriori-Wahrscheinlichkeiten – die Sparsamkeitsterme entstehen dabei gewissermaßen als „Abfallprodukt“ bei der Evaluierung der eigentlichen Zielsetzung. Es zeigt sich daher, dass die Konzepte statistischer Modellselektion fast ausschließlich datengestützt und empirisch motiviert sind und keine rationalistische Rechtfertigung benötigen.

Eine Ausnahme bildet die von Jorma Rissanen begründete Theorie der Minimum Description Length. Hierbei bildet der Kompromiss zwischen Anpassung und Sparsamkeit das Fundament aller von Rissanen (1978) erarbeiteten Verfahren. Konzepte aus der Informationstheorie helfen dabei, die Länge wissenschaftlicher Theorien, im Speziellen statistischer Modelle, zu beschreiben und dadurch Modelle zu wählen, die Wissen am besten „verschlüsseln“ können und dennoch anschließend dieses Wissen am besten zu reproduzieren vermögen. Dadurch wird nicht nur das Prinzip der Sparsamkeit direkt bei der Konstruktion von Modellwahlkriterien verwendet, sondern auch erstmals eine konkrete Ausarbeitung davon präsentiert, wie Sparsamkeit zu messen ist. Eine überschaubare Einführung bietet Abschnitt 3.5.1; die dort angegebenen Referenzen erlauben darüber hinaus einen tieferen Einblick in die Thematik, die in den folgenden Kapiteln keine zentrale Rolle einnehmen wird.

Wahrheit

Ein Vergleich und eine Beurteilung der einzelnen Verfahren, insbesondere der Selektionskriterien, erfolgt meist durch die Betrachtung asymptotischer Güteeigenschaften, wie der Konsistenz und der Effizienz. Abschnitt 3.6 beschäftigt sich hiermit ausführlich. Es zeigt sich, dass eine der entscheidenden Voraussetzungen zur Optimalität eines Modellwahlkriteriums in der Tatsache liegt, ob in der Menge der Kandidatenmodelle $\mathcal{M} = \{M_1, \ldots, M_k\}$ das wahre, datengenerierende Modell M_κ^* enthalten ist oder nicht. Es zeigt sich ferner, dass das Konzept eines wahren Modells als solches problematisch ist. Obgleich Voraussetzung in der Konstruktion vieler populärer Modellwahlkriterien, ist seine Existenz nicht unumstritten. Burnham und Anderson (2002) bemerken

> *„The words 'true model' represent an oxymoron"*

und de Leuuw (1988) meint in diesem Zusammenhang lapidar

> *„Truth is elusive"*

Die Diskussion dieses Aspekts verlagert sich in der statistischen Literatur jedoch meist in Richtung der nahezu gleichwertigen Frage der Dimensionalität eines wahren Modells. Ist es von unendlicher Dimension, so ist es implizit nicht in der Menge der Kandidatenmodelle enthalten; ist es von endlicher Dimension, so kann es durchaus Bestandteil dieser Menge sein. Abschnitt 3.6 präsentiert wichtige Resultate zur Optimalität von Modellwahlkriterien und diskutiert ihre Nützlichkeit anhand ausgewählter Aspekte.

In gewisser Weise ist diese Diskussion auch Bestandteil einer alten Realismus-Debatte: Die Annahme einer denkunabhängigen Realität, einer Realität die sich erkennen und erfassen lässt und damit auch letztlich zu Wissen unabhängig von menschlichen Theorien und Konventionen führt, kann als Position für einen erkenntnistheoretischen, wissenschaftlichen Realismus verstanden werden.[5] Eine solche Sichtweise impliziert, dass eine Wirklichkeit, eine Wahrheit existiert und wir diese erfahren und beschreiben können und dass die Annäherung an diese Wahrheit insofern auch Ziel der Wissenschaft ist. Im Gegensatz dazu existieren viele nicht-realistische Positionen, die sich in philosophischen Denkweisen, wie etwa dem Relativismus, dem Instrumentalismus oder dem konstruktiven Empirismus äußern. Letzterer geht auf van Fraaasen (1980) zurück und verneint das

[5] Tatsächlich umfasst der Begriff des Realismus eine Vielzahl philosophischer Positionen, die sich auf unterschiedliche Gegenstandsbereiche beziehen. Diese sind im Kontext dieser Arbeit jedoch nicht weiter relevant; der Kerngedanke und das Stichwort einer „denkunabhängigen Realität" genügt für die folgende Diskussion.

Ziel der Wissenschaft als Approximation von Wahrheit. Die Doktrin eines konstruktiven Empirismus sieht vor, die Wissenschaft als reinen Beobachter zu betrachten, der wahre Aussagen über beobachtbare Phänomene und Experimente machen kann, ausdrücklich aber nicht über unbeobachtbare Phänomene und damit über eine den Beobachtungen zugrundeliegende Wahrheit. Es ist fragwürdig, ob solche oder andere (beispielsweise relativistische) Standpunkt hilfreich sind. Sober (2000) quittiert die Diskussion mit den Worten:

> *„Realism says that the goal of science is to discover which theories are true; [constructive] empiricism maintains that the goal is to discover theories that are empirically adequate [...] In both cases, truth is the property that matters“*

Tatsächlich ist die Realismus-Debatte im Kontext statistischer Modellselektion nicht entscheidend. Auch wenn aufgrund ihrer Konzeption viele der in Kapitel 3 vorgestellten Methoden eine zugrundeliegende Wahrheit, zumindest in Form eines datengenerierenden Prozesses, benötigen und damit auch eine prinzipiell realistische Sichtweise angenommen wird, so steht vor allem die Identifizierung relevanter Effekte eines Phänomens im Vordergrund. Ob Aussagen über eine Wahrheit und ihre Existenz getroffen werden müssen, ist fraglich. Dies macht auch Abschnitt 3.6 deutlich: Ob die oben erwähnten Qualitätsmerkmale von Konsistenz und Effizienz im Kontext statistischer Modellwahl sinnvoll sind, wird dort diskutiert.

Grenzen des Wissens

Die Suche geeigneter Modelle zur Charakterisierung von Phänomenen unterliegt häufig gewissen Beschränkungen, insbesondere solchen, die sich aus den Grenzen empirischen Wissens ergeben. Dies betrifft vor allem die mangelnde Verfügbarkeit und das Fehlen von Daten: Die Herausforderung, Modelle zu bilden, zu wählen und zu schätzen, auch unter Beachtung der oben diskutierten Sachverhalte der Sparsamkeit und Optimalität, erfährt im Kontext fehlender Werte eine zusätzliche Dimension. Die Diskussion und Evaluierung von statistischen Methoden zur Modellselektion unter Berücksichtigung dieser Problematik ist Schwerpunkt dieser Arbeit und wird weitgehend in den Kapiteln 5-7 erörtert. Zusätzlich berücksichtigt werden dabei auch die Grenzen statistischer Modellselektionsverfahren. Die Unsicherheit bezüglich der Wahl eines geeigneten Modells ist ein weiterer Faktor, dem besondere Beachtung geschenkt wird. Die Kombination verschiedener Konzepte aus verschiedenen Teilgebieten der Statistik sollen helfen, ein relativ allgemein angelegtes Sammelsurium an Methoden zur Bewältigung dieser Probleme zu liefern. Die Illustration dieser Methoden beschränkt sich aus Gründen der Übersichtlichkeit dabei auf

lineare und logistische Regressionsmodelle, wie auch exemplarisch auf die Faktorenanalyse. Wie aus der obigen Diskussion bereits zu erkennen, wird dabei stets eine zu einem gewissem Maße realistische Grundhaltung angenommen, da stets eine Annäherung an eine zugrundeliegende Wahrheit, zumindest in Form eines datengenerierenden Prozesses, vorausgesetzt wird. Alle vorgestellten Methoden, mit Ausnahme der MDL-Methodik, sind dabei rein empirisch zu motivieren und bedürfen keiner zusätzlich rationalistischen Rechtfertigung.

3. Modellselektion

Die Ausführungen aus Kapitel 2 lassen erkennen, dass der Begriff der Modellselektion sehr weit gefasst ist und sich auf eine Vielzahl an Verfahren unterschiedlichster Motivation bezieht, die allesamt die Entscheidungsfindung für eine konkrete statistische Modellierung unterstützen sollen. Im Fokus dieses Kapitels sowie der gesamten Arbeit stehen – wie bereits erwähnt – vor allem Methoden, die insbesondere für die Selektion von Variablen in linearen und generalisierten linearen Modellen, die Bestimmung von Faktoren in der Faktorenanalyse und die Wahl der Ordnung autoregressiver Prozesse konzipiert wurden; wenn möglich werden diese in allgemeinster Form erläutert. Hierfür wird im Folgenden meist, jedoch nicht immer, die $n \times p + 1$ Datenmatrix

$$D_* = \begin{pmatrix} y_1 & x_{11} & \dots & \dots & x_{1p} \\ \vdots & \vdots & * & & \vdots \\ \vdots & \vdots & & & * \\ \vdots & \vdots & & * & \vdots \\ y_n & x_{n1} & \dots & \dots & x_{np} \end{pmatrix} \tag{3.1}$$

betrachtet, wobei $y = (y_1, \dots, y_n)'$ den Response und $X_j = (x_{1j}, \dots, x_{nj})'$, $j = 1, \dots, p$, mögliche Einflussgrößen einer multivariaten Zusammenhangsanalyse bezeichnen. Der $1 \times p$ Vektor $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ enthält die Werte der Kovariablen für die i-te Beobachtung. Durch das '$*$'-Symbol wird angedeutet, dass einige $x_{ij} \in D_*$ fehlen können; dieser Sachverhalt wird in Kapitel 5 ausführlich betrachtet.

Es wird angenommen, dass die Beobachtungen $y_1, \dots, y_n$ unabhängig sind und der Dichte f entstammen. Wie in Kapitel 2 bereits ausführlich erläutert, wird im Rahmen der Modellselektion eine parametrisierte Familie von Wahrscheinlichkeitsverteilungen $\mathcal{F} = \{f(y;\theta),\, \theta \in \Theta\}$ betrachtet. Im Kontext linearer Regressionsmodelle wird die bedingte, parametrisierte Dichte $f(y|X_1, \dots, X_p;\theta)$ auch häufig als $f(y|X_1, \dots, X_p;\beta,\sigma^2)$ notiert. Sofern nicht anders angegeben ist $\beta = (\beta_0, \dots, \beta_p)'$ dabei der dem Intercept und den Regressoren zugehörige $p + 1 \times 1$ Parametervektor und σ^2 steht für die Varianz. In Teilen der Literatur erfolgt die zusätzliche Aufteilung von β in $\beta = (\alpha, \gamma)$,

wobei α die ersten d Parameter bezeichnet, die aus inhaltlichen oder anderen Gründen auf jeden Fall in das Endmodell aufgenommen werden sollten und γ die weiteren q Parameter, die potentiell in dieses Modell integriert werden könnten. Dieses Vorgehen mag dem eingangs diskutierten Prinzip der Sparsamkeit in gewisser Weise widersprechen, wird aber insbesondere zur Vergleichbarkeit mit der Originalliteratur, falls nötig, übernommen. Im Kontext generalisierter linearer Regressionsmodelle wird die bedingte, parametrisierte Dichte $f(y|X_1, \ldots, X_p; \theta)$ häufig auch in Form einer Exponentialfamilie gemäß $f(y|X_1, \ldots, X_p; \vartheta, \phi)$ beschrieben, wobei ϑ den kanonischen und ϕ den Dispersionsparameter beschreibt. Der den Kovariablen zugehörige Parametervektor wird auch dort weiterhin mit β bezeichnet.

Die Menge aller konkurrierenden Modelle ist $\mathcal{M} = \{M_1, \ldots, M_k\} \subset \mathcal{F}$; die Dimension eines Modells $M_\kappa \in \mathcal{M}$, $\dim(M_\kappa)$, entspricht der Anzahl der Elemente von θ. Im Kontext linearer Regressionsmodelle wird die Varianz häufig als nuisance-Parameter behandelt, weswegen sich die Dimension von M_κ hier meist auf $\beta \subset \theta$ bezieht.

Im Folgenden werden eine Vielzahl an Modellwahlverfahren unterschiedlichster Konzeption vorgestellt und diskutiert: Abschnitt 3.1 erläutert dabei Selektionsprozeduren, die ihre Entscheidung auf Parameterschätzungen zurückführen, so etwa das klassische, sukzessive Hypothesentesten bzw. die Verwendung von Shrinkage-Verfahren. Abschnitt 3.2 beschreibt einige ausgewählte Ansätze aus dem Bereich der Vorhersagefehler, beispielsweise das Konzept der Kreuzvalidierung; Kriterien, die informationstheoretisch motiviert werden können, zum Beispiel Akaikes Informationskriterium, werden in Abschnitt 3.3 behandelt. Abschnitt 3.4 zeigt das bayesianische Selektionsverständnis auf und Abschnitt 3.5 verweist schließlich auf viele weitere, bekannte und unbekannte Selektionsmethoden.

Relevant für das Verständnis der Modellmittelungsmethoden aus Kapitel 4 und die in Kapitel 5 angedachten zentralen Korrekturen im Kontext fehlender Daten sind dabei im Besonderen das Kriterium von Akaike (Abschnitt 3.3.1) und Mallows C_p (Abschnitt 3.2.1) sowie mit Abstrichen auch die Kreuzvalidierungskriterien aus Abschnitt 3.2.3 und das Bayesianische Selektionskriterium nach Schwarz (Abschnitt 3.4.1). Die Vielfalt der in diesem Kapitel vorgestellten Konzepte demonstriert, wo geeignete Modifikationen für spezielle Fragestellungen liegen können und bildet damit die Grundlage für mögliche Erweiterungen der an späterer Stelle eingeführten Modellmittelungsverfahren bei fehlenden Werten. Einige diesbezüglich interessante Aspekte werden im weiteren Verlauf und auch zu Ende der Arbeit noch einmal angeschnitten.

3.1 Modellselektion durch Betrachtung der Parameterschätzungen

Der möglicherweise klassischste Ansatz in der Modellselektion besteht im Kontext von einfachen Regressionsanalysen meist darin, die den Regressoren zugehörige Parameterschätzung $\hat{\beta}$ und deren Standardfehler zu betrachten und abhängig davon Variablen zu selektieren. Dies geschieht in der Regel über das sukzessive Testen von Hypothesen bzw. durch die Verwendung neuerer statistischer Schätz- und Lernverfahren, wie beispielsweise Shrinkage oder Boosting. In den folgenden beiden Abschnitten 3.1.1 und 3.1.2 werden einige ausgewählte, populäre Methoden vorgestellt und kritisch diskutiert; die bereits an dieser Stelle gewonnen Erkenntnisse und Gedanken liefern die erste wertvolle Motivation für die in dieser Arbeit zentralen Modellmittelungsverfahren aus Kapitel 4 und Abschnitt 5.2.

3.1.1 Sukzessives Testen von Hypothesen

Gegeben sei eine Menge an Regressionsmodellen $\mathcal{M} = \{M_1, \ldots, M_k\}$, wobei X_κ die Designmatrix und β_κ den Parametervektor eines Modells $M_\kappa \in \mathcal{M}$ bezeichnet. Die Kandidatenmodelle seien bezüglich ihrer Dimension geordnet, also $\dim(M_1) \leq \ldots \leq \dim(M_k)$; dann testen Selektionsprozeduren auf Basis statistischer Hypothesentests sukzessive Hypothesen der Form $\beta_{\tilde{k}} = 0$, wobei $\beta_{\tilde{k}}$ einen Subvektor des Parametervektors β_k bezeichnet, um den Einfluss der Regressoren $X_1, \ldots, X_p$ auf y und damit die geeignete Dimension eines „besten“ Modells M_{κ^*} zu bestimmen.

Modellselektion auf Basis statistischer Hypothesentests benötigt neben Aufnahme- und Ausschlusskriterien zur Wahl geeigneter Regressoren auch einen Algorithmus (im folgenden als Selektionsprozedur bezeichnet), dessen Konvergenz mit einem passenden Stoppkriterium sichergestellt werden soll. Populäre und in den gängigen Softwarepaketen implementierte Prozeduren sollen im folgenden vorgestellt und diskutiert werden:

(a) **Selektionsprozeduren**. Einige der am häufigsten verwendeten Selektionsprozeduren sind die *Vorwärtsselektion*, die *Rückwärtsselektion* sowie die *Schrittweise Selektion*, die dem Algorithmus von Efroymson (1960) entspricht. Sie lassen sich wie folgt beschreiben:

Vorwärtsselektion

1. Beginne mit dem Modell kleinster Dimension, M_1.
2. Nehme die Variable $X_\kappa \in \{X_1, \ldots, X_p\}$ in das Modell auf, die ein Aufnahmekriterium aus (b) optimiert.
3. Wiederhole Schritt 2 bis ein Stoppkriterium aus (c) erfüllt ist.

Rückwärtsselektion

1. Beginne mit dem Modell größter Dimension, M_k.
2. Entferne die Variable $X_\kappa \in \{X_1, \ldots, X_p\}$, die ein Ausschlusskriterium aus (b) optimiert.
3. Wiederhole Schritt 2 bis ein Stoppkriterium aus (c) erfüllt ist.

Schrittweise Selektion

1. Beginne mit dem Modell kleinster Dimension, M_1.
2. Nehme die Variable $X_\kappa \in \{X_1, \ldots, X_p\}$ in das Modell auf, die ein Aufnahmekriterium aus (b) optimiert.
3. Entferne, wenn nötig, die bereits aufgenommene Variable $X_\kappa \in \{X_1, \ldots, X_p\}$, die ein Ausschlusskriterium aus (b) optimiert.
4. Wiederhole Schritt 2 und 3 bis ein Stoppkriterium aus (c) erfüllt ist.

(b) **Aufnahme- und Ausschlusskriterien**. Ob eine Einflussgröße $X_\kappa \in \{X_1, \ldots, X_p\}$ in das Endmodell M_{κ^*} aufgenommen wird oder nicht, kann durch eine Vielzahl an Kriterien bestimmt werden. Eine inzwischen weitreichend verwendete Möglichkeit besteht darin, eine Variable dann aufzunehmen, wenn sie ein Selektionskriterium der Abschnitte 3.2-3.5 optimiert, vergleiche hierzu insbesondere auch Abschnitt 3.5.3. Klassischerweise werden jedoch Teststatistiken beziehungsweise die dem Test zugehörigen p-Werte als Kriterium verwendet:

- Gegeben seien zwei verschachtelte lineare Modelle M_κ, M_λ, mit $M_\kappa : y = X_\kappa\beta_\kappa + \epsilon_\kappa$ und $M_\lambda : y = X_\lambda\beta_\lambda + \epsilon_\lambda$, wobei $\dim(M_\kappa) = p$, $\dim(M_\lambda) = k$, $p < k \leq n$, $(\beta'_\kappa, \beta'_{\tilde{\kappa}})' = \beta_\lambda$ und somit $\beta_\kappa \subset \beta_\lambda$. Das Testen der Hypothese

$$H_0 : \text{Submodell } M_\kappa \text{ ist gültig} \quad \text{vs.} \quad H_1 : \text{Submodell } M_\kappa \text{ ist nicht gültig}$$

entspricht dem Test $H_0 : \beta_{\tilde{\kappa}} = 0$ vs. $H_1 : \beta_{\tilde{\kappa}} \neq 0$ und wird mit Hilfe der F_{change}-*Statistik*,

$$F_{\text{change}} = \frac{SSE_{M_\kappa} - SSE_{M_\lambda}}{SSE_{M_\lambda}} \cdot \frac{n-k}{k-p} \sim F_{k-p,n-k}\,, \tag{3.2}$$

überprüft, wobei $SSE_{\text{M}\cdot}$ die Residuenquadratsummen des entsprechenden linearen Modells bezeichnet.

Bei einer Vorwärtsselektion wird aus der Menge $\mathcal{X} = \{X_1, \ldots, X_p\}$ diejenige Variable $X_\kappa \in \mathcal{X}$ aufgenommen, bei der der Test auf $H_0 : \beta_\kappa = 0$ die F_{change}-Statistik maximiert bzw. den zugehörigen p-Wert minimiert. Analog wird bei einer Rückwärtsselektion die Variable $X_\kappa \in \mathcal{X}$ entfernt, bei der der Test auf $H_0 : \beta_\kappa = 0$ die F_{change}-Statistik minimiert bzw. den p-Wert maximiert.

Die Verwendung der F_{change}-Statistik entspricht in diesem Zusammenhang einem Wald-, LQ- bzw. t-Test. Die Details der formalen Äquivalenz dieser Testentscheidungen finden sich in einschlägigen Lehrbüchern, so etwa bei Searle (1971, Seite 110ff.).

- Die für andere Regressionsmodelle verwendeten Teststatistiken basieren weitgehend auf Wald-, Score- und Likelihood-Quotienten-Tests. Im Kontext generalisierter linearer Regressionsmodelle sind diese Teststatistiken asymptotisch äquivalent und asymptotisch χ^2-verteilt, vergleiche in etwa Fahrmeir, Kneib und Lang (2007, Seite 204 ff.).

(c) **Stoppkriterien**. Verwendet man Teststatistiken als Aufnahme- bzw. Ausschlusskriterium, so sind kritische Werte einer Prüfverteilung das passende Stoppkriterium; verwendet man die zugehörigen p-Werte als Aufnahme- bzw. Ausschlusskriterium, so ist ein vorgegebenes Signifikanzniveau das passende Stoppkriterium.

Die Wahl eines kritischen Wertes bzw. Signifikanzniveaus ist gewissermaßen adhoc, weswegen sich die Voreinstellungen in statistischen Programmpakete teilweise stark voneinander unterscheiden. Hinweise zur Wahl „sinnvoller“ Werte findet man unter anderem bei Draper und Smith (1998).

Es herrscht keine generelle Übereinkunft über die genaue Gestalt der oben beschriebenen Prozeduren. Mögliche Erweiterungen und Modifikationen der in diesem Abschnitt vorgestellten Verfahren finden sich zahlreich, so etwa bei Sen und Srivastava (1990) sowie Miller (1990). Auch können einzelne Ideen des sukzessiven Hypothesentestens mit anderen Ansätzen der Modellselektion verknüpft werden. Die Algorithmen von Thall, Simon und Grier (1992) und Thall, Russell und Simon (1997) greifen eben diesen Ansatz auf.

Kernstück ihrer Idee ist die Verwendung der Kreuzvalidierung (vgl. Abschnitt 3.2.3) im Zuge einer aufwändigen Rückwärtsselektion, um ein unter bestimmten Gesichtspunkten optimales Signifikanzniveau zu bestimmen.

Selektionsprozeduren auf Basis sukzessiver Hypothesentests enthalten eine Vielzahl an Gefahren, Problemen und Unwägbarkeiten. In diesem Zusammenhang wird in der Literatur am häufigsten die fehlende Stabilität der Verfahren genannt. Verschiedene Algorithmen und Kriterien führen zu verschiedenen Ergebnissen und selbst kleinste Änderungen bezüglich der Gestalt der Selektionsprozeduren können die Wahl eine Endmodells M_{κ^*} entscheidend beeinflussen. Ausführliche Beispiele und Betrachtungen dieser Problematik finden sich unter anderem bei Breimann (1996a), Miller (1990) sowie Draper und Smith (1998).

Im Fokus der Kritik steht ferner die scheinbar willkürliche Wahl der kritischen Werte bzw. Signifikanzniveaus, die dadurch verursachte zusätzliche Unsicherheit sowie die Formulierung der Tests in Form multipler Alternativhypothesen. Akaike (1969) sieht darin die entscheidende Schwäche der oben vorgestellten Verfahren und betont die Notwendigkeit einer entscheidungstheoretisch orientierten Modellselektion. Akaike (1974) betont ferner:

> *„The use of a fixed α-level of significance for the comparison of models is wrong, because it does not take into account the increase of the variability of the estimates when the number of parameters is increased“*

Die oben erwähnten Algorithmen von Thall, Simon und Grier (1992) und Thall, Russell und Simon (1997) versuchen eben dieser ad-hoc Wahl des Signifikanzniveaus vorzubeugen. Ihre Prozeduren sind jedoch äußerst rechenintensiv und haben in der Literatur bislang kaum Beachtung gefunden.

Freedman (1983) beschreibt das häufig auftretende Phänomen, dass bei einer großen Anzahl an Regressoren nicht vorhandene Zusammenhänge als signifikant eingestuft werden. Er untermauert seine Hypothese durch Simulationsstudien und asymptotische Betrachtungen. Diese beschränken sich jedoch ausschließlich auf das lineare Modell und sind insofern mit Vorsicht zu genießen.

Ein letzter, wichtiger Kritikpunkt betrifft die Beschränkung der Prozeduren auf eine einzige Verteilungsfamilie. Zur Modellierung von Zähldaten verwendet man beispielsweise häufig Poisson-, Quasi-Poisson- oder Negativ-Binomial-Modelle. Ein Vergleich zwischen diesen Ansätzen ist jedoch nur mit Hilfe einiger Selektionskriterien der Abschnitte 3.2-3.5 möglich, nicht aber unter Verwendung sukzessiver Hypothesentests.

3.1.2 Shrinkage

Ähnlich dem Vorgehen beim sukzessiven Hypothesentesten werden bei der Verwendung von Shrinkage-Verfahren im Rahmen der Modellselektion die Parameterschätzungen β_κ als Grundlage der Entscheidung zugunsten eines Endmodells M_{κ^*} gesehen. Entscheidender Unterschied gegenüber dem klassischen Ansatz ist die Art der Parameterschätzung. Betrachtet man das lineare Modell $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, so weist die gewöhnliche Kleinste-Quadrate-Schätzung (OLS) für β – und damit auch die ML-Schätzung – in einigen Situationen Defizite auf, die mit Hilfe von Shrinkage-Verfahren teilweise oder sogar ganz umgangen werden können:

- Besitzt die Designmatrix X_κ keinen vollen Rang, so ist $X'_\kappa X_\kappa$ nicht mehr positiv definit und somit existiert auch keine eindeutige OLS-Schätzung. Insbesondere bei Multikollinearität tritt dieses Problem auf.
- Die OLS-Schätzung besitzt unter den unverzerrten Schätzern die kleinste Varianz. In Situationen bei denen die Qualität der Vorhersage von Bedeutung ist, kann jedoch ein verzerrter Schätzer, der eine noch kleinere Varianz besitzt, von größerem Nutzen sein.
- Ist die Anzahl der Variablen größer als die Stichprobe selbst, so ist ebenfalls keine OLS-Schätzung möglich. Speziell im Bereich hochdimensionaler Daten muss jedoch mit solch einer Datensituation gerechnet werden. Hier ist eine Methode die mehr als n Variablen und/oder Gruppen hochkorrelierter Variablen in das Endmodell M_{κ^*} aufnehmen kann erwünscht.

Einige bekannte Shrinkage-Verfahren, wie beispielsweise Ridge (Hoerl und Kennard (1970)), Bridge (Frank und Friedmann (1993)), Lasso (Tibsharani (1996)) und Elastic Net (Zou und Hastie (2005)) greifen einen oder mehrere der oben genannten Punkte auf, indem sie zur Parameterschätzung die Minimierung der Residuenquadratsummen unter geeigneten linearen Nebenbedingungen vorschlagen. Dies entspricht einer pönalisierten KQ-Schätzung und besitzt den angenehmen Nebeneffekt, dass bei einer passenden Pönalisierung einige Parameter implizit auf Null gesetzt werden, was einer Selektion von Variablen und damit gewissermaßen einem Modellwahlverfahren gleichgesetzt werden kann. Prinzipiell gilt, dass eine stärkere Pönalisierung zu einer stärkeren Schrumpfung aller Parameterschätzungen führt und damit zu einer weitreichenden Selektion von Variablen. Einen detaillierten Überblick über Shrinkage-Verfahren, die genaue

Gestalt der Pönalisierungen, wie auch Hinweise zur Implementierung finden sich unter anderem bei Hastie, Tibsharani und Friedman (2001) und Ulbricht (2010).

Unter den oben genannten Gesichtspunkten stellen Shrinkage-Verfahren eine sinnvolle Alternative zum klassischen Hypothesentesten, wie auch zur Modellselektion im Allgemeinen dar. Speziell bei der Betrachtung hochdimensionaler Daten können so Parameter geschätzt und selektiert, Vorhersagen verbessert sowie Gruppen hochkorrelierter Einflussgrößen in ein Endmodell aufgenommen werden. Eben dieser Punkt zeigt jedoch auch, dass die Verwendung der oben vorgestellten Verfahren vor allem interessant für Spezialfälle ist und insofern nicht zwingenderweise als ein Ansatz allgemeiner und übergreifender Modellselektion verstanden werden muss: Entsprechend dem eingangs vorgestellten Prinzip der Sparsamkeit ist es gar nicht nötig, die Aufnahme von mehr als n Variablen zu fordern. Unter diesem Gesichtspunkt ergibt es auch im allgemeinen Kontext keinen Sinn, Gruppen stark korrelierter Variablen in ein Endmodell aufzunehmen.

Ferner ist der Vergleich von Modellen verschiedener Verteilungsfamilien unmöglich; die meisten Untersuchungen befassen sich zudem ausschließlich mit linearen und generalisierten linearen Modellen. Die Bestimmung geeigneter Tuning-Parameter im Rahmen der Pönalisierungen erfordert einen Selektionsschritt, der neue Unsicherheit mit sich bringt. Meist wird hierfür das generalisierte Kreuzvalidierungskriterium verwendet, vergleiche auch Abschnitt 3.2.3.

3.2 Modellselektion auf Basis von Vorhersagefehlern

Die Beurteilung eines Modells $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$ kann über seine Vorhersagequalität erfolgen. Einige ausgewählte Kriterien, die unter diesem Gesichtspunkt konstruiert wurden, werden in den folgenden Abschnitten 3.2.2-3.2.5 vorgestellt.

3.2.1 Mallows Kriterium (C_p)

Colin Mallows entwickelte Mitte der 1960er Jahre das Selektionskriterium C_p (Mallows (1964), Gorman und Toman (1966), Mallows (1973)), das, obschon seiner Beschränkung auf das lineare Regressionsmodell, immer noch sehr häufig verwendet wird. Speziell bei der Entwicklung neuerer Ansätze der Modellmittelung (Hansen (2007), Hansen (2008a), Hansen (2008b), Hansen (2009), Hansen und Racine (2009), Abschnitt 4.2) spielt es eine tragende Rolle, weswegen es an dieser Stelle ausführlich motiviert werden soll.

Gegeben sei ein quasi-wahres, datengenerierendes lineares Regressionsmodell

$$y = X^{(k)}\beta^{(k)} + \epsilon^{(k)}, \qquad \epsilon^{(k)} \sim N(0, \sigma^2 I)\,, \tag{3.3}$$

mit $\mathrm{rg}(X^{(k)}) = k \leq n$ sowie ein Submodell als Approximation an (3.3),

$$y = X^{(p)}\beta^{(p)} + \epsilon^{(p)}, \qquad \epsilon^{(p)} \sim N(0, \sigma^2 I)\,, \tag{3.4}$$

mit $\mathrm{rg}(X^{(p)}) = p < k \leq n$. Die Störgrößen $\epsilon^{(k)}$ und $\epsilon^{(p)}$ seien dabei unabhängig und $\mathbb{E}(y) = X^{(k)}\beta^{(k)} = \mu$. Der Kleinste-Quadrate-Schätzer $\hat{\beta}^{(p)}$ von $\beta^{(p)}$ ergibt sich zu

$$\hat{\beta}^{(p)} = (X^{(p)\prime}X^{(p)})^{-1}X^{(p)\prime}y\,.$$

Der entsprechende Vorhersagewert ist damit

$$\hat{y}^{(p)} = X^{(p)}\hat{\beta}^{(p)} = P^{(p)}y\,,$$

wobei

$$P^{(p)} = X^{(p)}(X^{(p)\prime}X^{(p)})^{-1}X^{(p)\prime}$$

die Projektionsmatrix aus Modell (3.4) bezeichnet. Den mittleren quadratischen Vorhersagefehler (Mean Squared Prediction Error, MSPE) bezüglich des Erwartungswerts bei Verwendung von (3.4) anstelle von (3.3) erhält man zu

$$\begin{aligned}
\mathrm{MSPE} &= \mathbb{E}\Big[\sum_{i=1}^{n}(\hat{\mu}_i^{(p)} - \mu_i)^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{n}(\mathbf{x}_i^{(p)}\hat{\beta}^{(p)} - \mathbf{x}_i^{(k)}\beta^{(k)})^2\Big] \\
&= \sum_{i=1}^{n}\Big[\mathbb{E}(\mathbf{x}_i^{(p)}\hat{\beta}^{(p)}) - \mathbf{x}_i^{(k)}\beta^{(k)}\Big]^2 + \sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(\mathbf{x}_i^{(p)}\hat{\beta}^{(p)})\,,
\end{aligned} \tag{3.5}$$

wobei $\mathbf{x}_i^{(\cdot)}$ den i-ten Zeilenvektor der Designmatrix $X^{(\cdot)}$ bezeichnet. Da

$$\mathbb{E}(X^{(p)}\hat{\beta}^{(p)}) = \mathbb{E}(P^{(p)}y) = P^{(p)}X^{(k)}\beta^{(k)}\,, \tag{3.6}$$

erhält man für (3.5)

$$\begin{aligned}
\mathrm{MSPE} &= (X^{(k)}\beta^{(k)})'(I - P^{(p)})'(I - P^{(p)})(X^{(k)}\beta^{(k)}) \\
&\quad + \sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(\mathbf{x}_i^{(p)}\hat{\beta}^{(p)})\,.
\end{aligned} \tag{3.7}$$

Unter Verwendung von

$$(I - P^{(p)})y \sim N((I - P^{(p)})X^{(k)}\beta^{(k)}, \sigma^2(I - P^{(p)}))\,,$$

ergibt sich der erste Term aus (3.7) zu

$$\Big[\sum_{i=1}^{n}(y_i - \hat{y}_i^{(p)})^2\Big] - \sigma^2 \cdot \mathrm{sp}(I - P^{(p)})\,. \tag{3.8}$$

Der zweite Term aus (3.7) ist

$$\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(\mathbf{x}_i^{(p)}\hat{\beta}^{(p)}) \;=\; \mathrm{sp}(\mathbb{V}\mathrm{ar}(\hat{y}^{(p)})) = p\sigma^2\,, \tag{3.9}$$

da

$$\hat{y}^{(p)} = X^{(p)}\hat{\beta}^{(p)} \sim N(P^{(p)}X^{(k)}\beta^{(k)}, \sigma^2 P^{(p)})\,.$$

Unter Verwendung von (3.8) und (3.9) ergibt sich ein geschätzter MSPE zu

$$\widehat{\mathrm{MSPE}} = \sum_{i=1}^{n}(y_i - \hat{y}_i^{(p)})^2 - \sigma^2(n - 2p)\,.$$

Nach Division mit $\hat{\sigma}^2$, dem KQ-Schätzer von σ^2 aus dem vollen Modell, erhält man Mallows C_p-Kriterium zu

$$\begin{aligned} C_p &= \sum_{i=1}^{n}(y_i - \hat{y}^{(p)})^2/\hat{\sigma}^2 - n + 2p \\ &= SSE/\hat{\sigma}^2 - n + 2p\,. \end{aligned} \tag{3.10}$$

Da $\mathbb{E}(SSE) = \hat{\sigma}^2(n-p)$, und damit $\mathbb{E}(C_p) = p$, ist nicht nur ein Modell zu wählen, das C_p (und damit den MSPE) minimiert, sondern auch eines für das $C_p \approx p$ gilt.

3.2.2 Erwarteter Vorhersagefehler (EPE)

Gegeben seien ein Trainingsdatensatz $D_1 = \{(y_i^{(1)}, \mathbf{x}_i^{(1)}),\, i = 1, \ldots, n\}$ mit $\mathbf{x}_i^{(1)} = (x_{i1}^{(1)}, \ldots, x_{ip}^{(1)})$ sowie ein Testdatensatz $D_2 = \{(y_j^{(2)}, \mathbf{x}_j^{(2)}),\, j = 1, \ldots, m\}$ mit $\mathbf{x}_j^{(2)} = (x_{j1}^{(2)}, \ldots, x_{jp}^{(2)})$, deren Beobachtungen einer unbekannten Verteilung $F(y, \mathbf{x})$ entstammen. Gegeben sei ferner ein Regressionsmodell $y = f(X_1, \ldots, X_p; \theta) + \epsilon$ mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathbb{V}\mathrm{ar}(\epsilon_i) = \sigma^2$ sowie eine Regressionsschätzung $\hat{f}_{D_1}$ für f auf Basis des Trainingsdatensatzes D_1. Dann ist der erwartete Vorhersagefehler (Expected Prediction Error, EPE) durch

$$EPE \;=\; \mathbb{E}[L(y^{(2)}, \hat{f}_{D_1}(\mathbf{X}^{(2)}))] \tag{3.11}$$

gegeben, wobei $y^{(2)} = (y_1^{(2)}, \ldots, y_m^{(2)})'$, $\mathbf{X}^{(2)} = (\mathbf{x}'^{(2)}_1, \ldots, \mathbf{x}'^{(2)}_m)'$ und $L(\cdot)$ eine passende Verlustfunktion bezeichnet. Er kann durch Kreuzvalidierung (vgl. Abschnitt 3.2.3) geschätzt werden und dient als Maß der Vorhersagegüte eines Modells. Aus einer Menge an Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\}$ wird dasjenige Modell $M_{\kappa^*} \in \mathcal{M}$ gewählt, das den geringsten erwarteten Vorhersagefehler besitzt.

Für einen willkürlichen Kovariablenvektor $\mathbf{x}_0 \in D_2$ mit zugehörigem Responsewert y_0 ergibt sich der EPE bei quadratischer Verlustfunktion zu

$$\begin{aligned} EPE(\mathbf{x}_0) &= \mathbb{E}(y_0 - \hat{f}_{D_1}(\mathbf{x}_0))^2 \\ &= \sigma^2 + \mathrm{Bias}^2(\hat{f}_{D_1}(\mathbf{x}_0)) + \mathbb{V}\mathrm{ar}(\hat{f}_{D_1}(\mathbf{x}_0))\,, \end{aligned} \tag{3.12}$$

vergleiche auch Hastie, Tibsharani und Friedman (2001). Der erste Term beschreibt dabei das dem System angehörende, unvermeidbare Rauschen, das durch die Qualität der Schätzung $\hat{f}_{D_1}$ nicht beeinflusst werden kann. Der zweite und dritte Term von (3.12) widerspiegeln den quadrierten Bias sowie die Varianz von $\hat{f}_{D_1}(\mathbf{x}_0)$. Die Möglichkeit der Aufsplittung in einen Varianz- und einen Biasterm ist dabei typisch für Kriterien auf Basis von Vorhersagefehlern und findet sich bei genauerer Betrachtung auch in den Konzepten von Mallows C_p (Abschnitt 3.2.1) und dem finalen Vorhersagefehler (Abschnitt 3.2.4) wieder.

Komplexe Regressionsmodelle mit vielen Einflussgrößen führen zu einem geringen Bias, jedoch zu einer hohen Varianz; simple Regressionsmodelle mit wenigen Einflussgrößen zu hohem Bias, jedoch zu geringer Varianz. Die Minimierung von (3.11) beinhaltet somit einen Kompromiss bezüglich der Komplexität und greift damit, zumindest auf Basis einer a-posteriori-Interpretation, auch das eingangs in Kapitel 2 formulierte Prinzip der Sparsamkeit auf.

3.2.3 Kreuzvalidierungskriterium (CV)

Die Kreuzvalidierung ist ein algorithmisches Verfahren, das die Vorhersagefähigkeit eines Modells oder einer Regel prüft und kann als Schätzer des erwarteten Vorhersagefehlers (3.11) aufgefasst werden. Dabei wird ein Datensatz $D = \{(y_i, \mathbf{x}_i),\, i = 1\ldots, n\}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, in $1 < k \leq n$ disjunkte Partitionen $D = \{D_1, \ldots, D_k\}$ derselben Größe zerlegt und die ersten $k-1$ Partitionen $D_{/k} = D/\{D_k\}$ werden zum Anpassen eines Modells verwendet, die andere zur Validierung der Vorhersage. Diese Prozedur wird für alle möglichen Partitionen $D_{/j} = D/\{D_j\}$, $j = 1, \ldots, k-1$, wiederholt und das Modell mit dem geringsten mittleren Vorhersagefehler ausgewählt. Für $k = n$ ergibt sich das Kreuzvalidierungskriterium (Stone (1974), Geisser (1975)) zu

$$\mathrm{CV} \;=\; \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}_i(D_{/i})]\,, \tag{3.13}$$

wobei $L[y_i, \hat{y}_i(D_{/i})]$ eine passende Verlustfunktion zwischen dem i-ten Responsewert und dessen Vorhersage auf Basis des gefitteten Modells aus dem Teildatensatz $D_{/i}$ bezeichnet. Das Kriterium (3.13) wird oft auch als *Leave One Out - Methode* bezeichnet, die Verwendung einer Partition mit $k < n$ als k-fache Kreuzvalidierung.

Gegeben sei ein Datensatz $D = \{(y_i, \mathbf{x}_i),\, i = 1\ldots, n\}$, seine Partition $D = \{D_1, \ldots, D_n\}$ sowie ein Modell der Form $y = f(X_1, \ldots, X_p; \theta) + \epsilon$ mit $\mathbb{E}(\epsilon) = 0$. Dann ergibt sich mit einer Schätzung $\hat{f}$ für f, der Kreuzvalidierungsvorhersage $\hat{y}_i(D_{/i}) = \hat{f}^{-i}(\mathbf{x}_i)$ und quadratischer Verlustfunktion ein Kreuzvalidierungskriterium zu

$$\mathrm{CV}^* \;=\; \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}^{-i}(\mathbf{x}_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(\mathbf{x}_i)}{1 - P_{ii}}\right]^2\,,$$

wobei P_{ii} das i-te Diagonalelement der Projektionsmatrix P gemäß $\hat{y} = Py$ bezeichnet. Das generalisierte Kreuzvalidierungskriterium (GCV, Golub, Heath und Wahba (1979)) entspricht einer Schätzung von CV^* und definiert sich zu

$$\mathrm{GCV} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(\mathbf{x}_i)}{1 - \mathrm{sp}(P)/n}\right]^2\,. \tag{3.14}$$

Da im Gegensatz zum klassischen Kreuzvalidierungskriterium für das GCV nicht mehrere, sondern nur eine Regressionsschätzung angepasst werden muss, kann es einfacher und

schneller berechnet werden. Das generalisierte Kreuzvalidierungskriterium wird meist zur Bestimmung von Tuning-Parametern in der nonparametrischen Regression oder bei Shrinkage-Verfahren verwendet, vergleiche hierzu auch Eubank (1999).

3.2.4 Finaler Vorhersagefehler (FPE)

Gegeben sei ein Trainingsdatensatz $D_1 = \{(y_i^{(1)}, \mathbf{x}_i^{(1)}),\, i = 1, \ldots, n\}$, dessen Beobachtungen einer unbekannten Verteilung $F(y, \mathbf{x})$ entstammen. Gegeben sei ferner ein neuer, von $y^{(1)}$ unabhängiger Testvektor $y^{(2)}$, der derselben Verteilung folgt und eindeutig mit $\mathbf{x}^{(1)}$ assoziiert ist, so dass $D_1^* = \{(y_i^{(2)}, \mathbf{x}_i^{(1)}),\, i = 1, \ldots, n\}$. Für ein Modell $y = f(X_1, \ldots, X_p; \theta) + \epsilon$ mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathrm{Var}(\epsilon_i) = \sigma^2$ sowie einer Schätzung $\hat{f}_{D_1}$ für f auf Basis des Trainingsdatensatzes D_1 ergibt sich der finale Vorhersagefehler[6] (Final Prediction Error, FPE, Akaike (1969)) zu

$$FPE \;=\; \mathbb{E}(y^{(2)} - \hat{f}_{D_1}(\mathbf{x}^{(1)}))^2\,. \tag{3.15}$$

Sein Konzept basiert auf der Ein-Schritt-Vorhersage im Kontext autoregressiver Prozesse; er kann jedoch für Längs- und Querschnittsdaten gleichermaßen verwendet werden. Im linearen Modell, $y = X\beta + \epsilon$, $\beta = (\beta_1, \ldots, \beta_p)'$, $\epsilon \sim N(0, \sigma^2 I)$, erhält man den finalen Vorhersagefehler zu

$$FPE_{LM} \;=\; \sigma^2(1 + p/n)\,,$$

vergleiche auch Rao und Wu (2001). Die Varianz σ^2 wird dabei über $\hat{\sigma}^2 = SSE/(n-p)$ geschätzt. Es existieren zahlreiche Modifikationen des FPE, meist im Kontext autoregressiver Prozesse, die Effizienz bzw. Konsistenz verbessern sollen (Akaike (1970), Bhansali und Downham (1977), Shibata (1980), Shibata (1984)).

3.2.5 Weitere Ansätze

Es existieren eine Vielzahl anderer Selektionskriterien, die sich auf Vorhersagefehler stützen. Eines der bekanntesten ist *PRESS* von Allen (1974):

$$\mathrm{PRESS} \;=\; \sum_{i=1}^{n} (y_i - \hat{y}_i(D_{/i}))^2\,. \tag{3.16}$$

[6] Der *In-Sample Error* (vgl. Hastie, Tibsharani und Friedman (2001, Seite 202)) verwendet dasselbe Konzept wie der FPE und kann als dessen Verallgemeinerung angesehen werden.

Es entspricht offensichtlich dem n-fachen Kreuzvalidierungskriterium CV bei quadratischer Verlustfunktion.

Breimann und Freedmann (1983) motivieren ihr Selektionskriterium S_p als asymptotisch effizienten Schätzer für den mittleren quadratischen Vorhersagefehler im linearen Regressionsmodell. Unter einer Menge von Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\}$, ist dasjenige Modell $M_{\kappa^*} \in \mathcal{M}$ zu wählen, das durch

$$S_p \quad = \quad \{M_{\kappa^*} | M_{\kappa^*} = \arg \min_{\mathcal{M}, k \leq n/2} \hat{\sigma}^2_\kappa (1 + k/(n-k-1))\} \tag{3.17}$$

bestimmt wird. Dabei beschreibt $\hat{\sigma}^2_\kappa$ den KQ-Schätzer der Varianz unter Modell M_κ, $k = \mathrm{rg}(X_\kappa)$.

Es existieren ferner zahlreiche Vorschläge zur Schätzung von Vorhersagefehlern unter Verwendung von Bootstrap-Verfahren. Grundlegende und weit verbreitete Konzepte finden sich unter anderem in den Arbeiten von Efron (1979), Efron (1983), Efron (1986), Breimann (1992), Shao (1996) sowie Efron und Tibsharani (1997).

Jorma Rissanen verknüpft Ideen der Kodierungstheorie und der Statistik und begründet darauf aufbauend eine eigenständige Theorie, die er Ende der 1970er Jahre – wie in Kapitel 2 bereits erwähnt – unter dem Namen *Minimum Description Length* (MDL) zusammenfasst. Eine Eigenschaft dieser Theorie ist ihre Interpretation als Vorhersagegüte im Rahmen statistischer Modellselektion, vergleiche auch Grünwald (2005). Basierend auf diesem Konzept schlägt Rissanen (1986) ein Kriterium mit dem Namen *Predictive Least Squares* (PLS) vor. Dieses hat in der Literatur bislang jedoch kaum Beachtung gefunden; bekanntere Ansätze der MDL-Methodik werden in Abschnitt 3.5.1 erläutert.

3.3 Informationstheoretische Selektionskriterien

Mitte des 20. Jahrhunderts entwickelten Shannon (1948) und Wiener (1948) unabhängig voneinander eine der Theorie der Kommunikation, heutzutage auch besser bekannt als *Informationstheorie*. Ihr Begriff von Information formte die Grundlage für die Arbeit von Kullback und Leibler (1951), die einen neuen Distanzbegriff zum Vergleich zweier Funktionen, im Besonderen zweier Wahrscheinlichkeitsdichten, formulierten. Dieser beinhaltet eine Vielzahl technischer Vorteile, vergleiche hierzu insbesondere Kullback (1959, Kapitel 2). Die Verknüpfung eines verallgemeinerten Likelihoodprinzips mit der Informationstheorie durch die Brücke der *Kullback-Leibler-Distanz*, war zu Beginn der

1970er Jahre der Startschuss einer fundamental neuen, entscheidungstheoretisch motivierten Herangehensweise der Modellselektion. In diesem Kapitel soll Akaikes Informationskriterium, neben einer Vielzahl seiner Erweiterungen und Modifikationen, vorgestellt und diskutiert werden.

3.3.1 Akaikes Informationskriterium (AIC)

Hirotugu Akaike postulierte eine Erweiterung des klassischen Maximum-Likelihood-Prinzips (Akaike (1971)), das *Entropy Maximization Principle* (EMP)[7], auf dessen Grundlage er ein Kriterium zur Modellselektion entwickelte (Akaike (1973)). Dabei stellte er einen formalen Zusammenhang zwischen dem von ihm formulierten Prinzip und der Informationstheorie her.

Unter Akaikes Erweiterung des Maximum-Likelihood-Prinzips ist diejenige Parameterschätzung $\hat{\theta}$ für eine Wahrscheinlichkeitsverteilung $f(y;\theta)$ zu wählen, die die erwartete log-Likelihoodfunktion

$$\mathbb{E}_{\hat{\theta},Y}\{\log f(Y|\hat{\theta})\} \;=\; \mathbb{E}_{\hat{\theta}}\left\{\int f(y|\theta)\log f(y|\hat{\theta})\,dy\right\} \tag{3.18}$$

maximiert. Dabei bezeichnet Y eine Zufallsvariable, die der Verteilung mit Dichtefunktion $f(y;\theta)$ folgt und unabhängig von $\hat{\theta}$ ist.

Blackwell (1953) zeigte, dass zur Unterscheidung zweier Dichten $f(y;\theta)$ und $g(y;\theta)$ sämtlich benötigte Information (im entscheidungstheoretischen Kontext) in der zugehörigen Likelihood-Ratio-Statistik

$$T(y) \;=\; f(y|\theta)/g(y|\theta) \tag{3.19}$$

enthalten ist. Sei nun $f(y;\theta_k)$, $\theta_k \in \Theta_k$, $\Theta_k \subset \mathbb{R}^k$, das quasi-wahre, datengenerierende Modell (o.B.d.A.), $f(y;\theta_p)$, $p < k$, $\theta_p \in \Theta_p$, $\Theta_p \subset \mathbb{R}^p$, das Model basierend auf dem Parametervektor θ_p als Projektion von θ_k in den Θ_p, $g(y;\theta_p)$ ein Modell zur Approximation von $f(y;\theta_k)$, $g(y;\theta_0)$, $\theta_0 \in \Theta_p$ das approximative Modell mit $g(y;\theta_0) \equiv f(y;\theta_p)$ sowie

[7] Ludwig Eduard Boltzmann (1844–1906) entwickelte den Begriff der Entropie im Rahmen der theoretischen Physik. Negative Entropie ist Information im Sinne von Kullback und Leibler (1951). Eine Minimierung der Information entspricht daher gerade einer Maximierung der Entropie (siehe auch Seite 33) und motivierte Akaike in seiner Namensgebung. Einen genaueren Einblick hierzu geben unter anderem Burnham und Anderson (2002) und Shibata (1989).

$g(y;\hat{\theta}_p)$ ein approximatives Modell basierend auf der ML-Schätzung $\hat{\theta}_p$ von θ_0. Dann ist unter Verwendung des Likelihood-Ratios (3.19) die mittlere Menge an Information zur Unterscheidung des quasi-wahren Modells $f(y;\theta_k)$ und des approximativen Modells $g(y;\theta_0)$ durch

$$\mathbb{E}_Y\{\Phi(T(y))\} \;=\; \int f(y|\theta_k)\,\Phi\left(\frac{g(y|\theta_0)}{f(y|\theta_k)}\right) dy$$

gegeben. Dabei beschreibt $\Phi(\cdot)$ eine Funktion, die folgende Regularitätsvoraussetzungen erfüllen soll:

(i) Wenn $f(y|\theta_k) = g(y|\theta_0)$ für alle y, dann soll $\mathbb{E}_Y\{\Phi(T(y))\} = 0$ gelten und damit

$$\Phi(1) = 0\,,$$

(ii) wenn $f(y|\theta_k) \neq g(y|\theta_0)$ für mindestens ein y, dann soll $\mathbb{E}_Y\{\Phi(T(y))\} > 0$ gelten. Ist Φ zweimal differenzierbar, so entspricht diese Forderung unter angemessenen Regularitätsvoraussetzungen genau

$$\Phi^{''}(1) < 0\,.$$

Die Forderungen werden für $\Phi(\cdot) = \log(\cdot)$ erfüllt, vergleiche Akaike (1973) und insbesondere auch de Leeuw (1992). Aufgrund dieser und der vorhergehenden Überlegungen und unter Verwendung des Entropy Maximization Principle (3.18) angewandt auf die Likelihood-Ratio-Statistik (3.19) können folgende Verlustfunktion

$$L(\theta_k, \hat{\theta}_p) \;=\; \int f(y|\theta_k) \log \frac{g(y|\hat{\theta}_p)}{f(y|\theta_k)}\, dy \tag{3.20}$$

bzw. folgende Risikofunktion

$$\begin{aligned} R(\theta_k, \hat{\theta}_p) &= \mathbb{E}_{\hat{\theta}_p}\left[L(\theta_k, \hat{\theta}_p)\right] \\ &= \mathbb{E}_{\hat{\theta}_p}\left[\int f(y|\theta_k) \log \frac{g(y|\hat{\theta}_p)}{f(y|\theta_k)}\, dy\right] \end{aligned} \tag{3.21}$$

zur Bestimmung eines besten approximativen Modells $g(y;\hat{\theta}_p)$ definiert werden. Die Verlustfunktion (3.20) ist dabei formal äquivalent zur informationstheoretisch motivierten Kullback-Leibler-Distanz[8] (Kullback und Leibler (1951)),

$$\mathrm{KL}(f(y;\theta), g(y;\theta)) \;=\; \int f(y;\theta)\log\frac{f(y;\theta)}{g(y;\theta)}\,dy\,, \tag{3.22}$$

die als Pseudodistanz[9] zweier Wahrscheinlichkeitsdichten $f(y;\theta)$ und $g(y;\theta)$ angesehen werden kann. Die Risikofunktion (3.21) entspricht damit auch der erwarteten Kullback-Leibler-Distanz $\mathbb{E}_{\hat{\theta}_p}[\mathrm{KL}(f(y|\theta_k), g(y|\hat{\theta}_p))]$ und stellt den eingangs erwähnten Zusammenhang der (erweiterten) Likelihoodtheorie mit der Informationstheorie her. Die Maximierung der Risikofunktion (3.21),

$$\mathcal{Z} \;=\; R(\theta_k,\hat{\theta}_p) = \mathbb{E}_{\hat{\theta}_p}\int f(y|\theta_k)\log\left(\frac{g(y|\hat{\theta}_p)}{f(y|\theta_k)}\right)\,dy \to \max\,, \tag{3.23}$$

auf Grundlage des erweiterten Maximum-Likelihood-Prinzips (3.18) entspricht der Minimierung der erwarteten Kullback-Leibler-Distanz und damit der Zielfunktion für die Wahl eines unter diesen Gesichtspunkten besten approximativen Modells $g(y;\hat{\theta}_p)$.

Da $f(y;\theta_k)$ nicht bekannt ist, muss (3.23) geschätzt werden. In der Literatur existieren hierfür viele Wege: Amemiya (1980), Linhart und Zucchini (1986), Bozdogan (1987) und de Leeuw (1992) orientieren sich dabei stark an der ursprünglichen Version von Akaike (1973); Takeuchi (1976), Shibata (1989) und Burnham und Anderson (2002) wählen dagegen einen Ansatz, der einen stärkeren Bezug zu anderen informationstheoretischen Selektionskriterien ermöglicht und dadurch den Rahmen für eine einheitliche Theorie und Notation schafft. Hierfür trifft Shibata (1989) vier notwendige Annahmen:

A1. Sowohl der Gradientenvektor

$$g(\theta_p)' \;=\; \left(\frac{\partial}{\partial\theta_l}\log g(y|\theta_p),\; l=1,\ldots,p\right)$$

als auch die Hesse Matrix

$$H(\theta_p) \;=\; \left(\frac{\partial^2}{\partial\theta_l\partial\theta_m}\log g(y|\theta_p),\; 1\le l,m\le p\right)$$

[8] Die Kullback-Leibler-Distanz wird in der Literatur auch als Kullback-Leibler-Information, Kullback-Leibler-Divergenz, Kullback-Leibler-Diskrepanz, Negentropie und negative Entropie bezeichnet.

[9] Die Kullback-Leibler-Distanz ist keine Distanz im eigentlichen Sinne, da sie weder die Dreiecksungleichung erfüllt noch symmetrisch ist.

der log-Likelihoodfunktion $\log g(y|\theta_p)$ seien wohldefiniert.

A2. $\mathbb{E}|g(\theta_p)| < \infty$ und $\mathbb{E}|H(\theta_p)| < \infty$, wobei $|\cdot|$ den Absolutwert jeder Vektor- bzw. Matrixkomponente bezeichnet.

A3. Es existiert ein θ_0 in Θ_p, welches die Lösung von $\mathbb{E}\, g(\theta_0) = 0$ ist. Für jedes $\varepsilon > 0$ divergiert

$$\sup_{||\theta_p - \theta_0|| > \varepsilon} \log g(y|\theta_p) - \log g(y|\theta_0)$$

gegen $-\infty$ f.s..

A4. Für jedes $\varepsilon > 0$, existiert ein $\delta > 0$, so dass

$$\sup_{||\theta_p - \theta_0|| < \delta} |\mathbb{E}(\hat{\theta}_p - \theta_0)' J(\theta_0)(\hat{\theta}_p - \theta_0) - \mathrm{sp}(I(\theta_0)J(\theta_0)^{-1})| < \varepsilon\,,$$

wobei

$$I(\theta_0) = \mathbb{E}\, g(\theta_0) g(\theta_0)' \qquad \text{und} \qquad J(\theta_0) = -\mathbb{E}H(\theta_0)\,.$$

Annahme A3 sichert für $n \to \infty$, dass $\hat{\theta}_p - \theta_0$ gegen 0 konvergiert f.s., also dass $\hat{\theta}_p$ ein konsistenter Schätzer von θ_0 ist. Die Annahmen A2 und A3 implizieren, dass die Risikofunktion $R(\theta_k, \hat{\theta}_p)$ aus (3.21) ihr Maximum (bzw. die erwartete Kullback-Leibler-Distanz $\mathbb{E}_{\hat{\theta}_p}[\mathrm{KL}(f(y|\theta_k), g(y|\hat{\theta}_p))]$ ihr Minimum) an der Stelle θ_0 annimmt. Annahme A4[10] wird in Fußnote 10 diskutiert.

Die zu schätzende Zielfunktion (3.23) kann wie folgt notiert werden:

$$\begin{aligned} \mathcal{Z} &= \mathbb{E}_{\hat{\theta}_p} \int f(y|\theta_k) \log\left(\frac{g(y|\hat{\theta}_p)}{f(y|\theta_k)}\, dy\right) \\ &= \mathbb{E}_{\hat{\theta}_p} \int f(y|\theta_k) \log g(y|\hat{\theta}_p)\, dy - \int f(y|\theta_k) \log f(y|\theta_k)\, dy\,. \end{aligned}$$

[10] Annahme A4 ist eine entscheidende Voraussetzung in der Herleitung von Shibata (1989) und ein wichtiges Element der in den folgenden Abschnitten näher erläuterten Kriterien von Takeuchi (1976), Shibata (1989) und Konishi und Kitagawa (1996). Als Plausibilitätsbeweis für die vorliegende Annahme lässt sich anführen, dass unter Verwendung von (i) $E_z[z'Az] = \mathrm{sp}[A\Sigma]$ für eine nichtstochastische Matrix A, einen Vektor z mit Mittelwert Null und der Kovarianzmatrix $\Sigma = \mathbb{E}_z[zz']$ sowie (ii) $\Sigma = J(\theta_0)^{-1} I(\theta_0) J(\theta_0)^{-1}$ folgt, dass $\mathbb{E}_{\hat{\theta}_p}[(\hat{\theta}_p - \theta_0)' J(\theta_0)(\hat{\theta}_p - \theta_0)] = \mathrm{sp}[J(\theta_0)\Sigma] = \mathrm{sp}[I(\theta_0)J(\theta_0)^{-1}]$. Dies widerspiegelt qualitativ die Aussage von Annahme A4.

Eine Taylorentwicklung zweiter Ordnung von $\log g(y|\hat{\theta}_p)$ um θ_0 ergibt

$$\begin{aligned}\mathcal{Z} &\approx \mathbb{E}_{\hat{\theta}_p}\left[\int f(y|\theta_k)\,[\log g(y|\theta_0) + \frac{1}{2}(\hat{\theta}_p - \theta_0)' H(\theta_0)(\hat{\theta}_p - \theta_0)]\,dy\right] \\ &\quad - \int f(y|\theta_k)\log f(y|\theta_k)\,dy\,.\end{aligned}$$

Unter Verwendung von Annahme A4 erhält man

$$\begin{aligned}\mathcal{Z} &\approx \mathbb{E}_{\hat{\theta}_p}\left[\int f(y|\theta_k)\,\log g(y|\theta_0)\right] - \frac{1}{2}\,\mathrm{sp}(I(\theta_0)J(\theta_0)^{-1}) \\ &\quad - \int f(y|\theta_k)\log f(y|\theta_k)\,dy\,.\end{aligned}$$

Eine Taylorentwicklung zweiter Ordnung von $\log g(y|\theta_0)$ um $\hat{\theta}_p$ ergibt

$$\begin{aligned}\mathcal{Z} &= \mathbb{E}_{\hat{\theta}_p}\left[\int f(y|\theta_k)\left(\log g(y|\hat{\theta}_p) + \frac{1}{2}(\theta_0 - \hat{\theta}_p)' H(\theta_{00})(\theta_0 - \hat{\theta}_p)\right)\right] \\ &\quad -\frac{1}{2}\,\mathrm{sp}(I(\theta_0)J(\theta_0)^{-1}) - \int f(y|\theta_k)\log f(y|\theta_k)\,dy\,.\end{aligned}$$

Dabei bezeichnet θ_{00} einen Wert zwischen $\hat{\theta}_p$ und θ_0. Unter erneuter Verwendung von Annahme A4 erhält man

$$\begin{aligned}\mathcal{Z} &= \mathbb{E}_{\hat{\theta}_p}\mathbb{E}_Y\left\{\log g(y|\hat{\theta}_p)\right\} - \frac{1}{2}\mathrm{sp}(I(\theta_0)J(\theta_0)^{-1}) - \frac{1}{2}\mathrm{sp}(I(\theta_0)J(\theta_0)^{-1}) \\ &\quad - \int f(y|\theta_k)\log f(y|\theta_k)\,dy\,. \qquad (3.24)\end{aligned}$$

Sei $\widehat{\mathrm{sp}}(I(\theta_0)J(\theta_0)^{-1})$ eine Schätzung des zweiten und dritte Terms aus (3.24) und bezeichne $\mathcal{L}(\hat{\theta})$ die log-Likelihoodfunktion $\log g(y|\theta_p)$ an der Stelle $\hat{\theta}_p$ als Schätzung des ersten Terms; dann ergibt sich als Selektionskriterium

$$\mathcal{Z} \propto \mathcal{L}(\hat{\theta}) - \widehat{\mathrm{sp}}(I(\theta_0)J(\theta_0)^{-1})\,, \qquad (3.25)$$

da der letzte Term aus (3.24) beim Vergleich verschiedener Modelle der Form $g(y;\theta_p)$ als konstant angesehen werden kann. Akaike (1973) wählte als Risikofunktion nicht (3.21), sondern das minus Zweifache davon[11] und definierte folglich sein Selektionskriterium als

$$AIC = -2\mathcal{L}(\hat{\theta}) + 2K\,. \qquad (3.26)$$

[11] Das minus Zweifache des logarithmierten Likelihoodratios aus (3.19), also $-2\log T(y)$, folgt unter üblichen Regularitätsvoraussetzungen einer χ^2-Verteilung. Diese Tatsache machte sich Akaike bei der Wahl einer geeigneten Risikofunktion zu Nutze. Die Literatur bezeichnet dies oft als „historischen Grund“.

Dabei bezeichnet K die Anzahl der zu schätzenden Parameter und folgt aus der Annahme, dass $I(\theta_0) \approx J(\theta_0)$ für $f(y|\theta_k) \approx g(y|\theta_p)$. Die Maximierung von (3.23) entspricht der Minimierung von (3.26). Entsprechend wird das Modell $g(y;\hat{\theta}_p)$ gewählt, das den geringsten AIC-Wert besitzt. Für den Spezialfall eines linearen Regressionsmodells ergibt sich

$$\begin{aligned} AIC_{LM} &= -2\mathcal{L}(\hat{\theta}) + 2K \\ &= -2\left[n\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right) - \frac{1}{2\hat{\sigma}^2}||y-\hat{\mu}||^2\right] + 2K\,. \end{aligned}$$

Mit den Maximum-Likelihood Schätzern $\hat{\mu}_{\mathrm{ML}} = X(X'X)^{-1}X'y$ und $\hat{\sigma}^2 = ||y-\hat{\mu}||^2 = SSE/n$ folgt:

$$AIC_{LM} = -2\left[-n\log\frac{SSE}{n} - n\log\sqrt{2\pi} - \frac{1}{2}\right] + 2K\,.$$

Unter Vernachlässigung des konstanten Terms $n\log\sqrt{2\pi} - \frac{1}{2}$ erhält man damit:

$$AIC_{LM} \propto n\log\left(\frac{SSE}{n}\right) + 2K\,. \tag{3.27}$$

Akaikes Informationskriterium ist unter den Selektionskriterien eines der bekanntesten und am häufigsten verwendeten. Obgleich es auch allein über Akaikes erweitertes Maximum-Likelihood-Prinzip – angewandt auf den Likelihood-Ratio – motiviert werden kann, so wird vor allem die Nähe zur Informationstheorie und insbesondere das Konzept der Kullback-Leibler-Distanz mit all seinen technischen Vorteilen als überzeugender Aspekt in der Literatur genannt. Häufig kritisiert werden jedoch vor allem die zahlreichen Approximationen, insbesondere die Taylor-Entwicklungen und die Schätzung der Spur in (3.25). Die Kriterien der Kapitel 3.3.2-3.3.6 greifen eben diesen Punkt auf und schlagen unter verschiedenen Gesichtspunkten Modifikationen von Akaikes Informationskriterium vor.

3.3.2 Takeuchis Informationskriterium (TIC)

Takeuchis Informationskriterium (TIC, Takeuchi (1976)) unterscheidet sich von Akaikes Informationskriterium in der Schätzung der Spur aus (3.25). Für eine i.i.d. Stichprobe $y_1, \ldots, y_n$ bezeichne

$$g_i(\hat{\theta}_p)' \;=\; (\frac{\partial}{\partial\theta_l}\log g(y_i|\hat{\theta}_p),\; l=1,\ldots,p)$$

den Gradientenvektor und

$$H_i(\hat{\theta}_p) \;=\; (\frac{\partial^2}{\partial\theta_l\partial\theta_m}\log g(y_i|\hat{\theta}_p),\; 1\le l,m\le p)$$

die Hessematrix der log-Likelihoodfunktion $\log g(y|\theta_p)$ an der Stelle $\hat{\theta}_p$ der i-ten Beobachtung. Unter Verwendung von

$$\hat{I}(\theta_0) \;=\; \sum_{i=1}^{n} g_i(\hat{\theta}_p)g_i(\hat{\theta}_p)' \qquad \text{und} \qquad \hat{J}(\theta_0) = -\sum_{i=1}^{n} H_i(\hat{\theta}_p)$$

erhält man Takeuchis Informationskriterium als

$$TIC \;=\; -2\mathcal{L}(\hat{\theta}) + 2\,\mathrm{sp}(\hat{I}(\theta_0)\hat{J}(\theta_0)^{-1})\,. \tag{3.28}$$

Die Schätzung des zweiten Terms aus (3.28) ist technisch sehr aufwändig und bringt neue Unsicherheiten mit sich. Die ausführlichen Simulationsstudien von Burnham und Anderson (2002, Seite 384 ff.) legen nahe, dass Takeuchis Informationskriterium dem von Akaike nur dann überlegen ist, wenn das beste Kandidatenmodell bezüglich der Kullback-Leibler-Distanz weit entfernt vom wahren Modell liegt. Es ist jedoch nahezu unmöglich dies in einer realen Situation zu entscheiden, weswegen der (geringe) Qualitätsverlust des AIC aufgrund der deutlich einfacheren Implementierung meist in Kauf genommen wird.

3.3.3 Regularisiertes Informationskriterium (RIC)

Shibata (1989) bezeichnet die Restriktion der beiden Kriterien AIC und TIC auf das gewöhnliche Maximum-Likelihood-Prinzip und damit der Betrachtung von $\mathcal{L}(\hat{\theta})$ als überflüssig, da sich vorteilhafte Eigenschaften des ML-Schätzers, im Speziellen seine asymptotische Effizienz, auf die Klasse der unverzerrten Schätzer beschränkt. Im informationstheoretischen Kontext zur Unterscheidung zweier Dichten $f(y;\theta_k)$ und $g(y;\theta_p)$ ist eine solche Beschränkung jedoch überhaupt nicht nötig; es existieren Situationen in

denen die Verwendung anderer, eventuell verzerrter Schätzer durchaus einen Sinn ergibt. Die Betrachtung der pönalisierten log-Likelihoodfunktion

$$L_\nu(\theta) = \log g(y|\theta_p) + \nu k(\theta_p)\,, \qquad k(\theta_p) \leq 0,\ \nu \geq 0$$

bietet sich somit als mögliche Erweiterung für informationstheoretische Kriterien an. Bezeichne $\hat{\theta}_p^{(\nu)}$ die pönalisierte ML-Schätzung von θ_p, $y_1, \ldots, y_n$ eine i.i.d. Stichprobe und sei

$$g_{\nu i}(\hat{\theta}_p^{(\nu)})' \;=\; (\frac{\partial}{\partial \theta_l} \log g(y_i|\hat{\theta}_p^{(\nu)}) + \nu k_i(\theta_p),\ l = 1, \ldots, p)$$

der Gradientenvektor und

$$H_{\nu i}(\hat{\theta}_p^{(\nu)}) \;=\; (\frac{\partial^2}{\partial \theta_l \partial \theta_m} \log g(y_i|\hat{\theta}_p^{(\nu)}) + \nu k_i(\theta_p),\ 1 \leq l, m \leq p)$$

die Hessematrix der pönalisierten log-Likelihoodfunktion $L_\nu(\theta)$ an der Stelle $\hat{\theta}_p^{(\nu)}$ der i-ten Beobachtung. Dann ergibt sich

$$\hat{I}_\nu(\theta_0) \;=\; \sum_{i=1}^{n} g_{\nu i}(\hat{\theta}_p^{(\nu)}) g_{\nu i}(\hat{\theta}_p^{(\nu)})' \qquad \text{und} \qquad \hat{J}_\nu(\theta_0) = -\sum_{i=1}^{n} H_{\nu i}(\hat{\theta}_p^{(\nu)})\,.$$

Unter Adaption der Ideen aus den Abschnitten 3.3.1 und 3.3.2 erhält man das Regularisierte Informationskriterium (RIC) nach Shibata (1989) zu

$$RIC = -2L_\nu(\hat{\theta}_p^{(\nu)}) + 2\ \mathrm{sp}(\hat{I}_\nu(\theta_0)\hat{J}_\nu(\theta_0)^{-1})\,. \tag{3.29}$$

Die Verwendung von (3.29) erlaubt sowohl die Wahl des Pönalisierungsgewichts ν als auch eines geeigneten Modells $g(y; \hat{\theta}_p^{(\nu)})$. Für $\nu = 0$ entspricht das RIC dem TIC.

3.3.4 Korrigiertes Informationskriterium (AIC_c)

Bei der Schätzung der Risikofunktion (3.21) werden sowohl im Ansatz von Akaike (1973) als auch von Takeuchi (1976) Approximationen verwendet, die sich für kleine Stichproben als problematisch erweisen. Eine für Regressionsmodelle und einige autoregressive

Prozesse korrigierte Version des AIC (Sugiura (1978), Hurvich und Tsai (1989)) ergibt sich zu

$$AIC_c \;=\; -2\mathcal{L}(\hat{\theta}) + 2K\left(\frac{n}{n-K-1}\right). \tag{3.30}$$

Die Korrektur basiert dabei auf den Annahmen, dass das wahre Modell von unendlicher Dimension ist (also $k = \infty$) und für das approximative Modell die richtige Verteilungsannahme getroffen wurde. Diese Voraussetzungen sind stark, werden jedoch häufig erfüllt. Eine ausführliche Diskussion dieses Aspekts bieten unter anderem Burnham und Anderson (2002).

3.3.5 Informationskriterium bei Überdispersion (QAIC)

Im Rahmen generalisierter linearer Regressionsmodelle wird die bedingte Dichte des Response y bekanntermaßen häufig in Form einer Exponentialfamilie gemäß

$$f(y_i|\mathbf{x}_i;\vartheta_i,\phi) \;=\; \exp\left\{\frac{y_i\vartheta_i - b(\vartheta_i)}{a(\phi)} + c(y_i,\phi)\right\}, \quad i = 1,\ldots,n\,, \tag{3.31}$$

dargestellt. Dabei sind $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ bekannte Funktionen, ϑ beschreibt den kanonischen und ϕ den Dispersionsparameter. Dieser kann aus den Daten z.B. über

$$\hat{\phi} \;=\; \chi^2/df$$

geschätzt werden. Dabei beschreibt χ^2 die generalisierte χ^2-Statistik nach Pearson (vgl. (3.45)) und *df* steht für die Anzahl der Freiheitsgrade. Übersteigt die in den Daten beobachtete Varianz die von dem Modell (3.31) vorgesehene, so spricht man von Überdispersion. Um diesem Sachverhalt Rechnung zu tragen, kann die entsprechende Likelihood adjustiert werden (McCullagh und Nelder (1989)). Ein für Überdispersion adjustiertes Selektionskriterium (Lebreton et al. (1992)) ergibt sich somit zu

$$QAIC \;=\; -2\mathcal{L}(\hat{\theta})/\hat{\phi} + 2K\,. \tag{3.32}$$

Zur Schätzung von ϕ werden die Freiheitsgrade des saturierten Modells verwendet; die Anzahl der zu schätzenden Parameter K beinhaltet die Schätzung von ϕ. Eine für kleine Stichproben korrigierte Version des $QAIC$ findet sich bei Burnham und Anderson (2002, Seite 70).

3.3.6 Weitere Ansätze

Es existieren eine Vielzahl weiterer Ansätze, die entweder auf einer Schätzung der Kullback-Leibler-Distanz beruhen oder im weitesten Sinn Modifikationen der oben vorgestellten informationstheoretischen Kriterien sind. Diese umfassen neben Kriterien unter Berücksichtigung der Problematik fehlender Daten aus Abschnitt 5.1 und dem CAICF-Kriterium aus Abschnitt 3.5.2 auch insbesondere das *Generalisierte Informationskriterium (GIC)* von Konishi und Kitagawa (1996). Die Beschränkung des AIC auf das Likelihood-Prinzip sowie die Ungenauigkeit bei der Schätzung des zweiten Terms aus (3.25) bei grober Misspezifikation der Kandidatenmodelle sind Hauptkritikpunkte der Autoren. Sie greifen damit genau die Punkte auf, die auch Takeuchi (1976) und Shibata (1989) bereits kritisieren und als Anlass zur Modifikation des AIC nehmen. Insofern kann das GIC nicht nur als Verallgemeinerung des AIC, sondern auch des TIC und des RIC verstanden werden. Der für die konkrete Implementierung hohe technische Aufwand machen das Generalisierte Informationskriterium jedoch nur für theoretische Überlegungen brauchbar. So verwendet Shao (1997) für zahlreiche Optimalitätsbetrachtungen ausschließlich das GIC um Optimalitätseigenschaften namhafter Kriterien, wie AIC, SBC, C_p, CV oder GCV, nachzuweisen.

3.4 Bayesianische Modellselektion

Das bayesianische Paradigma kann im Rahmen der statistischen Modellselektion prinzipiell adaptiert werden. Modelle der Form (2.1) werden dabei schlicht als ganzheitliches Konstrukt interpretiert und bezüglich ihrer posteriori-Wahrscheinlichkeit $p(M_\kappa|y)$ verglichen. Grundlegende Ideen bayesianischer Modellselektion werden im Folgenden beschrieben und ausführlich diskutiert.

3.4.1 Schwarzsches Bayes-Kriterium (SBC)

Aus einer Menge von Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\}$, ist unter bayesianischen Gesichtspunkten dasjenige Modell $M_{\kappa^*} \in \mathcal{M}$ auszuwählen, welches die posteriori-Wahrscheinlichkeit

$$\begin{aligned} p(M_\kappa|y) &= p(y|M_\kappa) \cdot p(M_\kappa)/p(y) \propto p(y|M_\kappa) \cdot p(M_\kappa) \\ &\propto p(M_\kappa) \int_{\Theta_\kappa} p(y|M_\kappa, \theta_\kappa) \cdot p(\theta_\kappa|M_\kappa)\, d\theta_\kappa \end{aligned} \tag{3.33}$$

maximiert. Dabei bezeichnet $p(M_\kappa)$ die a-priori-Wahrscheinlichkeit für das Modell M_κ, $p(\theta_\kappa|M_\kappa)$ die a-priori-Wahrscheinlichkeit für den Parametervektor $\theta_\kappa \in \Theta_\kappa$, $\Theta_\kappa \subset \mathbb{R}^\kappa$, im Modell M_κ und $p(y|M_\kappa, \theta_\kappa)$ die entsprechende Likelihood. Da sich das Integral in (3.33) analytisch nicht immer einwandfrei berechnen lässt, ist eine geeignete Approximation nötig. Unter Verwendung der Laplace-Methode, und der damit verbundenen Taylor-Entwicklung um die MAP-Schätzung θ^*_κ, erhält man (vgl. Lantermann (2001, Seite 190))

$$\begin{aligned} p(M_\kappa|y) &= \mathcal{L}(\hat{\theta}) - \ln n \cdot (K/2) + \ln p(M_\kappa) + \ln p(\theta_\kappa|M_\kappa) \\ &\quad + \ln 2\pi \cdot (K/2) - \frac{1}{2} \ln \det \mathbb{E}[J(y; \theta^*_\kappa|M_\kappa)] \,, \end{aligned} \tag{3.34}$$

wobei $J(y; \theta^*_\kappa|M_\kappa)$ die beobachtete Fischer-Informationsmatrix der logarithmierten Likelihood $\ln p(y|M_\kappa, \theta_\kappa)$ an der Stelle θ^*_κ bezeichnet. Für $n \to \infty$ dominieren die ersten beiden Terme aus (3.34) die anderen und damit ergibt sich nach Multiplikation mit minus Zwei[12] das Schwarzsche Bayes-Kriterium[13] (Schwarz (1978)) zu

$$SBC = -2\mathcal{L}(\hat{\theta}) + \ln n \cdot K \,. \tag{3.35}$$

Eine Maximierung der posteriori-Wahrscheinlichkeit entspricht einer Minimierung des SBC, weswegen das Modell $M_{\kappa^*} \in \mathcal{M}$ gewählt wird, das den geringsten SBC-Wert besitzt. Draper (1995) schlägt vor, den Term $2\pi \cdot (K/2)$ aus (3.34) nicht zu vernachlässigen und somit das modifizierte Kriterium

$$SBC^* = -2\mathcal{L}(\hat{\theta}) + \ln \frac{n}{2\pi} \cdot K \tag{3.36}$$

zu verwenden. Tatsächlich ist die von Draper (1995) vorgeschlagene Modifikation dem SBC nicht zwingenderweise überlegen. Zahlreiche Beispiele hierfür finden sich in der Diskussion des Artikels von Draper (1995), vgl. Raftery, S. 78–79 und Kass und Wassermann, S. 84–85.

Ungeachtet seiner Popularität werden viele der Annahmen und Näherungen des SBC in der Literatur häufig kritisiert, vergleiche unter anderem Burnham und Anderson (2002),

[12] Die Multiplikation mit minus Zwei erfolgt zur Vergleichbarkeit mit Akaikes Informationskriterium.

[13] Das Schwarzsche Bayes-Kriterium (SBC) wird in der Literatur oft auch als *Bayesianisches Informationskriterium* (BIC) oder *Schwarzsches Informationskriterium* (SIC) bezeichnet. Dies ist jedoch irreführend, da die Ideen der bayesianischen Modellselektion nicht informationstheoretisch motiviert sind und mit der frequentistischen Interpretation der Informationskriterien teilweise in Widerspruch stehen.

Kockelkorn (2000) und Fahrmeir, Kneib und Lang (2007). Zu den wichtigsten Kritikpunkten soll im Folgenden kurz Stellung bezogen werden:

- *Die Annahme gleicher a priori Wahrscheinlichkeiten $p(M_\kappa)$ für alle Modelle $M_\kappa \in \mathcal{M}$ ist unrealistisch.* Diesem Argument ist entgegenzuhalten, dass in der Herleitung des SBC an keiner Stelle gleiche a-priori-Wahrscheinlichkeiten angenommen werden. Lediglich bei der Approximation von (3.33) durch (3.34) und der damit verbundenen Vernachlässigung des Terms $\ln p(M_\kappa)$ entfallen die a-priori-Wahrscheinlichkeiten und führen zu einer Ungenauigkeit, die leicht durch Hinzunahme des entsprechenden Terms umgangen werden kann.

- *Die verwendeten Approximationen sind zu ungenau.* Für kleine Stichprobengrößen (also $n \nrightarrow \infty$) bzw. für geringe a-priori-Wahrscheinlichkeiten $p(M_\kappa)$ sind die von Schwarz verwendeten Approximationen äußerst kritisch bzw. nicht zu vertreten. Dies folgt offensichtlich aus (3.34).

- *Nur wenn das wahre Modell in den Kandidatenmodellen enthalten ist, ist das SBC korrekt.* Die Annahme, dass das wahre Modell (implizit endlicher Dimension) in den Kandidatenmodellen enthalten ist, ist technisch gesehen nicht notwendig. Das SBC ist aber nur dann konsistent, wenn obige Annahme zutrifft, vergleiche Shao (1997) sowie Abschnitt 3.6.

Eine Vielzahl weiterer, interessanter Aspekte dieses Bayes-Ansatzes werden in Kapitel 4.1 im Kontext der dort vorgestellten Modellmittelungsverfahren diskutiert.

3.4.2 Weitere Ansätze

Es existieren einige weitere Möglichkeiten der Modellselektion basierend auf Bayes-Ansätzen. Für zwei konkurrierende Modelle $M_\kappa, M_\lambda \in \mathcal{M} = \{M_1, \ldots, M_k\}$ werden klassischerweise die posteriori-Odds

$$\frac{p(M_\kappa|y)}{p(M_\lambda|y)} = \frac{p(M_\kappa)}{p(M_\lambda)} \cdot \frac{p(y|M_\kappa)}{p(y|M_\lambda)} \tag{3.37}$$

verglichen. Bei einem Verhältnis größer Eins wird das Modell M_κ gewählt, andernfalls das Modell M_λ. Bei gleichen a-priori-Wahrscheinlichkeiten vereinfachen sich die posteriori-Odds zum sogenannten *Bayes-Faktor*

$$BF(y) = \frac{p(y|M_\kappa)}{p(y|M_\lambda)}\,. \tag{3.38}$$

Die exakte, analytische Berechnung der posteriori-Odds bzw. des Bayes-Faktors ist oft jedoch nicht möglich, weswegen andere Methoden in bayesianischen Fragestellungen häufig bevorzugt werden (vgl. Fahrmeir, Kneib und Lang (2007)). Des weiteren ergeben sich Schwierigkeiten bei schwacher a-priori-Information, insbesondere bei uneigentlichen a-priori-Verteilungen. O'Hagan (1995) schlägt in Bezug auf diese Problematik die Verwendung partieller Bayes-Faktoren vor, wobei ein Teil des Datensatzes als Trainingsdatensatz zur Bestimmung zusätzlicher a-priori-Information verwendet wird und der andere zur eigentlichen Berechnung des Bayes-Faktors.

Zahlreiche Arbeiten beschäftigen sich mit der problematischen Definition von a-priori-Wahrscheinlichkeiten für die Kandidatenmodelle. George und McCullogh (1993) sowie George und McCullogh (1997) motivieren eine differenzierte Betrachtung der a-priori-Information und schlagen unter Verwendung hierarchisch gemischter Prioris eine bayesianische Prozedur mit dem Namen *Stochastic Search Variable Selection (SVSS)* vor. Laud und Ibrahim (1995) verzichten ebenfalls auf die Spezifikation von a-priori-Wahrscheinlichkeiten und schlagen unter prädiktiven Gesichtspunkten drei – weithin unbekannte – bayesianische Selektionskriterien vor.

Eine weitere Möglichkeit bayesianischer Modellselektion bietet das *Deviance Information Criterion (DIC)* von Spiegelhalter et al. (2002), das den Einsatz von MCMC-Verfahren benötigt und in der einschlägigen Literatur durchaus häufig verwendet wird.

3.5 Weitere Ansätze

Weitere populäre und interessante Ansätze der Modellselektion, die nicht durch die Abschnitte 3.1-3.4 abgedeckt werden, sollen im Folgenden motiviert und erläutert werden.

3.5.1 Minimum Description Length

Die auf Rissanen (1978) zurückgehende Theorie der *Minimum Description Length (MDL)* vertritt das zentrale Konzept, dass jede Art von Regelmäßigkeit in den Daten $\mathcal{D}$ durch eine Komprimierung derselben ausgedrückt werden kann. Die Komprimierung erfolgt dabei durch einen Code, der anhand einer Computersprache, beispielsweise C oder Pascal, beschrieben wird. Die MDL-Methodik betrachtet diese Komprimierung als ein Lernprozess bezüglich $\mathcal{D}$ und sucht in seiner idealisierten Form nach dem kürzesten Computerprogramm, das die Daten über den Code erzeugen kann. Die Länge des Programms

wird dabei durch die sogenannte Kolmogorov-Komplexität, üblicherweise in Bits, gemessen. Dieses Grundkonzept kann praktisch nicht umgesetzt werden, da zum einen kein Computerprogramm existiert, das für beliebige Daten $\mathcal{D}$ das kürzeste Programm ermittelt um $\mathcal{D}$ zu reproduzieren und zum anderen die – möglicherweise willkürliche – Wahl der Computersprache das Ergebnis entscheidend beeinflussen kann, vergleiche Li und Vitànyi (1997).

Die MDL-Methodik in ihrer nicht idealisierten, praktikableren Form setzt die Grundkonzepte durch die Reduzierung erlaubter Codes dennoch um. Entscheidend aus statistischer Sicht ist der auf die Kraft-McMillan-Ungleichung (Kraft (1949)) zurückgehende Zusammenhang von Codelängenfunktionen und Wahrscheinlichkeitsfunktionen, wodurch sich die Einschränkung erlaubter Codes auf die Wahl geeigneter Verteilungsfamilien und damit Kandidatenmodelle $M_1, \ldots, M_k$ reduziert, vergleiche auch Grünwald (2007). Im Folgenden sollen einige unter statistischen Gesichtspunkten relevante Aspekte erläutert werden.

Beschreibe der Ausdruck der *Punkthypothese* ($\mathcal{H}$) ein Modell $f(y;\theta)$ mit einer konkreten Ausprägung von θ, sei $L(\mathcal{H})$ die Länge (in Bits) um $\mathcal{H}$ zu beschreiben und $L(\mathcal{D}|\mathcal{H})$ die Länge (in Bits) zur Beschreibung der Daten auf Basis der Punkthypothese; dann besagt das (zweiteilige) *MDL-Prinzip* das Modell $M_{\kappa^*} \in \mathcal{M} = \{M_1, \ldots, M_k\}$ zu wählen, für welches

$$L(\mathcal{H}) + L(\mathcal{D}|\mathcal{H}) \quad \rightarrow \quad \min. \tag{3.39}$$

Der erste Term kann aus statistischer Sicht als Versuch verstanden werden, ein möglichst sparsames Modell zu wählen. Der zweite Term beschreibt eine Art Reproduzierbarkeit der Daten $\mathcal{D}$ unter Verwendung einer gegebenen Punkthypothese und kann als ein Maß der Anpassungs- bzw. Vorhersagegüte eines Modells interpretiert werden. Die Minimierung beider Terme greift damit explizit das eingangs formulierte Prinzip der Sparsamkeit auf und genügt somit auch Occam‘s Razor.

Um die Umsetzung des MDL-Prinzips in der Statistik vollständig zu verstehen, ist ein detailliertes Wissen im Gebiet der Kodierungstheorie notwendig. Nur so erschließen sich dem Betrachter die weitreichenden Zusammenhänge von Codelängen und Wahrscheinlichkeitsfunktionen sowie die zentrale Rolle der Likelihood, um die Daten unter einer gegebenen Hypothese zu beschreiben. Unter Verwendung einer zweiteiligen Codelänge und unter Vernachlässigung aller Terme der Ordnung $\mathcal{O}(1)$ ergibt sich das Kriterium nach Rissanen (1978),

$$MDL \;=\; -2\mathcal{L}(\hat{\theta}) + \ln n \cdot K\,, \tag{3.40}$$

das formell dem Schwarzschen Bayes-Kriterium entspricht. Die Verwendung anderer Codelängen führt zu weiteren möglichen Selektionskriterien und ermöglicht ferner die Einbeziehung von Unsicherheit, die durch die Modellselektion verursacht wird.

Die MDL-Methodik weist viele Gemeinsamkeiten und Ähnlichkeiten zu anderen Theorien, wie dem maschinellen Lernen, der Informationstheorie und Bayes-Ansätzen auf, vergleiche auch Grünwald (2007). Einer der größten Schnittstellen liegt jedoch im Bereich der *Minimum Message Length*, begründet durch Wallace und Boulton (1968). Lantermann (2001) beschreibt diesen Zusammenhang ausführlich und diskutiert Konzepte und Formalismen dieser beiden Prinzipien sowie deren Verhältnis zum Schwarzschen Bayes-Kriterium ausführlich.

Die Theorie der Minimum Description Length greift wesentliche Punkte, die an ein Selektionsprinzip gestellt werden, auf und verzichtet auf die möglicherweise problematische Formulierung wahrer Modelle, die bei der Konzipierung klassischer Selektionskriterien oftmals möglichst genau approximiert werden sollen. Insofern bietet sich das MDL-Prinzip zur Modellwahl an. Andererseits ist das grundlegende Konzept fachfremd; Begriffe der Informations- und vor allem der Kodierungstheorie spielen in der modernen Statistik kaum eine Rolle, weswegen die Ideen der MDL-Methodik in der Literatur bisher weitgehend ignoriert bzw. nur rudimentär erwähnt werden. Was die „Komprimierung von Daten“ für die Statistik wirklich bedeutet und inwiefern diese Philosophie mit anderen Bereichen der Statistik in Einklang steht, ist bisher noch weitgehend terra incognita.

3.5.2 Dimensionskonsistente Kriterien

Als dimensionskonsistent werden Kriterien bezeichnet, bei denen die Wahrscheinlichkeit das wahre Modell M_κ^* aus der Menge $\mathcal{M} = \{M_1, \ldots, M_k\}$ der Kandidatenmodelle auszuwählen für $n \to \infty$ gegen Eins konvergiert, vergleiche hierzu auch Abschnitt 3.6. Dies impliziert, dass ein wahres Modell existiert, endlicher Dimension ist und sich in der Menge $\mathcal{M}$ befindet. Im Folgenden sollen zwei Kriterien kurz umrissen werden, die ausschließlich unter dem Gesichtspunkt der Konsistenz konstruiert wurden.

Das Kriterium von Hannan und Quinn (1979) wurde ursprünglich zur Wahl der Ordnung autoregressiver Prozesse und als Weiterentwicklung des SBC konstruiert. Da grundlegende Konzepte und Eigenschaften bei einer Verallgemeinerung nicht verloren gehen, wird es häufig in der Form

$$HQ \;=\; -2\mathcal{L}(\hat{\theta}) + c \cdot \ln\ln n \cdot K\,, \quad c > 1\,, \tag{3.41}$$

zitiert, wobei c in der Regel auf Zwei gesetzt wird. Auf diese Weise ist es mit Akaikes Informationskriterium und dem Schwarzschen Bayes-Kriterium vergleichbar. Die Autoren betonen, dass ihr Kriterium aufgrund des zweiten Terms in (3.41) gegenüber dem SBC bei kleinen Stichproben weniger dazu neigt, unterangepasste Modelle zu wählen. Obgleich häufig zitiert, spielt es in der angewandten Literatur meist eine eher untergeordnete Rolle.

Bozdogan (1987) betont, dass Akaikes Informationskriterium keineswegs für sich in Anspruch nimmt, ein konsistentes Kriterium zu sein. Dennoch schlägt er vor, das AIC hinsichtlich seiner Konsistenz zu verbessern und motiviert unter diesen Gesichtspunkten das Kriterium

$$CAICF \;=\; -2\mathcal{L}(\hat{\theta}) + (\ln n + 2) \cdot K + \ln\det(\hat{J}(\theta_0))\,, \tag{3.42}$$

mit $\hat{J}(\theta_0)$ wie in Abschnitt 3.3.2. Dieses Kriterium wird jedoch äußerst selten verwendet.

3.5.3 Ad-hoc Ansätze

Anpassungsstatistiken

Häufig wird die Zielgröße mit ihrem Fit verglichen, um so eine Aussage über die Qualität der Modelle zu gewinnen. Die unter dieser Prämisse konstruierten Anpassungsstatistiken werden entweder absolut oder in Form einer Teststatistik zur Modellselektion verwendet.

Gegeben sei eine Menge $\mathcal{M} = \{M_1, \ldots, M_k\}$ generalisierter linearer Regressionsmodelle in Form einer Exponentialfamilie gemäß

$$f(y_i|\mathbf{x}_i; \vartheta_i, \phi) \;=\; \exp\left\{\frac{y_i\vartheta_i - b(\vartheta_i)}{a(\phi)} + c(y_i, \phi)\right\}\,, \quad i = 1, \ldots, n\,,$$

für bekannte Funktionen $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ und $\mathbb{E}(y) = \mu$. Bezeichne $M_{\lambda^*} \in \mathcal{M}$ das volle Modell mit $\dim(M_{\lambda^*}) = n$, $\hat{\mu}_{\lambda^*} = y$, sowie $M_\kappa, M_\lambda \in \mathcal{M}$ zwei verschachtelte, konkurrierende Kandidatenmodelle mit $\dim(M_\kappa) = p$, $\dim(M_\lambda) = k$, $p < k < n$. Dann definiert sich die Anpassungsstatistik der *Devianz* zu

$$D(y, \hat{\mu}_\kappa) \;=\; 2\sum_{i=1}^{n}\{y_i(\hat{\vartheta}_i^{(n)} - \hat{\vartheta}_i^{(p)}) - b(\hat{\vartheta}_i^{(n)}) + b(\hat{\vartheta}_i^{(p)})\}\,, \tag{3.43}$$

wobei $\hat{\vartheta}^{(n)}$ die Maximum-Likelihood-Schätzung von ϑ im Modell M_{λ^*} bezeichnet und $\hat{\vartheta}^{(p)}$ die Maximum-Likelihood-Schätzung von ϑ im Modell M_κ. Sie entspricht dem logarithmierten Likelihood-Quotienten zwischen M_{λ^*} und M_κ und misst den Verlust an Anpassung, wenn anstelle des vollen Modells M_{λ^*} das Submodell M_κ verwendet wird. Zum Vergleich der Kandidatenmodelle M_κ und M_λ wird die Teststatistik

$$\Delta D \;=\; \frac{D(y, \hat{\mu}_\kappa) - D(y, \hat{\mu}_\lambda)}{a(\phi)} \;\overset{a}{\sim}\; \chi^2_{k-p} \tag{3.44}$$

verwendet. Analog zu den in Abschnitt 3.1.1 vorgestellten Verfahren, können so sukzessive Hypothesen bezüglich einer Veränderung der Devianz getestet werden, wobei die Nullhypothese zur Gültigkeit des Submodells M_κ für $\Delta D > \chi^2_{1-\alpha,k-p}$ verworfen wird.

Obgleich häufig verwendet, unterliegt die Modellselektion auf Basis der Devianz einigen Beschränkungen. Insbesondere gilt $\Delta D \overset{a}{\sim} \chi^2_{k-p}$ nur unter strengen Voraussetzungen, vergleiche McCullagh und Nelder (1989, Seite 118 ff.). Des weiteren erscheint es äußerst idealisiert für das volle Modell $\hat{\mu}_{\lambda^*} = y$ anzunehmen. Ein perfekter Fit wird in der Regel nur für ein Modell unendlicher Dimension erreicht; die des vollen Modells kann jedoch maximal n entsprechen, weswegen die Annahme nur für $n \to \infty$ hält.

Eine weitere, in generalisierten linearen Modellen häufig verwendete Anpassungsstatistik, ist die generalisierte χ^2-*Statistik*:

$$\chi^2 \;=\; \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathbb{V}\mathrm{ar}(\hat{\mu}_i)}\,. \tag{3.45}$$

Neben ihrer Verwendung als Gütekriterium spielt sie insbesondere bei der Definition des für Überdispersion adjustierten Informationskriteriums QAIC (Abschnitt 3.3.5) eine tragende Rolle.

Sonstige Ad-hoc Ansätze

Die in Abschnitt 3.1.1 beschriebenen Selektionsprozeduren bedienen sich häufig verschiedenster statistischer Kenngrößen, um zu entscheiden welche Variablen $X_\kappa \in \{X_1, \ldots, X_p\}$ in ein Modell aufgenommen bzw. aus einem Modell ausgeschlossen werden. Im Rahmen eines linearen Regressionsmodells, $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, verwenden statistische Softwarepakete hierfür insbesondere die Residuenquadratsummen $SSE = (y-\hat{y})'(y-\hat{y})$, die geschätzte Varianz $\hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/(n-p)$ oder (partielle) Korrelationen zwischen $X_\kappa \in \{X_1, \ldots, X_p\}$ und y. Auch wird das adjustierte Bestimmtheitsmaß (Fisher (1924)), $R^2_{adj} = \text{MSM/MST}$ mit $\text{MSM} = (\hat{y} - \bar{y})'(\hat{y} - \bar{y})/p$ und $\text{MST} = (y - \bar{y})'(y - \bar{y})/n - 1$, gelegentlich verwendet. Neben der oft äußerst willkürlichen Auswahl dieser Größen bleibt auf die bestehenden Schwierigkeiten der Modellwahl durch statistische Hypothesentests und den dort angewandten Selektionsprozeduren, wie in Abschnitt 3.1.1 beschrieben, hinzuweisen.

3.5.4 Robuste Verfahren

Ein Großteil der in den Abschnitten 3.1-3.4 vorgestellten Verfahren beruhen direkt oder indirekt auf der KQ- bzw. ML-Schätzmethodik. Es ist bekannt, dass diese Schätzungen im Allgemeinen gegenüber Ausreißern und einflussreichen Beobachtungen nicht robust sind. Insofern liegt es nahe, bekannte Prozeduren und Kriterien soweit zu modifizieren, dass sie mit diesen Schwierigkeiten gegebenenfalls umgehen können. Im Rahmen statistischer Modellselektion geschieht dies häufig unter Verwendung der von Huber (1964) vorgeschlagenen *M-Schätzer*. Ursprünglich für den einfachen und sehr allgemeinen Fall eines zu schätzenden Lageparameters konzipiert, lautet ihr Minimierungsprinzip in einer etwas allgemeineren Form: Wähle für eine parametrisierte Dichte $f(y;\theta)$ sowie eine gegebene, nicht konstante Funktion $\varphi(y,\theta)$ den Schätzer $\tilde{\theta}$ für θ, so dass

$$\sum_{i=1}^{n} \varphi(y_i, \theta) \quad \rightarrow \quad \min\,. \tag{3.46}$$

Existiert die erste Ableitung $\varphi'(y,\theta) = \partial\varphi(y,\theta)/\partial\theta = \psi(y,\theta)$, dann entspricht die Minimierung von (3.46) der Lösung der Schätzgleichung

$$\sum_{i=1}^{n} \psi(y_i, \theta) \quad \stackrel{!}{=} \quad 0\,. \tag{3.47}$$

Dieses Prinzip beinhaltet als Spezialfall die ML-Schätzung ($\varphi(y_i, \theta) = \log f(y_i|\theta)$) und ermöglicht es unter Verwendung einer geeigneten Funktion $\varphi(y, \theta)$ einflussreiche Beobachtungen bzw. Ausreißer herunterzugewichten, wodurch robuste Selektionskriterien konzipiert werden können. Details zu M-Schätzern, Ansätze zur Lösung der obigen Schätzgleichungen und die Wahl geeigneter φ- bzw. ψ-Funktionen finden sich unter anderem bei Huber (1981) sowie Maronna, Martin und Yohai (2006).

Ronchetti (1997) definiert eine robuste Version von Akaikes Informationskriterium unter Verwendung von $\varphi(y, \tilde{\theta})$ anstelle der maximierten Likelihood $\mathcal{L}(\hat{\theta})$ und erhält folglich

$$AICR = -2\sum_{i=1}^{n} \varphi(y_i, \tilde{\theta}) + 2K\,. \tag{3.48}$$

Gegeben sei ein lineares Modell $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$. Unter Verwendung von

$$\varphi(y, \beta) = \begin{cases} ((y - X\beta)/\sigma)^2/2 & \text{für } |(y - X\beta)/\sigma| < c \\ c|((y - X\beta)/\sigma)| - c^2/2 & \text{sonst} \end{cases}$$

zur Bestimmung eines M-Schätzers $\tilde{\beta}$ für β entspricht (3.48) einem Kriterium von Ronchetti (1985), das in diesem Fall die in Abschnitt 3.6 angeführten Optimalitätseigenschaften des AIC behält. Adaptiert man die Ideen des $AICR$ auf das Schwarzsche Bayes-Kriterium, so erhält man das Kriterium von Machado (1993):

$$SBCR = -2\sum_{i=1}^{n} \varphi(y_i, \tilde{\theta}) + \ln n \cdot K\,. \tag{3.49}$$

Ronchetti und Staudte (1994) schlagen eine auf M-Schätzern basierende robuste Version von Mallows C_p vor. Sie lautet

$$RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p), \tag{3.50}$$

wobei

$$W_p = \sum_{i=1}^{n} \hat{w}_i^2 \tilde{\epsilon}_i^2 = \sum_{i=1}^{n} \hat{w}_i^2 (y_i - \mathbf{x}_i^{(p)} \tilde{\beta}^{(p)})^2$$

die gewichteten Residuenquadratsummen des Submodells (3.4) beschreibt, $\tilde{\beta}^{(p)}$ ein auf $\psi(\cdot)$ beruhender M-Schätzer ist, $\hat{w}_i = \psi(\tilde{\epsilon}_i)/\tilde{\epsilon}_i$, $\hat{\sigma}^2$ eine robuste und konsistente Schätzung von σ^2 aus dem vollen Modell (3.3) bezeichnet und

$$U_p \;=\; \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\{\hat{w}_i\tilde{\epsilon}_i\}\,, \qquad V_p = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\{\hat{w}_i\mathbf{x}_i^{(p)}(\tilde{\beta}^{(p)} - \beta^{(k)})\}\,.$$

Sind die geschätzten Gewichte $\hat{w}_i$ jeweils Eins, so wird $V_p = p$, $U_p = n - p$ und damit entspricht (3.50) genau (3.10). Das Kriterium W_FC_p von Agostinelli (2002) entspricht dem robusten Mallows Kriterium RC_p mit dem Unterschied, dass die Identifizierung der Ausreißer und dadurch auch die Bestimmung der Gewichte aus dem vollen Modell (3.3) und nicht aus dem Submodell (3.4) erfolgt.

Weitere Ansätze robuster Modellselektion finden sich bei Sommer und Huggins (1996), die ein Kriterium auf Basis der Wald-Teststatistik vorschlagen sowie Ronchetti, Field und Blanchard (1997), deren Prozedur als robuste Kreuzvalidierung interpretiert werden kann.

Die in diesem Kapitel vorgestellten Verfahren bieten eine gute Möglichkeit, die Problematik von einflussreichen Beobachtungen zu berücksichtigen. Die Grundkonzepte entstammen dabei meist denen der klassischen robusten Statistik und beinhalten deren Vor- wie auch deren Nachteile. Als etwas problematisch erweist sich jedoch vor allem die unter Umständen äußerst computerintensive Berechnung der vorgestellten Kriterien. Verfahren, die auf M-Schätzern beruhen, benötigen häufig numerische Algorithmen (iteratively re-weighted least squares, Newton-Raphson) um die Schätzgleichung (3.47) zu lösen.

3.6 Asymptotische Optimalität

Die Entscheidung zu Gunsten eines Verfahrens oder eines Kriteriums kann unter verschiedenen Gesichtspunkten erfolgen. Häufig spielen inhaltliche oder konzeptuelle Erwägungen eine Rolle. Abhängig davon, ob Prädiktion oder die Quantifizierung der wesentlichen Effekte im Vordergrund steht, abhängig von thematischen Aspekten, der Gegebenheit der Daten, der Philosophie des Anwenders und abhängig von der Komplexität und dem Aufwand der Implementierung werden Modellwahlkriterien ausgewählt und verwendet. Unabhängig davon stellt sich in der Statistik prinzipiell die Frage, ob Verfahren, Methoden und Kriterien in einem gewissen Sinn optimal sind, also ob sie „schöne“ Eigenschaf-

ten besitzen, die einen Vorteil darstellen können. Seit Beginn der 1980er Jahre werden auch im Rahmen der statistischen Modellselektion Optimalitätseigenschaften diskutiert: Hierbei werden nahezu ausschließlich Effizienz und Konsistenz von Modellwahlkriterien verglichen; Eigenschaften, so wird die untenstehende Diskussion zeigen, die im vorliegenden Kontext nicht immer nützlich sein müssen.

Gegeben sei ein datengenerierendes, wahres Modell M_κ^* mit $\mathbb{E}(y) = \mu$ sowie ein Modell $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$ mit der zugehörigen Schätzung $\hat{\mu}(M_\kappa)$ für μ. Ferner sei

$$L(M_\kappa) \quad = \quad ||\mu - \hat{\mu}(M_\kappa)||^2/n \tag{3.51}$$

eine Verlustfunktion mit $||\cdot||$ als euklidischer Norm und $M_\kappa^L \in \mathcal{M}$ das quasi-wahre Modell, welches die Verlustfunktion über alle $M_\kappa \in \mathcal{M}$ minimiert. Ein Kriterium wird dann als *konsistent*[14] bezeichnet, wenn es ein Modell $M_{\kappa^*} \in \mathcal{M}$ wählt, so dass

$$\lim_{n\to\infty} P_{n,\theta}(M_{\kappa^*} = M_\kappa^L) \quad = \quad 1 \tag{3.52}$$

für alle $\theta \in \Theta$.

Ein Selektionskriterium wird als *asymptotisch effizient* bzw. *asymptotisch verlusteffizient* bezeichnet, wenn es ein Modell M_{κ^*} wählt, so dass

$$L(M_{\kappa^*})/L(M_\kappa^L) \quad \xrightarrow{p} \quad 1\,. \tag{3.53}$$

Der Nachweis der asymptotischen Effizienz erfolgt ausschließlich unter der Annahme, dass $\dim(M_\kappa^*) = \infty$, vergleiche unter anderem Leeb und Pötscher (2008a, Abschnitt 3) sowie Shao (1997).

Für den Spezialfall eines linearen Regressionsmodells, $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, sind wesentliche asymptotisch effiziente Kriterien gemäß (3.53) das AIC (Akaike (1973), Abschnitt 3.3.1), AIC_c (Sugiura (1978), Abschnitt 3.3.4), C_p (Mallows (1964), Abschnitt 3.2.1), CV (Stone (1974), Abschnitt 3.2.3), FPE (Akaike (1969), Abschnitt 3.2.4), GCV (Golub, Heath und Wahba (1979), Abschnitt 3.2.3), PRESS (Allen (1974), Abschnitt

[14] Häufig wird in der Literatur eine etwas stärkere Forderung an die Definition der Konsistenz gestellt, nämlich das wahre Modell M_κ^* mit Wahrscheinlichkeit Eins auszuwählen, also $\lim_{n\to\infty} P_{n,\theta}(M_{\kappa^*} = M_\kappa^*) = 1$; vergleiche auch Leeb und Pötscher (2005). Die vorgestellten Grundaussagen verändern sich durch diese Definition nicht, jedoch ist der Nachweis der Konsistenz unter diesen Umständen anders zu führen.

3.2.5) und S_p (Freedman (1983), Abschnitt 3.2.5), vergleiche hierzu insbesondere auch Shibata (1981), Nishii (1984), Li (1987) und Shao (1997). Diese Kriterien sind nicht konsistent.

Konsistente Kriterien im Kontext linearer Regressionsmodelle sind unter anderem das SBC (Schwarz (1978), Abschnitt 3.4.1), MDL (Rissanen (1978), Abschnitt 3.5.1), HQ (Hannan und Quinn (1979), Abschnitt 3.5.2) und CAICF (Bozdogan (1987), Abschnitt 3.5.2). Diese Kriterien sind nicht asymptotisch effizient.

In der Grundaussage ähnliche bzw. gleichwertige Ergebnisse bezüglich Konsistenz und Effizienz für AIC, AIC_C, FPE, C_p, HQ und SBC im Bereich verschiedener stationärer und nichtstationärer AR- und ARMA-Prozesse finden sich bei Hannan und Quinn (1979), Hannan (1980), Quinn (1980), Shibata (1980), Hannan (1981), Paulsen (1984), Tsay (1984) und Karagrigoriou (1997). Für den etwas allgemeineren Fall einer stochastischen Regression finden sich Konsistenzaussagen bezüglich PLS (Rissanen (1986), Abschnitt 3.2.5) bei Wei (1992), notwendige Erweiterungen und Bedingungen zur Konsistenz bekannter Kriterien bei Pötscher (1989).

Aus (3.52) wird klar, dass eine Definition der Konsistenz nur dann einen Sinn ergibt, wenn sich in der Menge der Kandidatenmodelle das wahre Modell M_κ^* (bzw. ein quasiwahres Modell M_κ^L) befindet. Dies impliziert, dass ein wahres Modell existiert und endlicher Dimension ist. In solchen, eher seltenen Situationen scheinen konsistente Kriterien einen Vorteil zu besitzen. Wie oben erwähnt, bauen ausnahmslos alle Nachweise der asymptotischen Effizienz auf der Annahme eines wahren Modells unendlicher Dimension auf. Dies impliziert, dass sich das wahre Modell nicht in der Menge der Kandidatenmodelle befindet. Diese, meist realistischere Annahme suggeriert einen Vorteil zu Gunsten der effizienten Modellwahlkriterien.

Ein endgültiger Vergleich aufgrund dieser Überlegungen ist dennoch kritisch. Es stellt sich die Frage, ob es überhaupt Ziel der Modellselektion ist, das wahre Modell M_κ^* zu finden und möglichst genau zu approximieren. Wie auch in der Literatur häufig bemerkt, ist die Annahme eines wahren Modells unendlicher Dimension in vielen Situationen zwar weitaus realistischer als die Annahme eines wahren Modells endlicher Dimension; diesem Modell im Sinne von (3.53) jedoch möglichst nahe kommen zu wollen oder gemäß (3.52) gar mit Wahrscheinlichkeit Eins auszuwählen, widerspricht dem eingangs formulierten Prinzip der Sparsamkeit. Wie in Kapitel 2 bereits diskutiert, sollen Modelle in den empirischen Wissenschaften, insbesondere auch in der Statistik, meist abstrahieren und idealisieren; sie sollen grundlegende Elemente eines Phänomens in Beziehung setzen

und nur äußerst selten ein detailliertes Abbild der Realität – inklusive aller geringen, möglicherweise zu vernachlässigenden Effekte – liefern. Dies wird auch in der Diskussion des bedeutenden Artikels von Shao (1997) deutlich: Zhang, S. 254-258, betont an dieser Stelle:

> *„Even if a true model does exist, there is still ample reason to choose simplicity over correctness knowing perfectly well that the selected model might be untrue. The practical advantage of a parsimonious model often overshadows concerns over the correctness of the model. After all, the goal of statistical analysis is to extract information rather than to identify the true model.“*

und auch Leeb und Pötscher (2008a) bemerken

> *„Since it seems to be rarely the case that the estimation of M_0 [M_κ^*] is the ultimate goal of the analysis, the consistency property of $\hat{M}$ [M_{κ^*}] may not be overly important.“*

und ergänzen

> *„The true model may be infinite-dimensional, suggesting an advantage of AIC or a related procedure over BIC. At a given sample size, however, one of the finite-dimensional candidate models may provide a very good approximation to the true data-generating process, and hence the [...] loss-efficiency result favoring AIC may not be relevant.“*

Tatsächlich ergeben sich noch viele weitere, vorwiegend technische Probleme der oben genannten Optimalitätsbegriffe. Exemplarisch zu nennen sind dabei sicherlich die üblichen Nachweise der Verlusteffizienz, etwa bei Shibata (1981), Li (1987) oder Shao (1997), die nur punktweise asymptotisch in θ gelten und somit für eine feste und endliche Stichprobengröße in Frage zu stellen sind: So zeigt Kabaila (2002), dass bereits in sehr einfachen Situationen für ein festes n die Forderung (3.53) insofern verletzt werden kann, als dass das konsistente Kriterium SBC hierbei verlusteffizienter ist wie das verlusteffiziente Kriterium AIC.

Diese und viele andere technische Sachverhalte, die oben angeführte Diskussion, wie auch die Tatsache, dass sich nahezu alle der oben genannten Resultate auf das lineare Modell bzw. auf autoregressive Prozesse beziehen, unterstreichen die Frage, ob die Konsistenz und Effizienz im Kontext der Modellselektion hilfreiche Optimalitätseigenschaften sind oder nicht. Das folgende Kapitel 4 zeigt, dass in der Tat für eine „gute“ Schätzung $\hat{\theta}$ weit mehr als die bisher angeführten Überlegungen beachtet werden müssen, nämlich auch noch zusätzlich die Unsicherheit bezüglich des gewählten Modells. Einige erste Lösungsansätze aktueller Literatur werden insbesondere in den Abschnitten 4.2.2 und 4.2.3 vorgestellt.

4. Modellmittelung

In der Statistik werden Inferenz und Modellselektion meist als zwei separate Prozesse betrachtet: Wird aufgrund eines Kriteriums oder Verfahrens ein einziges, bestes Modell M_{κ^*} aus einer Menge von Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\}$ ausgewählt, so wird dieses Modell implizit als korrekt angenommen und die dazugehörigen Parameterschätzungen und Konfidenzintervalle werden so bestimmt, als wäre das Modell *a priori* festgelegt worden. Dies ist äußerst problematisch, da der Selektionsschritt in der Regel datenbasiert ist, die Daten oft mehrere Modelle stützen und die Konstruktion eines korrekten Schätzers *post model selection*[15] somit die Selektionsprozedur und die dadurch verursachte Unsicherheit bezüglich der Wahl von M_{κ^*} berücksichtigen sollte. Die Vernachlässigung dieses Aspekts kann ernste Folgen haben; insbesondere muss damit gerechnet werden, dass die Parameterschätzungen verzerrt sind, deren Varianz systematisch unterschätzt wird und die entsprechenden Konfidenzintervalle damit durchweg zu optimistisch sind, vergleiche hierzu auch Chatfield (1995), Draper (1995), Candolo, Davison und Demétrio (2003), Hjort und Claeskens (2003), Leeb und Pötscher (2005) sowie die dort angegebenen Referenzen.

Die Unsicherheit in der Modellwahl bezieht sich dabei meist auf zwei Aspekte: 1) die Wahl einer geeigneten, möglicherweise multivariaten Wahrscheinlichkeitsverteilung $f(\cdot)$ und 2) die Wahl einer dieser Verteilung zugehörigen Parametrisierung. Betrachtet man beispielsweise im Regressionskontext eine Zielgröße y, die aus Zähldaten besteht, sowie eine Menge potentieller Kovariablen $X_1, \ldots, X_p$, so kann zur Modellierung der Abhängigkeitsstruktur $f(y|X_1, \ldots, X_p)$ etwa ein 1) Poisson-, Negativ-Binomial- bzw. Quasi-Poisson-Ansatz verfolgt werden; die Wahl einer 2) entsprechend geeigneten Parametrisierung $f(y|X_1, \ldots, X_p; \theta)$ ist in diesem Kontext äquivalent mit der Wahl geeigneter Regressoren. Unabhängig vom gewählten Verteilungsmodell führen verschiedene Kovariablenkombinationen und -transformationen zu einer unterschiedlichen Parametrisierung

[15] Geht dem Inferenzprozess ein datenbasierter Modellselektionsschritt voraus, so wird der zugehörige Parameterschätzer in der Literatur häufig auch als *Post-Model-Selection-Estimator* (PMSE) bezeichnet.

und damit zu einer unterschiedlichen Interpretation der Datenstruktur. Insbesondere der zweite der eben genannten Aspekte, also die Unsicherheit bezüglich der Wahl einer geeigneten Parametrisierung anhand der Selektion von Variablen, ist Schwerpunkt dieses Kapitels.

Bereits seit Ende der 1970er Jahre wird die Problematik einer Vernachlässigung der Modellselektion in der Inferenz auch als solche erkannt (z.B. bei Leamer (1978), Hjort (1982), Hodges (1987), Pötscher (1991)); zu einer konkreten Ausarbeitung an Lösungsansätzen führten jedoch erst einige richtungsweisende Artikel Mitte der 1990er Jahre, im Speziellen die viel zitierten Arbeiten von Draper (1995) und Chatfield (1995), der hierzu bemerkt:

> *„In practice model uncertainty is a fact of life and likely to be more serious than other sources of uncertainty which have received far more attention from statisticians“*

Er ergänzt in diesem Zusammenhang folgerichtig:

> *„[...] we must get over the key message that the properties of an estimator may depend not only on the selected model, but also on the selection process“*

Eine Möglichkeit die Unsicherheit innerhalb der Modellselektion explizit zu berücksichtigen, besteht darin, sich bei der Inferenz nicht nur auf ein einziges Modell, sondern auf mehrere Modelle zu stützen. Das Kombinieren mehrerer Modelle ist Thema dieses Kapitels und wird auch als *Modellmittelung* bezeichnet. Die Literatur unterscheidet dabei zwischen bayesianischer Modellmittelung (Bayesian Model Averaging, BMA, Abschnitt 4.1) und frequentistischer Modellmittelung (Frequentist Model Averaging, FMA, Abschnitt 4.2).

Ungeachtet der zugrundeliegenden Philosophie beschäftigen sich die nächsten beiden Abschnitte mit der Konstruktion gewichteter Parameterschätzungen der Form

$$\hat{\bar{\theta}} \;=\; \sum_{\kappa=1}^{k} w_\kappa \hat{\theta}_\kappa \,, \tag{4.1}$$

und deren Varianz $\mathbb{V}\mathrm{ar}(\hat{\bar{\theta}})$. Dabei beschreibt $\hat{\theta}_\kappa$ die Parameterschätzung im Modell $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$, die gemäß ihrer Evidenz oder aufgrund anderer Überlegungen ein Gewicht w_κ, $\sum_\kappa w_\kappa = 1$, zugewiesen bekommt, wodurch explizit die Unsicherheit bezüglich der Modellwahl mitberücksichtigt wird. Das Hauptaugenmerk liegt dabei stets auf der Konstruktion der Gewichte, wie auch der Optimalität der daraus resultierenden gewichteten Parameterschätzung. Es ist zu beachten, dass für jede Vektorkomponente $\hat{\theta}_{\kappa i} \in \hat{\theta}_\kappa$ eines Modells M_κ, die nicht in einem konkurrierenden Modell M_λ, $\lambda \neq \kappa$, vertreten ist, stets angenommen wird, dass $\hat{\theta}_{\lambda i} = 0$.

Der Schätzer (4.1) wird auch als *BMA-Schätzer* bezeichnet wenn er bayesianisch motiviert ist bzw. als *FMA-Schätzer* wenn er frequentistisch motiviert ist. Er enthält als Spezialfall auch jeden Modellselektionsschätzer, da hierbei schlicht das Modell M_{κ^*}, das durch ein Verfahren oder Kriterium Γ gewählt wird, das Gewicht $w_{\kappa^*} = 1$ erhält, alle anderen Modelle das Gewicht 0. Diese Schätzer werden im Folgenden auch als *BMS-Schätzer* (Bayesian Model Selection) bzw. *FMS-Schätzer* (Frequentist Model Selection) bezeichnet.

4.1 Der bayesianische Ansatz

Die Idee bayesianischer Modellmittelung geht zurück auf Leamer (1978). Erst der rasante Fortschritt in der Entwicklung statistischer Software ermöglichte jedoch eine Implementierung und Weiterentwicklung dieser Ideen und resultierte in einer Vielzahl an Literatur innerhalb der letzten 15 Jahre. Eine ausführliche Beschreibung der Grundkonzepte sowie die Diskussion gängiger Probleme und Aspekte der Programmierung finden sich unter anderem bei Draper (1995), Kass und Raftery (1995), Chatfield (1995), Raftery, Madigan und Hoeting (1997) und Hoeting et al. (1999).

Der im bayesianischen Kontext intuitivste Gewichtungsansatz besteht darin, die Modelle $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$ gemäß ihrer posteriori-Wahrscheinlichkeit

$$p(M_\kappa|y) = \frac{p(M_\kappa) \int_{\Theta_\kappa} p(y|M_\kappa, \theta_\kappa) \cdot p(\theta_\kappa|M_\kappa)\, d\theta_\kappa}{\sum_{\kappa=1}^{k} p(M_\kappa) \int_{\Theta_\kappa} p(y|M_\kappa, \theta_\kappa) \cdot p(\theta_\kappa|M_\kappa)\, d\theta_\kappa}, \tag{4.2}$$

wie in (3.33) beschrieben, zu gewichten, also für (4.1) die Gewichte

$$w_\kappa^{(1)} = p(M_\kappa|y) \tag{4.3}$$

zu wählen. Aus (3.34) und (3.35) folgt, dass $p(M_\kappa|y) \approx -\frac{1}{2}\text{SBC}$ und somit ergibt sich als eine weitere Möglichkeit der Gewichtung

$$w_\kappa^{(2)} = \frac{\exp(-\frac{1}{2}SBC_\kappa)}{\sum_{\kappa=1}^{k} \exp(-\frac{1}{2}SBC_\kappa)}. \tag{4.4}$$

Bezeichne $\hat{\theta}_\kappa$ einen bayesianischen Punktschätzer im Modell M_κ, beispielsweise den Erwartungswert oder den Modus der posteriori-Verteilung $p(\theta_\kappa|M_\kappa, y)$, dann lässt sich unter Verwendung von (4.3) oder (4.4) ein gewichteter Parameterschätzer $\hat{\bar{\theta}}$ gemäß (4.1) bestimmen. Seine Varianz erhält man über

$$\mathbb{V}\mathrm{ar}(\hat{\bar{\theta}}) \;=\; \mathbb{E}_{\mathcal{M}}(\mathbb{V}\mathrm{ar}(\hat{\theta}_\kappa|y, M_\kappa)) + \mathbb{V}\mathrm{ar}_{\mathcal{M}}(\mathbb{E}(\hat{\theta}_\kappa|y, M_\kappa))\,, \tag{4.5}$$

vergleiche auch Draper (1995). Der erste Term repräsentiert dabei die gewichtete Varianz innerhalb der Modelle, der zweite Term die gewichtete Varianz zwischen den Modellen, wodurch die Unsicherheit bezüglich der Modellselektion implizit berücksichtigt wird.

Der bayesianische Ansatz unterliegt einigen generellen, vorwiegend technischen Beschränkungen, deren Relevanz im Folgenden diskutiert werden soll:

- Ist die Anzahl der Kandidatenmodelle sehr groß, so erweist sich die Umsetzung der Grundkonzepte als äußerst schwierig bzw. unmöglich. Madigan und Raftery (1994) schlagen daher eine Reduzierung der in die Summation aufzunehmenden Modelle vor: Es sollen nur die Modelle verwendet werden, die – gemessen an ihrer posteriori-Wahrscheinlichkeit – eine ausreichende Erklärungskraft besitzen, also die Menge

  $$\mathcal{A}' \;=\; \left\{ M_\kappa : \frac{\max_{M_\kappa \in \mathcal{M}}\{p(M_\kappa|y)\}}{p(M_\kappa|y)} \leq C \right\},$$

  wobei C eine Konstante beschreibt. Madigan und Raftery (1994) sowie Raftery, Madigan und Volinsky (1996) verwenden bei der Auswertung von Daten dabei einen Wert von $C = 20$. In Anlehnung an Occam's Razor (vgl. Kapitel 2) sollen sparsame Modelle bevorzugt, komplexere Modelle ausgeschlossen werden. Modelle der Menge

  $$\mathcal{B} \;=\; \left\{ M_\kappa : \exists M_\lambda \in \mathcal{A}', M_\lambda \subset M_\kappa,\ \frac{p(M_\lambda|y)}{p(M_\kappa|y)} > 1 \right\}$$

  werden daher nicht weiter betrachtet. Gemäß Madigan und Raftery (1994) umfasst eine statistische Analyse somit die oben vorgestellten bayesianischen Modellmittelungsmethoden, angewandt auf die Menge $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B} \subseteq \mathcal{M}$. Dieses Vorgehen wird auch als *Occam's Window* bezeichnet. Hinweise zur Implementierung finden sich bei Volinsky et al. (1997) und Hoeting et al. (1999).

- Die in (4.2) enthaltenen Integrale sind analytisch nicht immer zu bestimmen. Eine Approximation liefert jedoch meist die Laplace-Methode (vgl. Tierney und Kadane (1986)). In vielen Fällen kann auch die Approximation durch das Schwarzsche

Bayes-Kriterium verwendet werden, vergleiche (4.4). Dies ermöglicht eine einfache Implementierung und ist Grundlage des Pakets „BMA“ (Bayesian Model Averaging, Raftery et al. (2006)) für die statistische Software *R*. Wie in Abschnitt 3.4.1 detailliert beschrieben, ist diese Approximation unter Umständen jedoch äußerst ungenau. Insofern ist ein solches Vorgehen mit Vorsicht zu genießen.

- Die Bestimmung der priori-Wahrscheinlichkeiten $p(M_\kappa)$ ist unklar. In einem Großteil der Literatur sowie in zahlreichen Anwendungen der bayesianischen Modellmittelung wird schlicht auf die Spezifikation der Wahrscheinlichkeiten verzichtet, indem alle Modelle a priori als gleich wahrscheinlich eingestuft werden, also $p(M_\kappa) = \frac{1}{k}$, $\forall M_\kappa \in \mathcal{M}$, gewählt wird. Meist wird sogar die SBC-Approximation (4.4) verwendet, wodurch implizit keine priori-Information für die Modelle spezifiziert werden muss. Diese Vorgehensweise erlaubt eine einfache und zielgerichtete Implementierung. Das bayesianische Grundverständnis priori-Wissen zu nutzen geht dabei jedoch vollständig verloren.
  Nur sehr wenige Arbeiten beschäftigen sich mit der Frage, was die priori-Wahrscheinlichkeiten $p(M_\kappa)$ wirklich bedeuten, wie sie zu interpretieren sind und wie sie konstruiert werden können. Hoeting et al. (1999) greifen eine Idee von George und McCullogh (1993) auf, und schlagen vor, für parametrische Regressionsmodelle die priori-Wahrscheinlichkeiten wie folgt zu wählen:

$$p(M_\kappa) \;=\; \prod_{j=1}^{p} \pi_j^{\delta_{\kappa j}} (1-\pi_j)^{1-\delta_{\kappa j}} \,.$$

  Dabei beschreibt $\pi_j \in [0,1]$ die priori-Wahrscheinlichkeit für $\beta_j \neq 0$ im Regressionsmodell M_κ. Die Indikatorvariable $\delta_{\kappa j}$ gibt an, ob sich die Variable j im Modell M_κ befindet oder nicht. Dieser Ansatz ist sicherlich eine erste Möglichkeit Modelle als mehr oder weniger wahrscheinlich einzustufen. Problematisch erscheint jedoch die implizite Annahme, dass das Vorhandensein oder Fehlen der einzelnen Variablen voneinander unabhängig ist.

4.2 Frequentistische Ansätze

Der hohe computationale Aufwand zur Berechnung der posteriori-Wahrscheinlichkeiten (4.2), die daraus resultierende lange Rechenzeit, die unbefriedigende Approximation

über das SBC-Kriterium sowie die unklare Bestimmung und Interpretation der priori-Wahrscheinlichkeiten $p(M_\kappa)$ führten seit Ende der 1990er Jahre zu einer Diskussion über frequentistische Alternativen in der Modellmittelung und zur Veröffentlichung einiger Artikel, die konkrete Lösungsansätze für die Konstruktion eines FMA-Schätzers entwickeln (Buckland, Burnham und Anderson (1997), Hansen (2007), Liang et al. (2010)) oder ein einheitliches frequentistisches Paradigma entwerfen (Hjort und Claeskens (2003)). In den folgenden Abschnitten 4.2.1-4.2.3 werden diese Ansätze vorgestellt und diskutiert. Abschnitt 4.2.4 erläutert wie in diesem Zusammenhang Varianzschätzungen konzipiert werden können und Abschnitt 4.2.5 überträgt einige Kernkonzepte in den Kontext der Faktorenanalyse. Weitere Beiträge und Diskussionen zu dieser Thematik finden sich auch in den hier nicht näher kommentierten Arbeiten von Yang (2001, 2003), Yuan und Yang (2005), Hjort und Claeskens (2006), Leung und Barron (2006), Hansen (2008a,b, 2009) sowie Hansen und Racine (2009).

4.2.1 Kriteriumsbasierte Schätzungen

Eine Möglichkeit die Evidenz eines Modells zu beurteilen, liegt in der Betrachtung eines beliebigen Selektionskriteriums Γ wie in Kapitel 3 beschrieben. Diese simple Einsicht kann dazu verwendet werden, Gewichte für die Kandidatenmodelle zu konstruieren. Der in diesem Zusammenhang erste in der Literatur beachtete, nicht-bayesianische Ansatz zur Modellmittelung geht auf Buckland, Burnham und Anderson (1997) zurück. Die Autoren schlagen vor, die Gewichte[16]

$$w_\kappa^{(3)} = \frac{\exp(-\frac{1}{2}AIC_\kappa)}{\sum_{\kappa=1}^{k}\exp(-\frac{1}{2}AIC_\kappa)} \tag{4.6}$$

zu verwenden und damit Modelle mit einem geringeren AIC-Wert als plausibler einzustufen und folglich auch höher zu gewichten. Im Prinzip kann anstelle des Kriteriums von Akaike auch jedes andere sinnhafte Kriterium Γ verwendet werden, siehe hierzu auch Abschnitt 5.2.2. Dieser Ansatz erscheint insgesamt plausibel, auch unter Berücksichtigung von (4.4). Dennoch ist die Konstruktion ad-hoc; die Arbeiten von Hjort und Claeskens (2003), Hansen (2007) sowie Liang et al. (2010) zeigen ein in Simulationsstudien bezüglich dem MSE durchweg schlechtes Abschneiden der zugehörigen FMA-Schätzung $\hat{\hat{\theta}} = \sum_{\kappa=1}^{k} w_\kappa^{(3)}\hat{\theta}_\kappa$. Aufgrund der einfachen Implementierung hat das Konzept jedoch

[16] Die Gewichte $w_\kappa^{(3)}$ werden in der Literatur auch häufig als *Akaike Gewichte*, *AIC-Gewichte* oder *geglättete AIC-Gewichte* bezeichnet.

bereits eine breite Verwendung gefunden, insbesondere in Biologie, Ökonometrie und Medizin. Als Beispiele für eine erfolgreiche und anschauliche Umsetzung sind unter anderem die Arbeiten von Hayes, Weikel und Huso (2003), Reid et al. (2003), Candolo, Davison und Demétrio (2003) und Mackenzie et al. (2005) zu nennen.

Buckland, Burnham und Anderson (1997) beschreiben ferner die Möglichkeit, Gewichte über Bootstrapping zu erhalten. Sei B die Anzahl der Bootstrap-Stichproben und b_κ die Anzahl der Fälle in denen das Modell $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$ aufgrund von Akaikes Informationskriterium (oder eines anderen Verfahrens Γ) gewählt wird. Dann ergeben sich als mögliche Gewichte

$$w_\kappa^{(4)} \;=\; \frac{b_\kappa}{B}\,. \tag{4.7}$$

Burnham und Anderson (2002, Seite 90 ff.) bemerken, dass für verlässliche Schätzungen meist 10.000 Bootstrap-Stichproben oder mehr gezogen werden sollten. Aufgrund des hohen computationalen Aufwands und der großen Menge an alternativen Gewichtsschätzungen wird diese Methodik eher selten verwendet.

4.2.2 Der MMA-Schätzer

Gegeben sei eine Menge an Regressoren $\mathcal{X} = \{X_1, \ldots, X_k\}$ sowie eine Menge an linearen Modellen $\mathcal{M} = \{M_1, \ldots, M_k\}$, so dass jedes Modell $M_\kappa \in \mathcal{M}$, $\kappa = 1, \ldots, k$, die Gestalt

$$y \;=\; \sum_{j=1}^{\kappa} \beta_j X_j + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)\,,$$

besitzt. Dies bedeutet implizit, dass die k Regressoren geordnet vorliegen müssen und die Menge $\mathcal{M}$ der Kandidatenmodelle aus k verschachtelten Modellen besteht. Ein gewichteter Parameterschätzer zur Modellmittelung ist für diese Situation gemäß (4.1) durch $\hat{\tilde{\beta}} = \sum_{\kappa=1}^{k} w_\kappa \hat{\beta}_\kappa$ gegeben, wobei $\hat{\beta}_\kappa$ den Parameterschätzer aus Modell M_κ bezeichnet. Daraus ergibt sich der geschätzte, gewichtete Erwartungswert zu $\hat{\mu}_w = X_k \hat{\tilde{\beta}}$. Hansen (2007) definiert in Anlehnung an Mallows C_p (Abschnitt 3.2.1) das Kriterium

$$\tilde{C}_p \;=\; (y - X_k\hat{\tilde{\beta}})'(y - X_k\hat{\tilde{\beta}}) + 2\sigma^2 K_w\,, \tag{4.8}$$

wobei $K_w = \mathrm{sp}\,(P_w)$ mit $P_w = \sum_{\kappa=1}^{k} w_\kappa P_\kappa$[17], $P_\kappa = X_\kappa(X_\kappa' X_\kappa)^{-1} X_\kappa'$. Die Varianz σ^2 kann dabei über die KQ-Schätzung im vollen Modell M_k geschätzt werden. Er schlägt vor, die Gewichte zur Modellmittelung so zu wählen, dass

$$w_\kappa^{(5)} \;=\; \arg\min_{w_\kappa \in \mathcal{H}} \tilde{C}_p\,, \tag{4.9}$$

wobei $\mathcal{H} = \{w_\kappa \in [0,1]^k : \sum_{\kappa=1}^{k} w_\kappa = 1\}$. Der zugehörige FMA-Schätzer $\hat{\tilde{\beta}} = \sum_{\kappa=1}^{k} w_\kappa^{(5)} \hat{\beta}_\kappa$ wird in Anlehnung an Mallows C_p als Mallows-Model-Averaging-Schätzer, kurz *MMA-Schätzer*, bezeichnet. Da der erste Term in (4.8) offensichtlich quadratisch bezüglich w_κ ist, der zweite linear, lassen sich die Gewichte über quadratische Optimierung ermitteln. Statistische und mathematische Software beinhalten in der Regel passende numerische Algorithmen, so dass die Berechnung von (4.9) prinzipiell unproblematisch ist.

Der Ansatz von Hansen ist der erste, der Gewichte unter dem Gesichtspunkt der Optimalität konstruiert. Konkret sind damit die folgenden beiden Charakteristika gemeint:

(i) Die Verwendung der gewichteten Schätzung $\hat{\tilde{\beta}} = \sum_{\kappa=1}^{k} w_\kappa^{(5)} \hat{\beta}_\kappa$ und damit von $\hat{\mu}_w = X_k \hat{\tilde{\beta}}$ garantiert unter bestimmten Voraussetzungen ein asymptotisch effizientes Verfahren ähnlich (3.53)[18], vergleiche Hansen (2007, Theorem 1).

(ii) $\mathbb{E}(\tilde{C}_p) = \mathbb{E}[(\hat{\mu}_w - \mu)'(\hat{\mu}_w - \mu)] + n\sigma^2$, vergleiche Hansen (2007, Lemma 3). In Analogie zu Mallows C_p steht damit die Minimierung des mittleren quadratischen Vorhersagefehlers im Vordergrund. Der einzige Unterschied liegt in der Betrachtung des Risikos für den gemittelten Erwartungswert $\hat{\mu}_w$ anstelle des Erwartungswerts auf Basis von (3.4).

Gleichwohl ergeben sich Probleme, die eine praktische Umsetzung des MMA-Schätzers schwierig gestaltet: Eine natürliche Anordnung der Regressoren existiert fast nie und eine willkürliche Anordnung kann das Ergebnis entscheidend beeinflussen, vergleiche hierzu

[17] P_w kann als gewichtete Projektionsmatrix interpretiert werden. Diese ist zwar symmetrisch, i.d.R. jedoch nicht mehr idempotent.

[18] Es ist für (3.53) lediglich erforderlich, den gemittelten Erwartungswert $\hat{\mu}_w$ an Stelle von $\mu(M_\kappa)$ zu betrachten und die Optimalität für alle $w_\kappa \in \mathcal{H}$ anstatt für alle $M_\kappa \in \mathcal{M}$ zu fordern. Hansen (2007) zeigt damit, dass $L(\mathcal{M}; w^{(5)}) / \inf_{w \in \mathcal{H}} L(\mathcal{M}; w) \xrightarrow{p} 1$, wobei $L(\mathcal{M}; w) = ||\mu - \hat{\mu}_w||^2/n$. Sein Nachweis unterliegt insofern recht strengen Voraussetzungen, als dass gefordert wird, dass die Gewichte der diskreten Menge $\mathcal{H}_N = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$, $N \in \mathbb{N}$, entstammen. Wan, Zhang und Zou (2010) zeigen jedoch später, dass diese Einschränkung nicht nötig ist und der Nachweis der asymptotischen Effizienz für jede beliebige Menge $\mathcal{H}$ gilt.

auch Liang et al. (2010). Zudem beschränkt sich die Verwendung auf das lineare Modell, was ein entscheidender Nachteil gegenüber allen bisher vorgestellten bayesianischen wie auch frequentistischen Ansätzen darstellt. Eine ausführlichere Diskussion dieser und anderer Aspekte erfolgt in den Simulationsstudien in Kapitel 6.

4.2.3 Der OPT-Schätzer

Der OPT-Schätzer von Liang et al. (2010) ist – motiviert durch die Arbeit von Hansen (2007) – ebenfalls unter den Gesichtspunkten der Optimalität konstruiert, lehnt sich in Notation und Denkweise jedoch stark an die Pretest-Literatur, insbesondere die Artikel von Magnus und Durbin (1999), Danilov und Magnus (2004) und Magnus, Powell und Prüfer (2008) an. Gegeben sei ein lineares Regressionsmodell $y = X\alpha + Z\gamma + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, bei dem X eine $n \times d$ Designmatrix beschreibt, die diejenigen Regressoren enthält, die aus theoretischen oder anderen Gründen auf jeden Fall in ein Endmodell aufgenommen werden sollten und Z eine $n \times q$ Designmatrix, die diejenigen Regressoren enthält, die potentiell in ein Endmodell aufgenommen werden können. Der $d \times 1$ Vektor α ist dabei der Parameter von Interesse, der $q \times 1$ Vektor γ wird als nuisance Parameter behandelt, um eine „bessere“ Schätzung von α zu erhalten; insgesamt können also 2^q verschiedene Kandidatenmodelle betrachtet werden. Der restringierte ($\gamma = 0$) und der unrestringierte KQ-Schätzer von α ergibt sich jeweils zu $\hat{\alpha}_r = (X'X)^{-1}X'y$ bzw. $\hat{\alpha}_u = \hat{\alpha}_r - Q\hat{\zeta}$, wobei $Q = (X'X)^{-1}X'ZC$, $C = (Z'\{I - P\}Z)^{-\frac{1}{2}}$, $P = X(X'X)^{-1}X'$, $\hat{\zeta} = CZ'(I - P)y$, $\hat{\zeta} \sim N(\zeta, \sigma^2 I_q)$, vergleiche auch Magnus und Durbin (1999). Betrachtet man alle 2^q Kandidatenmodelle, die sich aus den q potentiellen Einflussfaktoren $Z = (Z_1, \ldots, Z_q)$ ergeben, so definiert sich der κ-te, $\kappa = 1, \ldots, 2^q$, restringierte KQ-Schätzer $\hat{\alpha}_\kappa$ von α als

$$\hat{\alpha}_\kappa \;=\; \hat{\alpha}_r - QW_\kappa\hat{\zeta}$$

mit $W_\kappa = I_q - P_\kappa$, $P_\kappa = CS_\kappa(S_\kappa'C^2S_\kappa)^{-1}S_\kappa'C$. Dabei bezeichnet S_κ' eine $r \times q$ Selektionsmatrix, $r = \dim(M_\kappa)$, die gemäß der Restriktion $S_\kappa'\gamma = 0$ angibt, welche Kovariablen $Z^* \subseteq Z$ im κ-ten Modell von Interesse sind. Für den nuisance Parameter γ erhält man entsprechend die KQ-Schätzung $\hat{\gamma}_\kappa = CW_\kappa\hat{\zeta}$. Der *OPT-Schätzer* ist derjenige FMA-Schätzer

$$\hat{\bar{\alpha}} \;=\; \sum_{\kappa=1}^{2^q} w_\kappa^{(6)}\hat{\alpha}_\kappa\,, \tag{4.10}$$

der gemäß dem schwachen MSE-Kriterium (Wallace (1972)) die Gewichte $w_\kappa^{(6)}$ so wählt, dass

$$w_\kappa^{(6)} = \arg\min_{w_\kappa \in \mathcal{H}} \text{sp}\{\widehat{MSE}(\hat{\tilde{\alpha}})\}, \tag{4.11}$$

wobei $\mathcal{H} = \{w_\kappa \in [0,1]^{2^q} : \sum_{\kappa=1}^{2^q} w_\kappa = 1\}$. Auf Basis dieser Gewichte erhält man entsprechend den gewichteten KQ-Schätzer $\hat{\tilde{\gamma}} = \sum_{\kappa=1}^{2^q} w_\kappa^{(6)} \hat{\gamma}_\kappa$ für den nuisance Parameter γ. Ein alternativer OPT-Schätzer, der unter Aspekten der Vorhersagequalität verwendet werden kann, entspricht gemäß Liang et al. (2010) demjenigen FMA-Schätzer, der die Gewichte (4.11) verwendet, die nicht die Spur des geschätzten MSE von $\hat{\tilde{\alpha}}$ minimieren, sondern die Spur des MSE von $\hat{\tilde{\mu}} = H \sum_{\kappa=1}^{2^q} w_\kappa^{(7)} \hat{\beta}_\kappa$, $\hat{\beta}_\kappa = (\hat{\alpha}_\kappa', \hat{\gamma}_\kappa')'$, $H = (X, Z)$.

Der OPT-Schätzer stellt gegenüber dem MMA-Schätzer insofern eine klare Weiterentwicklung dar, als dass auch ohne eine explizite Anordnung der Regressoren die geforderte Optimalitätseigenschaft (nämlich die Gültigkeit von (4.11)) erhalten bleibt; die Simulationen von Liang et al. (2010) unterstützen diese Annahme ausnahmslos und zeigen, dass willkürliche Umordnungen der Regressoren einen entscheidenden Einfluss auf das Verhalten des MMA-Schätzers besitzen.

Problematisch erscheint jedoch der dem Ansatz innewohnende, hohe computationale Aufwand. Zum einen erfordert die explizite Minimierung der Spur des MSE von $\hat{\tilde{\alpha}}$ einen großen technischen Einsatz, vergleiche hierzu auch Liang et al. (2010, Theorem 2); zum anderen ergibt sich aus der Notwendigkeit über alle 2^q Modelle zu mitteln, bei einer großen Anzahl an potentiellen Einflussgrößen ein Rechenaufwand, der für eine explizite Berechnung des OPT-Schätzers zu hoch sein kann.

4.2.4 Schätzung der Varianz

Um Konfidenzintervalle für einen FMA-Schätzer bestimmen zu können, wird eine Schätzung für dessen Varianz benötigt. Derzeit werden hierfür vor allem zwei Ansätze diskutiert und verwendet: Der erste findet seinen Ursprung in der Arbeit von Buckland, Burnham und Anderson (1997), die vorschlagen die Varianz von $\hat{\tilde{\theta}}$ gemäß

$$\widehat{\mathbb{V}\text{ar}}(\hat{\tilde{\theta}}) = \left\{ \sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\theta}_\kappa | M_\kappa) + (\hat{\theta}_\kappa - \hat{\tilde{\theta}})^2} \right\}^2 \tag{4.12}$$

zu schätzen, wodurch – wie im bayesianischen Kontext – sowohl die Varianz innerhalb der Modelle als auch die Varianz zwischen den Modellen berücksichtigt wird. Dieser Vorschlag ist in erster Linie pragmatisch motiviert und nicht zu 100% korrekt. So zeigen Hjort und Claeskens (2003), dass die aus (4.12) bestimmten Konfidenzintervalle die vorgegebene Sicherheit in der Regel nicht exakt einhalten. Insbesondere die Simulationsstudien in Kapitel 6 legen jedoch auch nahe, dass die Verwendung von (4.12) meist zu sehr guten und realistischen Ergebnissen führt, die den Ansprüchen die Unsicherheit bezüglich der Modellselektion zu berücksichtigen, gerecht wird.

Alternativ bieten sich Bootstrap-Verfahren zur Schätzung der Varianz an. Hierbei werden in der Regel die Beobachtungen (und nicht etwa – wie sonst häufig – die Residuen) zum Resampling verwendet und B Bootstrap-Stichproben von gleichem Umfang wie in den Originaldaten gezogen. In jeder Bootstrap-Stichprobe b_j, $j = 1, \ldots, B$, wird $\hat{\tilde{\theta}}_{b_j}$ berechnet und gemäß der Stichprobenvarianz ergibt sich

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\tilde{\theta}}) \;=\; \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\tilde{\theta}}_{b_j} - \hat{\bar{\tilde{\theta}}}_B)^2\,, \tag{4.13}$$

wobei $\hat{\bar{\tilde{\theta}}}_B = \frac{1}{B}\sum_{j=1}^{B}\hat{\tilde{\theta}}_{b_j}$. Sowohl die Datenbeispiele von Buckland, Burnham und Anderson (1997) als auch die Simulationen von Candolo, Davison und Demétrio (2003) legen nahe, dass die Verwendung von (4.13) zu vernünftigen Ergebnissen führt, die in etwa in der Größenordnung von (4.12) liegen.

4.2.5 Modellmittelung in der Faktorenanalyse

Der folgende Abschnitt soll am Beispiel einer explorativen Maximum-Likelihood-Faktorenanalyse exemplarisch aufzeigen, wie die Ideen frequentistischer Modellmittelung außerhalb traditioneller Regressionsbeispiele im Rahmen multivariater, statistischer Modellierung verwendet werden könnten. Einige Kernkonzepte und die daraus resultierenden Chancen und Risiken werden an dieser Stelle eher knapp erläutert; eine detaillierte Diskussion erfolgt in Abschnitt 7.3 anhand eines konkreten Datenbeispiels.

Der Grundgedanke der Faktorenanalyse besteht darin, die Information einer Datenmatrix, die sich aus p beobachtbaren Variablen zusammensetzt, in Form einer geringeren Anzahl von k latenten Faktoren zusammenzufassen; also die Dimension der Daten soweit zu reduzieren, dass möglichst wenige Faktoren möglichst viel der Variation der

ursprünglichen Datenmatrix wiedergeben und gegebenenfalls zur inhaltlichen Interpretation und für weitere Analysen verwendet werden können. Die Wahl einer geeigneten Anzahl an Faktoren ist ein klassisches Modellselektionsproblem, das in praktischen Fragestellungen meist über Likelihood-Quotienten-Tests, ad-hoc Ansätzen wie dem Eigenwertkriterium (Guttmann (1954)) oder dem Scree-Plot (Cattell (1966)) sowie auf Basis bekannter Kriterien (z.B. AIC, SBC, CV) gelöst wird, vergleiche hierzu auch Costa (1996). Unabhängig von der dafür gewählten Methodik und den damit eventuell individuell auftretenden Problemen verursacht auch hier der datengestütze Selektionsschritt eine Unsicherheit, die bei den klassischen Inferenzprozeduren nicht berücksichtigt wird.

Gegeben sei das faktoranalytische Modell

$$X' \;=\; \Gamma^{(k)} F^{(k)} + U\,, \tag{4.14}$$

wobei X die $n \times p$ Matrix der beobachteten Daten, $\Gamma^{(k)}$ die $p \times k$ Ladungsmatrix, F die $k \times n$ Matrix der k Faktoren und U die $p \times n$ Matrix der Störterme beschreibt, $k < p$. Die Matrizen F, U und X werden dabei jeweils als multivariat normalverteilt angenommen mit Erwartungswert 0 und den zugehörigen Kovarianzmatrizen I, $\Psi = \mathrm{diag}(\Psi_1^2, \ldots, \Psi_p^2)'$ und $\Sigma = \Gamma^{(k)}\Gamma^{(k)\prime} + \Psi$. Eine Möglichkeit die Anzahl der Faktoren in einer Maximum-Likelihood-Faktorenanalyse zu bestimmen, liegt in der Verwendung von Akaikes Informationskriterium,

$$\begin{aligned} AIC_{\mathrm{FA}} &= -2\mathcal{L}(\hat{\Gamma}^{(k)}, \hat{\Psi}) + 2K \\ &\propto \ln\det\hat{\Sigma} + \mathrm{sp}(S\hat{\Sigma}^{-1}) + 2p(k+1) - k(k-1)\,, \end{aligned} \tag{4.15}$$

wobei $\hat{\Gamma}^{(k)}$ und $\hat{\Psi}$ die ML-Schätzungen von $\Gamma^{(k)}$ und Ψ bezeichnen, S die empirische Kovarianzmatrix von X ist und $\hat{\Sigma}^{-1} = (\hat{\Gamma}^{(k)}\hat{\Gamma}^{(k)\prime} + \hat{\Psi})^{-1}$, vergleiche hierzu auch Akaike (1987). In Analogie zu dem Vorgehen in Abschnitt 4.2.1 und unter Verwendung der exponentiellen AIC-Gewichte (4.6) erhält man einen FMA-Schätzer für Γ bzw. Ψ zu

$$\hat{\bar{\Gamma}} \;=\; \sum_{\kappa=1}^{k^*} w_\kappa^{(3)}\, \hat{\Gamma}_\kappa^{(k)} \qquad \text{bzw.} \qquad \hat{\bar{\Psi}} = \sum_{\kappa=1}^{k^*} w_\kappa^{(3)}\, \hat{\Psi}_\kappa\,, \tag{4.16}$$

wobei k^* die Anzahl der Kandidatenmodelle beschreibt. Voraussetzung für die Verwendung von (4.16) ist dabei die Beachtung des Rotationsproblems bei der Schätzung von $\Gamma^{(k)}$: Nur wenn bei allen betrachteten k^* Modellen die gleiche Rotationsmethode für die entsprechende Ladungsmatrix verwendet wird, ergibt der zugehörige FMA-Schätzer einen Sinn. Auch wenn in der explorativen Faktorenanalyse üblicherweise die

Punktschätzungen, insbesondere diejenigen von Γ, von Interesse sind, so kann gegebenenfalls natürlich auch deren Varianz, beispielsweise über Bootstrapping wie in Abschnitt 4.2.4 beschrieben, geschätzt werden. Zu Ende von Abschnitt 7.3 wird diese Problematik noch einmal aufgegriffen und diskutiert.

Die Nachteile und Beschränkungen eines Modellmittelungsschätzers in der Faktorenanalyse ergeben sich größtenteils aus der Faktorenanalyse selbst: Die Verwendung des ML-Prinzips führt oftmals zu unzulässigen Lösungen; dies bedeutet, dass durch den iterativen Schätzprozess ein oder mehrere Ψ_i, $i = 1, \ldots, p$, der Einzelrestvarianz Ψ kleiner als Null geschätzt werden. Solche Resultate, in der Literatur öfter auch als *Heywood cases* bezeichnet, ergeben keinen Sinn und werden bei der Verwendung statistischer Software gesondert gekennzeichnet und oftmals auf $\Psi_i = 0$ gesetzt. Es besteht jedoch auch die Möglichkeit, dass die unzulässigen Lösungen nicht auf das Maximum-Likelihood-Prinzip, sondern auf eine geringe Stichprobengröße, wenige Variablen oder Ausreißer zurückzuführen sind, vergleiche insbesondere Khattree und Naik (2000, Seite 155 ff.) und die dort angegebenen Referenzen. Generell sind Resultate bei denen unzulässige Fälle auftreten mit Vorsicht zu genießen und sehr genau zu analysieren; es existieren jedoch zahlreiche Beispiele bei denen diese Problematik vernachlässigt werden kann und die Ergebnisse im Allgemeinen gut interpretiert werden können, vergleiche hierzu auch Abschnitt 7.3. Offensichtlich wird der gemittelte Schätzer $\hat{\bar{\Psi}}$ nur äußerst selten nichtpositive Elemente enthalten; seine Sinnhaftigkeit muss jedoch kritisch überprüft werden, wenn auch nur eine Einzelrestvarianz $\hat{\Psi}_i$ ein negatives (bzw. von der Software auf Null gesetztes) Element enthält.

Ein weiterer kritischer Punkt betrifft generell die Betrachtung kleiner Stichproben: Es können sich hier Probleme mit der Annahme der multivariaten Normalverteilung ergeben (Khattree und Naik (2000)); auch ist eine Korrektur des AIC, wie in Abschnitt 3.3.4 erläutert, im Kontext der Faktorenanalyse nicht möglich.

5. Berücksichtigung fehlender Werte

Statistische Verfahren, insbesondere auch im Kontext von Modellselektion und Modellmittelung, gehen meist von der vollständigen Kenntnis aller Beobachtungen einer Datenmatrix D aus. In der Praxis sind fehlende Werte jedoch ein häufig auftretendes Problem, das unterschiedlichste Ursachen besitzen kann: Werden Daten durch Interviews gewonnen, so besteht die Möglichkeit, dass die befragten Personen Antworten verweigern, Fragen nicht beantworten können oder bei Meinungsumfragen ihre Präferenz nicht ausdrücken können. Auch die Interviewer selbst können für das Fehlen von Datenmaterial verantwortlich sein; beispielsweise wenn Fragen übersehen oder unleserlich notiert werden. Bei klinischen Studien, die über einen längeren Zeitraum durchgeführt werden, ist es keine Seltenheit, dass Patienten nicht über die volle Zeitspanne beobachtet werden können oder erst zu einem späteren Zeitpunkt überhaupt in die Studie eintreten. Naturwissenschaftliche und industrielle Experimente benötigen oft den Einsatz von Messgeräten, die unter bestimmten Umständen keine Messungen liefern, so etwa bei sehr geringen Konzentrationen eines Stoffes oder ungewöhnlichen Versuchsbedingungen. Auch in der Biologie und Landwirtschaft kann die Zerstörung von Testfeldern zum Ausfall ganzer Datenblöcke führen.

Um das Auftreten fehlender Beobachtungen im Rahmen statistischer Modellierung und Inferenz zu berücksichtigen, bedarf es insbesondere einer genauen Analyse des Fehlendmechanismus, also der Frage, ob dem Fehlen der Werte eine bestimmte Systematik zugrunde liegt oder nicht. Sei D eine $n \times p$ Datenmatrix und F eine Indikatormatrix derselben Dimension, für die gilt:

$$F_{ij} \;=\; \begin{cases} 1, & \text{falls die Beobachtung } D_{ij} \text{ vollständig ist} \\ 0, & \text{falls die Beobachtung } D_{ij} \text{ fehlt} \end{cases} .$$

Unterscheidet man zwischen den beobachteten Daten D^{obs} und den fehlenden Daten D^{mis}, so lässt sich gemäß Rubin (1976) der Fehlendmechanismus über die bedingte

Verteilung $f(F|D;\xi) = f(F|D^{\text{obs}}, D^{\text{mis}};\xi)$ spezifizieren, wobei ξ die Parametrisierung des durch F charakterisierten Fehlendprozesses beschreibt. Gilt

$$f(F|D;\xi) \;=\; f(F|D^{\text{obs}}, D^{\text{mis}};\xi) = f(F|\xi) \qquad \forall D\,,$$

so wird der Fehlendmechanismus als *missing completely at random*, kurz *MCAR*, bezeichnet. Die Werte fehlen unter diesen Voraussetzungen rein zufällig; ihre Fehlwahrscheinlichkeit ist damit unabhängig von allen beobachtbaren wie auch unbeobachtbaren Faktoren. Hängt die Fehlwahrscheinlichkeit dagegen nur von beobachteten Werten ab, wodurch für die bedingte Dichte $f(F|D;\xi)$ offensichtlich

$$f(F|D;\xi) \;=\; f(F|D^{\text{obs}}, D^{\text{mis}};\xi) = f(F|D^{\text{obs}};\xi) \qquad \forall D^{\text{mis}}$$

gilt, so wird der Fehlendmechanismus als *missing at random (MAR)* bezeichnet. Liegt weder ein MCAR- noch ein MAR-Prozess vor, so wird der Fehlendmechanismus als *missing not at random (MNAR)* bzw. *nonignorable* bezeichnet. Die fehlenden Werte hängen dabei explizit von unbeobachtbaren Quantitäten ab und für die bedingte Dichte lässt sich keine Vereinfachung vornehmen:

$$f(F|D;\xi) \;=\; f(F|D^{\text{obs}}, D^{\text{mis}};\xi)\,.$$

In der Literatur existieren eine Vielzahl an Vorschlägen, wie mit der Problematik fehlender Werte umgegangen werden kann; detaillierte Erläuterungen zu den gängigsten Verfahren finden sich unter anderem bei Little und Rubin (2002), Molenberghs und Kenward (2007) sowie Rao et al. (2008, Kapitel 8) und den dort angegebenen Referenzen; inwieweit diese Verfahren in statistischer Software implementiert sind, erläutern Horton und Kleinman (2007). Grundsätzlich lassen sich nahezu alle Prozeduren in die folgenden drei Bereiche untergliedern:

1. **Betrachtung der vollständigen Fälle.** Hierbei wird jedwede statistische Methodik auf den Subdatensatz D_*^c, der nur die vollständigen Beobachtungen $(y_i^{\text{obs}}, \mathbf{x}_i^{\text{obs}}) \in D_*$ enthält, angewandt. Dieses Vorgehen wird gemeinhin auch als *Complete Case Analyse* bzw. *listwise deletion* bezeichnet und bedeutet keinen zusätzlichen Arbeitsaufwand für die statistische Analyse. Problematisch erscheint hierbei jedoch die Größe des Stichprobenumfangs, der sich insbesondere bei einer großen Anzahl an Variablen potentiell stark verringern kann, da bereits ein fehlender Wert je Beobachtungszeile $D_i = (y_i, \mathbf{x}_i)$ genügt, damit eine komplette Zeile des Datensatzes

entfernt werden muss. So erläutern etwa Little und Rubin (2002), dass bei einer Fehlendwahrscheinlichkeit von 10% je Wert D_{ij} und 20 Variablen für eine Analyse nur noch etwa 13% der Fälle zur Verfügung stehen.
Ist der Fehlendmechanismus MCAR, so sind konsistente Schätzer für den Datensatz der vollständigen Fälle D_*^c weiterhin konsistent, da effektiv ein repräsentativer Subdatensatz der Stichprobe vorliegt. Bei MAR-Szenarien gilt diese Aussage nicht mehr ohne Einschränkung; in der Regel muss mit verzerrten Schätzungen gerechnet werden. Eine Ausnahme bildet hierbei jedoch der interessante Spezialfall der Schätzung der Regressionsparameter bei linearen und logistischen Regressionsmodellen. Hängt hier das Fehlen der Werte in den Kovariablen nicht vom Response ab, so sind die Schätzungen weiterhin konsistent, vergleiche hierzu auch Little (1992) und Vach (1994). Bei MNAR können meist keine allgemeingültigen Aussagen bezüglich der Konsistenz getroffen werden. Zu erwähnen sind jedoch auch hier die Auswirkungen auf die Parameterschätzungen in der linearen Regression: Hängt das Fehlen eines Wertes in den Kovariablen nur von dieser Variable selbst ab, was einen MNAR-Fehlendmechanismus impliziert, so sind die Schätzungen der Regressionsparameter weiterhin konsistent, da der Wertebereich der Kovariablen zwar eingeschränkt ist, die Werte selbst jedoch repräsentativ bleiben; vergleiche auch Fieger (2001).
Aus den oben angeführten Überlegungen ist ferner klar, dass aufgrund der Verringerung des Stichprobenumfangs jeder MSE-konsistente Schätzer bei MCAR- und MAR-Szenarien für den Subdatensatz der vollständigen Fälle weniger effizient ist, als derselbe Schätzer für den vollständigen Datensatz ohne fehlende Werte. Bei MNAR-Fehlendmechanismen kann darüber keine Aussage getroffen werden.

2. **Imputationsansätze.** Anstatt fehlende Beobachtungen zu verwerfen, liegt es nahe, diese durch „sinnvolle“ Werte zu ersetzen und die statistische Methodik auf den aufgefüllten Datensatz anzuwenden. Ein solches Vorgehen wird als *Imputation* bezeichnet. In der Literatur existieren eine Vielzahl an Vorschlägen wie fehlende Werte ersetzt werden können; einige dieser Methoden können als ad-hoc bezeichnet werden, da sie vor allem pragmatisch motiviert sind und eine unkomplizierte Anwendung garantieren sollen: Dazu gehören unter anderem die Mittelwertsimputation, bei der fehlende Werte schlicht durch das arithmetische Mittel bzw. den Median oder Modus der entsprechenden Variablen ersetzt werden sowie LOCF (Last Observation Carried Forward), das in longitudinalen Studien fehlende Werte durch den letzten beobachteten Wert der Untersuchungseinheit auffüllt. Solche

Prozeduren führen in der Regel zu verzerrten Schätzungen und unterschätzen die Varianz, vergleiche beispielsweise Cook, Zeng und Yi (2004).
Andere Methoden setzen darin an, die Struktur der Beobachtungen untereinander bzw. die Abhängigkeiten zwischen den Variablen auszunutzen, um Werte zu imputieren: Bei der k-Nächste-Nachbarn (kNN) Methode (Chen und Shao (2000), Gottardo (2008)) werden über die euklidische Distanz oder die Stichprobenkorrelation die k nächsten Nachbarn der Beobachtung mit dem fehlenden Wert bestimmt. Diese Nachbarn dürfen keine fehlenden Werte enthalten. Das (gewichtete) Mittel der Nachbarn dient dann als Imputation für den fehlenden Wert. Regressionsimputationen (Little und Rubin (2002)) versuchen dagegen die Abhängigkeitsstruktur der Variablen durch Regressionsmodelle zu erfassen. Dabei werden alle vollständigen Fälle, also die Complete Cases, zum anpassen eines Regressionsmodells verwendet, bei dem die Variable, die den fehlenden Wert enthält, als Response betrachtet wird und alle anderen Variablen (oder eine Untermenge davon) als Kovariablen. Auf Basis dieses Modells wird dann der Vorhersagewert der Regression für die interessierende Beobachtung zum Imputieren verwendet. Dem Regressionsmodell können auch stochastische Fehlerterme hinzugefügt und/oder statt den geschätzten Regressionsparametern auch Realisationen ihrer geschätzten (posteriori-) Verteilung verwendet werden, vergleiche Schafer (1997) sowie Abschnitt 5.1.2 für eine ausführliche Diskussion dieser Methoden.
Häufig verwendet werden auch verteilungsbasierte Imputationen. Dabei wird ein multivariates Verteilungsmodell für die Daten spezifiziert und anschließend aus der prädiktiven a-posteriori-Verteilung der fehlenden Daten gegeben die beobachteten Daten

$$p(D^{\text{mis}}|D^{\text{obs}}) \;=\; \int p(D^{\text{mis}}|D^{\text{obs}};\theta)\; p(\theta|D^{\text{obs}})\; d\theta \tag{5.1}$$

die notwendigen Imputationswerte gezogen. Dabei beschreibt $p(D^{\text{mis}}|D^{\text{obs}};\theta)$ die bedingte prädiktive Verteilung der fehlenden Werte, gegeben die beobachteten Werte und Parametervektor θ, was ausdrücklich eine stochastische Imputation impliziert und im Spezialfall auch als die oben angedeutete stochastische Regressionsimputation aufgefasst werden kann. Entscheidender Unterschied ist jedoch die zusätzliche Berücksichtigung der posteriori-Verteilung $p(\theta|D^{\text{obs}})$ der Modellparameter. Diese wird aus einer, möglicherweise nicht-informativen, a-priori-Verteilung $p(\theta)$ sowie der Likelihoodfunktion $\mathcal{L}(\theta|D)$ bestimmt und berücksichtigt die Unsicherheit bezüglich (der Schätzung von) θ. Bei einem solchen Vorgehen ergeben sich unter anderem die folgenden, zu diskutierenden Punkte:

- Es stellt sich die Frage, welches multivariate Verteilungsmodell konkret für die Daten spezifiziert werden sollte. In der Regel wird die multivariate Normalverteilung verwendet. Auch wenn in den wenigsten Datensätzen von ausschließlich metrischen Variablen ausgegangen werden kann, so finden sich dennoch zahlreiche Beispiele in der Literatur, die bestätigen, dass die Verwendung der multivariaten Normalverteilung als Imputationsmodell in der Regel zu ähnlich guten Ergebnissen führt wie komplexere Alternativen, vergleiche hierzu auch Schafer (1997), King et al. (2001) und die dort angegebenen Referenzen. Alternativ ergibt sich auch die Möglichkeit, die Daten geeignet zu transformieren bevor die Verteilung festgelegt wird und erst nach dem Imputationsschritt zu retransformieren.
- Die analytische Berechnung von (5.1) ist – abgesehen von Spezialfällen – im Allgemeinen nicht möglich. Dies liegt insbesondere an der komplexen und aufwändigen Berechnung der Likelihood $\mathcal{L}(\theta|D)$, vergleiche hierzu auch den weiter unten angeführten Punkt 3., und damit der Ziehung der Modellparameter aus der posteriori-Verteilung $p(\theta|D^{\text{obs}})$. Es existieren jedoch mehrere Algorithmen mit denen sich die Ziehung aus (5.1) sehr gut simulieren lässt: An erster Stelle muss dabei sicherlich der weit verbreitete Imputation-Posterior-Algorithmus (IP, Schafer (1997)) genannt werden. Dieser benötigt in der Implementation MCMC-Verfahren, weswegen in einigen Situationen mit einem erheblichen Zeitaufwand gerechnet werden muss. In den letzten Jahren werden deshalb zunehmend schnellere Verfahren benötigt und verwendet, so etwa ein Bootstrap-Ansatz von Honaker, King und Blackwell (2008), vergleiche auch Abschnitt 5.1.2. Prinzipiell kann auch die Strategie verfolgt werden, anstatt einer multivariaten Verteilung für die komplette Datenmatrix, nur die bedingten Verteilungen für jede einzelne Variable zu spezifizieren und im Stile eines Gibbs-Samplers iterativ aus diesen Randverteilungen zu ziehen um damit Imputationen zu erhalten, die als Ziehungen aus der gemeinsamen, multivariaten (theoretischen) Verteilung aufgefasst werden können (sofern diese existiert). Ein Überblick für ein solches Vorgehen, das häufig auch als *Fully Conditional Specification* bezeichnet wird, findet sich unter anderem bei Drechsler und Rässler (2008).
- Ungeachtet der oben vorgestellten Möglichkeiten der Imputation, ergibt sich die Problematik der Unsicherheit bezüglich des Imputationsmodells und da-

mit auch der Imputationen selbst, die implizit von der Stichprobe abhängen. Verteilungsbasierte Ansätze ermöglichen die Berücksichtigung dieser Unsicherheit in der Inferenz. Hierfür werden M Imputationen aus der prädiktiven a-posteriori-Verteilung (5.1) gezogen. Daraus ergeben sich M neue Datensätze $D^{(m)}$, $m = 1, \ldots, M$. Der resultierende Schätzer

$$\hat{\theta}^M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)} \tag{5.2}$$

ist damit das Mittel der Schätzer $\hat{\theta}^{(m)}$ aus den imputierten Datensätzen $D^{(m)}$. Seine Varianz erhält man zu

$$\mathbb{V}\mathrm{ar}(\hat{\theta}^M) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{V}\mathrm{ar}(\hat{\theta}^{(m)}) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \hat{\theta}^M)^2 , \tag{5.3}$$

wobei der erste Term die Varianz innerhalb der imputierten Datensätze beschreibt und der zweite Term die Varianzen zwischen den Datensätzen. Für die theoretische Rechtfertigung der *multiplen Imputation* (Rubin (1978), Rubin (1996)) sei an dieser Stelle auf Little und Rubin (2002) und Molenberghs und Kenward (2007) verwiesen.

Aus den oben angeführten Überlegungen wird klar, dass die Auswahl geeigneter Imputationsmethoden größte Sorgfalt erfordert und insofern auch ein gewisses Gefahrenpotential beherbergt. Dempster und Rubin (1983) kommentieren dies wie folgt:

> *„The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all. It is dangerous because it lumps together situations where the problem is sufficiently minor that it can be handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases“*

3. **Explizite Berücksichtigung der fehlenden Werte in der Inferenz.** Anstatt die fehlenden Beobachtungen schlicht zu verwerfen oder durch Imputationen zu ersetzen, besteht auch die Möglichkeit in der Inferenz explizit auf diesen Umstand einzugehen:

   (i) *Gewichtungsansätze.* Das Konzept von Gewichtungsansätzen besteht darin, ein Modell für die Wahrscheinlichkeit einer vollständigen Beobachtung, $\pi_i = P(\mathbf{F}_i = \mathbf{1}_p)$, $\mathbf{F}_i = (F_{i1}, \ldots, F_{ip})$, für alle Beobachtungen D_i, $i = 1, \ldots, n$, anzupassen und die daraus gewonnenen geschätzten inversen Wahrscheinlichkeiten $1/\hat{\pi}_i$ im weiteren Inferenzprozess als Gewichte für die vollständigen Fälle

zu verwenden. Dieses Konzept basiert auf der Idee des Horvitz-Thompson-Schätzers (Horvitz und Thompson (1952)). Es ermöglicht fehlende Beobachtungen explizit im Inferenzprozess zu berücksichtigen und führt unter einem MAR- oder MCAR-Fehlendmechanismus und angemessenen Regularitätsvoraussetzungen zu konsistenten Schätzungen; aufgrund des Verwerfens der unvollständigen Beobachtungen muss jedoch – ähnlich dem Vorgehen bei einer Complete Case Analyse – mit einem Effizienzverlust gerechnet werden, vergleiche hierzu beispielsweise Clayton et al. (1998). Die Wahrscheinlichkeiten π_i können prinzipiell über logistische Regressionsmodelle bzw. generalisierte additive Modelle geschätzt werden, siehe auch Hens, Aerts und Molenberghs (2006) sowie Abschnitt 5.1.1 und Kapitel 6. Wie sich eine Misspezifikation eines solchen Modells bzw. die Unsicherheit in der Wahl geeigneter Kovariablen bei logistischen Regressionsmodellen und Glättungsparameter bei additiven Modellen auswirkt, ist bisher weitgehend ungeklärt. Kapitel 6 zeigt, dass zumindest im Kontext von Modellselektion und Modellmittelung mit einer Unterschätzung der Varianz gerechnet werden muss. Um die Effizienz von Schätzern auf der Basis der oben beschriebenen Gewichtungsansätze zu verbessern, existieren Verfahren, die auch die nicht-vollständigen Fälle explizit berücksichtigen. Allgemein sind solche Ansätze in der Literatur unter der Bezeichnung *double robustness* bekannt. Eine Einführung hierzu sowie zahlreiche Verweise auf die Originalliteratur findet man unter anderem bei Molenberghs und Kenward (2007, Kapitel 10). Die Umsetzung dieser Verfahren an konkreten Datenbeispielen ist aufgrund ihrer Komplexität bisher jedoch nur sehr beschränkt möglich. Ferner erfordern sie die Spezifikation eines zusätzlichen Fehlendmodells, wodurch weitere Unsicherheit entsteht.

(ii) *Modellbasiertes Vorgehen.* Prinzipiell gilt unter der MAR-Annahme, dass die Maximierung der beobachteten Likelihood

$$\mathcal{L}(\theta|D^{\text{obs}}) \;=\; \int f(D^{\text{obs}}, D^{\text{mis}}; \theta)\; dD^{\text{mis}} \tag{5.4}$$

bei einer korrekten Spezifizierung des Datenmodells zu korrekten ML-Schätzungen für θ führt, vergleiche in etwa Little und Rubin (2002) oder Fieger (2001). Um die ML-Schätzungen für ein konkretes Datenbeispiel zu bestimmen werden für die Maximierung jedoch meist der Newton-Raphson-Algorithmus bzw. Scoring-Methoden benötigt, die beide auf die Matrix der zweiten Ableitungen der Likelihood-Funktion angewiesen sind. Die Einträge

dieser Matrix sind für nicht monotone Fehlendmuster in den Daten in der Regel komplexe Funktionen von θ, weswegen die Berechnung der ML-Schätzung sehr aufwändig bzw. nicht möglich ist. Für die Bestimmung der Schätzungen werden daher üblicherweise iterative Methoden verwendet; der Gold-Standard ist dabei der EM-Algorithmus (Dempster, Laird und Rubin (1977)), der ohne die Matrix der zweiten Ableitungen auskommt; eine detaillierte Beschreibung hierfür findet sich unter anderem bei Schafer (1997). Generell lässt sich ein modellbasiertes Vorgehen auch bayesianisch motivieren und umsetzen; wie in Punkt 2. bereits angedeutet, ist es hierfür nur nötig der Likelihood $\mathcal{L}(\theta|D^{\text{obs}})$ eine priori-Verteilung $p(\theta)$ hinzuzufügen und jedwede Inferenz auf die entsprechende posteriori-Verteilung zu stützen. Insbesondere bei Regressionsproblemen, bei denen stets die bedingte Verteilung des Response von Interesse ist, führt die für fehlende Werte in den Kovariablen komplexe Likelihood jedoch meist dazu, dass die marginalen posteriori-Verteilungen nicht mehr bestimmbar sind und aufwändig approximiert werden müssen, vergleiche hierzu insbesondere Little (1992).

Im Allgemeinen lassen sich verteilungsbasierte Ansätze, die, wie oben beschreiben, explizit die Problematik fehlender Daten miteinbeziehen, natürlich nur dann durchführen, wenn eine Likelihood spezifiziert werden kann. Bei Quasi-Likelihood-Ansätzen, wie beispielsweise bei marginalen Modellen (GEE, Liang und Zeger (1986)) häufig verwendet, ist ein solches Vorgehen nicht möglich. Ausführliche Betrachtungen zu dieser Thematik findet man bei Kastner (2001) und Molenberghs und Kenward (2007).

In den folgenden Abschnitten 5.1 und 5.2 werden einige der oben angedeuteten Kernkonzepte noch einmal ausführlich erörtert, im Kontext von Modellselektion und Modellmittelung verwendet und hierfür gegebenenfalls auch erweitert.

5.1 Modellselektion bei fehlenden Daten

Wie in Kapitel 3 ausführlich beschrieben, existieren eine Vielzahl an Konzepten zur Konstruktion von Modellwahlkriterien. Die meisten dieser Ansätze sind durch gewisse Optimalitätsüberlegungen motiviert; so wird beispielsweise bei Kreuzvalidierungsansätzen der erwartete Vorhersagefehler minimiert, bei informationstheoretischen Kriterien wie dem AIC die Kullback-Leibler-Distanz, bei bayesianischen Ansätzen steht die

Maximierung der posteriori-Wahrscheinlichkeit eines Modells im Vordergrund und die MDL-Methodik sucht nach einem kürzesten Code um Daten zu ver- und entschlüsseln. Viele Kriterien und Verfahren besitzen darüber hinaus noch Optimalitätseigenschaften wie Effizienz bzw. Konsistenz, die in Abschnitt 3.6 ausführlich beschrieben werden. All diese Optimalitätsbetrachtungen gehen von vollständigen Beobachtungen aus. Es ist klar, dass insbesondere für nicht-MCAR Fehlendmechanismen in Frage gestellt werden muss, wie sehr die Konstruktionskonzepte bei fehlenden Daten von den beschriebenen Verfahren noch erfasst werden können und ob es nötig ist, Kriterien gegebenenfalls zu adjustieren. Im folgenden Abschnitt 5.1.1 soll exemplarisch erläutert werden, wie die Modifikation des AIC für fehlende Daten ausschauen kann. Die Idee ist der Einbezug eines Gewichtungsansatzes, wie oben rudimentär erläutert, und geht auf Hens, Aerts und Molenberghs (2006) zurück. In Abschnitt 5.1.2 wird die Möglichkeit der Verwendung von Imputationen diskutiert. Die Auswahl der Imputationsmethoden steht dabei repräsentativ für verschiedenste Herangehensweisen innerhalb der oben angedeuteten Ansätze und orientiert sich dabei primär an einer praktikablen Umsetzbarkeit für reelle Datenprobleme; zusätzlich wird das Konzept der Regressionsimputation verallgemeinert und in Form eines rekursiven Algorithmus vorgestellt. Um die Unsicherheit, die durch den Selektionsschritt verursacht wird, zu berücksichtigen, werden in Abschnitt 5.2 die vorgestellten Methoden adjustiert und erweitert.

5.1.1 Gewichtetes Akaike Kriterium ($\mathbf{AIC}_W$)

Das für fehlende Daten adjustierte, gewichtete Informationskriterium AIC_W (Hens, Aerts und Molenberghs (2006)) basiert auf dem Konzept des Horvitz-Thompson-Schätzers (Horvitz und Thompson (1952)). Jede Beobachtung wird dabei mit ihrer inversen Auswahlwahrscheinlichkeit gewichtet. Sei

$$\delta_i \;=\; \begin{cases} 1, & \text{für eine vollständige } i\text{-te Beobachtung } D_i = (y_i, \mathbf{x}_i) \\ 0, & \text{für eine unvollständige } i\text{-te Beobachtung } D_i = (y_i, \mathbf{x}_i) \end{cases}$$

sowie $\pi_i = P(\delta_i = 1)$ die zugehörige Wahrscheinlichkeit einer vollständigen Beobachtung, dann können die folgenden Gewichte

$$w_i \;=\; \frac{\delta_i}{\pi_i} \tag{5.5}$$

definiert werden. Man erhält das gewichtete Informationskriterium AIC_W zu

$$\mathrm{AIC}_W = -2\sum_{i=1}^{n} w_i \cdot \log f(y_i|\hat{\theta}_W) + 2K\,, \tag{5.6}$$

wobei $\hat{\theta}_W$ den gewichteten ML-Schätzer bezeichnet, also den Schätzer der die logarithmierte Likelihood $\log f_w(y|\theta) = \sum_i w_i \log f(y_i|\theta)$ auf Basis der Gewichte (5.5) maximiert. Für das lineare Regressionsmodell $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, $\mu = X\beta$, erhält man das Kriterium offensichtlich zu

$$\begin{aligned} AIC_{W,LM} &= -2\sum_{i=1}^{n} w_i \cdot \log f(y_i|\hat{\mu}_W, \hat{\sigma}^2_W) + 2K \\ &\propto \sum_{i=1}^{n} w_i \cdot \log\left(\frac{\sum_{i=1}^{n} w_i \hat{\epsilon}_i^2}{\sum_{i=1}^{n} w_i}\right) + 2K\,, \end{aligned} \tag{5.7}$$

wobei $\hat{\epsilon}_i$ die geschätzten Residuen des gefitteten Modells bezeichnen. Die Gewichte w_i sind in der Regel unbekannt und müssen, vorzugsweise nonparametrisch, beispielsweise über ein generalisiertes additives Modell, geschätzt werden.

Die Rechtfertigung des Kriteriums liegt in der Feststellung, dass die Verwendung der gewichteten Kullback-Leibler-Distanz bei fehlenden Daten und unter Verwendung der Gewichte (5.5) genau der Kullback-Leibler-Distanz (3.22) für den vollen Datensatz entspricht, also

$$\begin{aligned} \mathrm{KL}_W(f(y), f(y;\theta)) &= \mathbb{E}\left\{\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \log[f(y_i)/f(y_i;\theta)]\right\} \\ &= \sum_{i=1}^{n} \mathbb{E}\{\log[f(y_i)/f(y_i;\theta)]\} \end{aligned}$$

gilt (vgl. Hens, Aerts und Molenberghs (2006, Seite 2507)) und das gewichtete Kriterium AIC_W als Schätzer der KL-Distanz insofern als sinnvoll erscheint. Diese Überlegungen können jedoch nicht herangezogen werden, um für das gewichtete Akaike-Kriterium die asymptotische Effizienz nachzuweisen. Obgleich die Autoren mit diesem Gewichtungsansatz ein gängiges statistisches Prozedere für das Auftreten fehlender Daten aufgreifen, so existieren dennoch einige kritische Punkte, die in der konkreten Anwendung Probleme bereiten können:

- In der Regel sind die Gewichte (5.5) unbekannt und müssen geschätzt werden. Hens, Aerts und Molenberghs (2006) schlagen vor, dies nonparametrisch, bevorzugt über generalisierte additive Modelle, zu versuchen, da zu erwarten ist, dass $\sum_{i=1}^{n} \hat{w}_i = \sum_{i=1}^{n} \delta_i / \hat{\pi}_i \approx n$. Dies ist korrekt, allerdings entsteht dadurch ein weiterer Modellselektionsschritt, der zusätzliche Unsicherheit mit sich bringt: In der nonparametrischen Regression ist es nötig, sich für einen Glättungsparameter, beispielsweise über das generalisierte Kreuzvalidierungskriterium GCV, zu entscheiden, wodurch – wie in Kapitel 4 ausführlich erläutert – eine Unsicherheit entsteht, die korrekterweise in der Inferenz berücksichtigt werden müsste. Die Simulationsergebnisse aus Kapitel 6 bestätigen, dass dieser Effekt keineswegs vernachlässigbar ist. Auch ist nicht klar, welche Kovariablen in einem konkreten Datenbeispiel (bei dem in vielen Kovariablen Werte fehlen und bei dem der Fehlendmechanismus nicht eindeutig geklärt werden kann) potentiell in das entsprechende GAM aufgenommen werden sollten; Liegen n und p in einem ähnlichen Größenbereich ist es allein aus technischer Sicht oft nicht möglich ein generalisiertes additives Modell zu fitten. Eine willkürliche Auswahl der Kovariablen kann problematisch sein; dies verdeutlicht eine Simulation von Schomaker (2006, Abschnitt 4.2.3), der zeigt, dass sich die Wahl der Kovariablen in einem GAM auf die Resultate des AIC_W entscheidend auswirken kann.

- Hens, Aerts und Molenberghs (2006) verweisen darauf, dass das Verhalten des AIC_W am besten über Simulationsstudien erkundet werden kann. Die Autoren studieren darauf aufbauend verschiedene Modellselektionssituationen im Kontext linearer Modelle, designbasierter Stichproben und nonparametrischer Regression, bei denen allesamt ein MAR-Fehlendmechanismus angenommen wird. Sie kommen in allen betrachteten Situationen zum Schluss, dass der Gewichtungsansatz in der Regel bessere Ergebnisse liefert als eine Complete Case Analyse. Diese Aussage stellt sich bei einer detaillierten Betrachtung der Resultate, beispielsweise Hens, Aerts und Molenberghs (2006, Tabelle 2), als falsch heraus. Ähnlich dem Vorgehen von Claeskens und Consentino (2008) werden überangepasste Modelle als korrekt gewertet und gehen folglich mit in die Interpretation der Resultate ein. Es stellt sich heraus, dass die – durchweg guten – Ergebnisse vor allem diesem Umstand zu verdanken sind und deshalb mit Vorsicht genossen werden müssen. Die Simulationsstudien von Schomaker (2006) konstatieren dem AIC_W einen besonders stark ausgeprägten Hang zur Wahl überangepasster Modelle, was bei der Interpretation konkreter Anwendungsbeispiele zusätzlich berücksichtigt werden sollte.

5.1.2 Selektion nach Imputation

Um die Problematik fehlender Daten in der Modellselektion zu berücksichtigen, liegt es nahe, Werte auf Basis eines der oben beschrieben Verfahren zu imputieren und den Selektionsschritt auf dem aufgefüllten Datensatz D^{imp} durchzuführen. Für eine Menge an Kandidatenmodellen $\mathcal{M} = \{M_1 \ldots, M_k\}$ wird hierbei also dasjenige Modell $M_\kappa \in \mathcal{M}$ gewählt, das ein Kriterium oder Verfahren Γ, wie in Kapitel 3 beschrieben, auf D^{imp} minimiert (bzw. maximiert) und somit zu der Wahl des Parameterschätzers

$$\hat{\theta}_{\text{imp}} \;=\; \hat{\theta}_\kappa : \arg \min_{M_\kappa \in \mathcal{M}} \{\Gamma(M_\kappa; \hat{\theta}_\kappa) | D^{\text{imp}}\}\,, \qquad \kappa = 1, \ldots, k\,, \tag{5.8}$$

führt. Wie zu Beginn dieses Kapitels erläutert, ergibt sich die Möglichkeit, die Unsicherheit bezüglich der Imputation zu berücksichtigen, indem auf Basis der prädiktiven a-posteriori-Verteilung (5.1) mehrfach imputiert wird, also anstelle eines einzigen Datensatzes M imputierte Datensätze $D^{(m)}$, $m = 1, \ldots, M$, erzeugt werden. Dies resultiert in dem gemittelten Selektionschätzer

$$\hat{\theta}^M_{\text{imp}} \;=\; \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)}_{\text{imp}}\,, \tag{5.9}$$

dessen Varianz offensichtlich über

$$\widehat{\mathbb{V}\text{ar}}(\hat{\theta}^M_{\text{imp}}) \;=\; \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbb{V}\text{ar}}(\hat{\theta}^{(m)}_{\text{imp}}) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\theta}^{(m)}_{\text{imp}} - \hat{\theta}^M_{\text{imp}})^2 \tag{5.10}$$

geschätzt werden kann. Den Selektionsschritt auf einem imputierten Datensatz durchzuführen und damit die Schätzer (5.8) bzw. (5.9) zu berechnen und zu interpretieren, ist ein intuitives und gelegentlich verwendetes Verfahren, das jedoch in der Regel in der Literatur nicht genauer untersucht und beschrieben wird und somit eher einen ad-hoc Charakter – ähnlich einer Complete Case Analyse – zu besitzen scheint. Dies wird auch in den Arbeiten von Burton und Altmann (2004) sowie Horton und Switzer (2005) deutlich, die medizinische Artikel bezüglich des Umgangs mit fehlenden Daten bewerten: Burton und Altmann (2004) betrachten 100 Artikel aus dem Jahre 2002, die die Auswertungen verschiedener Krebsstudien darlegen. 81% der Artikel verwenden dabei Datensätze mit fehlenden Werten in ihrer Analyse; nur vier dieser Artikel verwenden Imputationsmethoden um der Problematik fehlender Daten gerecht zu werden, alle weiteren

Artikel ignorieren diesen Sachverhalt bzw. verwenden bewusst Complete Case Methoden. Von den vier Artikeln, die Imputationen verwenden, sind drei Ansätze als ad-hoc zu bewerten, vergleiche auch Horton und Kleinman (2007); nur ein Artikel verwendet multiple Imputationen und rechtfertigt diesen Schritt. Horton und Switzer (2005) bewerten 331 Artikel des *New England Journal of Medicine*, wovon 26 Artikel Datensätze mit fehlenden Werten analysieren. In keinem dieser Artikel wurden fehlende Fälle verworfen, stattdessen werden durchgängig Imputationsmethoden verwendet, wovon 24 allerdings als ad-hoc anzusehen sind. Meist wird hierbei eine Mittelwertsimputation oder LOCF verwendet. Die zwei Artikel, die multiple Imputation verwenden, spezifizieren dieses Vorgehen nicht im Detail.

Im Folgenden sollen drei Imputationsmöglichkeiten beschrieben werden, die sich prinzipiell im Zusammenhang mit statistischer Modellselektion als geeignet ansehen lassen: (i) eine verallgemeinerte Regressionsimputation, (ii) eine kNN-Methode und (iii) einfache und multiple verteilungsbasierte Imputationen auf Basis eines EM-Bootstrap-Ansatzes. Die Eigenschaften der daraus resultierenden Selektionschätzer im Sinne von (5.8) und (5.9) können in der Regel nicht analytisch untersucht werden, weswegen ausführliche Simulationsstudien in Kapitel 6 das Verhalten dieser Schätzer untersuchen und bewerten.

Verallgemeinerte Regressionsimputation

Betrachtet man die Datenmatrix D_* aus (3.1), so lässt sich die Kernidee einer einfachen Regressionsimputation wie folgt veranschaulichen: Für einen fehlenden Wert $x_{ij} \in D_*$ verwendet man die vollständige Datenmatrix D_*^c um das lineare Regressionsmodell

$$X_j = \alpha + \gamma_0 y + \sum_{\substack{l=1 \\ l \neq j}}^{p} \gamma_l X_l + \epsilon\,, \quad \epsilon \sim N(0, \sigma^2 I) \qquad \forall D_*^c \tag{5.11}$$

anzupassen und dann auf Basis der Parameterschätzungen $(\hat{\alpha}, \hat{\gamma}_0, \ldots, \hat{\gamma}_{j-1}, \hat{\gamma}_{j+1}, \ldots, \hat{\gamma}_p)$ einen Wert $\tilde{x}_{ij}$ für x_{ij} via

$$\tilde{x}_{ij} = \hat{\alpha} + \hat{\gamma}_0 y_i + \sum_{\substack{l=1 \\ l \neq j}}^{p} \hat{\gamma}_l x_{il} \tag{5.12}$$

zu imputieren. Dieses Grundkonzept enthält zwei wesentliche Beschränkungen, die seine Anwendung stark limitieren: Zum einen sollte man meist damit rechnen, dass ein Daten-

satz multinomiale, binäre bzw. Zähldaten enthält, wodurch die Verwendung eines linearen Modells als unzureichend angesehen werden muss, zum anderen ist in keinster Weise zu erkennen, wie mit mehreren fehlenden Werten in einer Beobachtung $D_i = (y_i, \mathbf{x}_i)$ umgegangen werden sollte. Die erste Problematik lässt sich leicht umgehen, indem man anstelle eines linearen Regressionsmodells ein generalisiertes lineares Regressionsmodell mit passender Linkfunktion wählt. Dieser Vorschlag ist nicht neu und wird in der Literatur öfters angedeutet, so etwa bei Little und Rubin (2002, Kapitel 12), jedoch selten konkret umgesetzt. Eine Ausnahme bildet die Idee der *chained equations* (van Buuren, Boshuizen und Knook (1999)), eine Art Gibbs-Sampler auf Basis iterativ gefitteter generalisierter Regressionsmodelle, der in dem *R*-Paket „MICE“ implementiert ist und multiple Imputationen ermöglicht. Um jedoch auch die möglicherweise sehr komplexen Zusammenhänge in den Daten passend abzubilden, können auch generalisierte additive Modelle (GAM, vgl. Hastie und Tibsharani (1990)) zur Regressionsimputation herangezogen werden. Dieses Konzept wurde in der einschlägigen Literatur bisher nicht aufgegriffen und soll in der im Folgenden beschriebenen verallgemeinerten Regressionsimputation verwendet werden. Die zweite Problematik lässt sich durch algorithmisches Vorgehen erfassen. Es stellt sich schlicht die Frage, für welche fehlenden Werte die Imputationsmodelle im Stile von (5.11) zuerst gefittet werden sollen und ob die darüber gewonnenen Imputationen zum Fitten der weiteren Imputationsmodelle verwendet werden oder nicht. Der folgende, in Tabelle 5.1 dargestellte, Algorithmus verwendet generalisierte additive Modelle zur Erfassung der Datenstruktur und fittet für jeden fehlenden Wert ein eigenes Imputationsmodell auf Basis des sukzessive aufgefüllten Datensatzes, also unter Einbezug der neu imputierten Werte.

Aus der Beschreibung des Algorithmus ist ersichtlich, dass die Reihenfolge mit der die Werte imputiert werden, sowohl von der Anzahl der fehlenden Werte je Beobachtung (Schritt 2) als auch der Anzahl der fehlenden Werten je Variable (Schritt 3) abhängt. Fehlt mehr als ein Wert für eine Beobachtung, so wird zunächst nur einer dieser Werte durch eine Imputation ersetzt; erst nachdem auch weitere Beobachtungen mit einer gleichen oder größeren Anzahl an fehlenden Werte berücksichtigt wurden, wird in einem der darauffolgenden Schritte der nächste fehlende Wert dieser Beobachtung ersetzt – dies folgt aus der sukzessiven Aktualisierung von D_s in Schritt 4 c).

Generalized **A**dditive **M**odel based **R**ecursive **I**mputation

Schritt 1: Sei s ganzzahlig und beginne mit $s = 1$.

Schritt 2: Bezeichne D_s den Subdatensatz, der nur die Zeilen mit s fehlenden Werten enthält.

Schritt 3: $\mathcal{X}^*$ ist die Menge aller Variablen, die fehlende Werte enthalten. Wähle $X_j \in \mathcal{X}^*$ mit der geringsten Anzahl an fehlenden Werten in D_s.

Schritt 4: Für jeden fehlenden Wert $x_{ij} \in X_j$, $i = 1, \ldots, n$, und $x_{ij} \in D_s$,

a) fitte das GAM,

$$\eta(X_j) \;=\; \alpha + f_0(y) + \sum_{\substack{l=1 \\ l \neq j,\, l \notin \Phi_l}}^{p} f_l(X_l) + \epsilon, \tag{5.13}$$

unter Verwendung der vollständigen Daten D_*^c, wobei $\Phi_l = \{l : x_{il} \text{ fehlend}\}$ und $\eta(\cdot)$ eine passende Linkfunktion ist. Wenn x_{ij} und x_{kj} gleichzeitig fehlen und $i < k$, dann erfolgt die Imputation von x_{ij} vor x_{kj}.

b) imputiere einen neuen Wert $\tilde{x}_{ij}$ für x_{ij} basierend auf dem geschätzten GAM

$$\eta(\tilde{x}_{ij}) \;=\; \hat{\alpha} + \widehat{f}_0(y_i) + \sum_{\substack{l=1 \\ l \neq j,\, l \notin \Phi_l}}^{p} \widehat{f}_l(x_{il}), \tag{5.14}$$

wobei $\hat{\alpha}$, $\hat{f}_o$ und $\hat{f}_l$ Schätzungen von α, f_o und f_l sind.

c) Aktualisiere D_s, D_*^c und $\mathcal{X}^*$ indem die imputierten Werte als vollständige Beobachtungen gewertet werden.

Schritt 5: Wiederhole ab Schritt 3 bis $D_s = \emptyset$.

Schritt 6: Setze s auf $s + 1$.

Schritt 7: Wiederhole ab Schritt 2 bis s die maximale Anzahl an fehlenden Werten in einer Zeile erreicht hat.

Schritt 8: Wiederhole die Schritte 1–7 bis $\mathcal{X}^* = \emptyset$.

Tab. 5.1: Algorithmus für eine verallgemeinerte Regressionsimputation

Für eine detaillierte Illustration des Algorithmus wird im Folgenden ein Beispieldatensatz (Tabelle 5.2) verwendet. Die Variablen y, X_1 und X_2 sind dabei metrisch, X_3 ist eine Zählvariable and X_4 binär. Insgesamt fehlen dem Datensatz fünf Werte: x_{41}, x_{23}, x_{24}, x_{34} und x_{54}.

i	y	X_1	X_2	X_3	X_4
1	1.2	24.2	-8.2	7	0
2	2.3	23.4	-3.2	$*$	$*$
3	2.2	30.0	-0.1	9	$*$
4	2.3	$*$	-4.1	10	1
5	2.6	22.0	-0.8	7	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
100	1.7	26.4	-2.9	7	1

Tab. 5.2: Beispieldaten zur Veranschaulichung des Algorithmus

Gemäß Schritt 1 und 2 wird $s = 1$ gesetzt und folglich all diejenigen Beobachtungen betrachtet, die exakt einen fehlenden Wert aufweisen; in Tabelle 5.2 sind dies genau die dritte, vierte und fünfte Zeile. Da X_1 weniger fehlende Werte besitzt als X_4 (Schritt 3) wird das generalisierte additive Modell $\eta(X_1) = \alpha + f_0(y) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \epsilon$ auf Basis der vollständigen Fälle D_*^c gefittet um – unter Verwendung der Gauß'schen Linkfunktion – eine Imputation für x_{41} wie in Schritt 4 b) beschrieben zu erhalten. Diese Imputation kann dabei der konkrete, eventuell gerundete, Vorhersagewert $\tilde{x}_{ij}$ aus (5.14) sein, oder aus der zugehörigen geschätzten Verteilung (hier: Normalverteilung) gezogen werden. Die vollständigen Fälle werden gemäß 4 c) aktualisiert und enthalten nun den imputierten Wert $\tilde{x}_{41}$. Schritt 5 gibt vor, im Weiteren den Wert x_{34} zu imputieren, anschließend x_{54}. Um dem Sachverhalt Rechnung zu tragen, dass es sich bei X_4 um eine binäre Variable handelt, betrachtet man das GAM $\eta(X_4) = \alpha + f_0(y) + f_1(X_1) + f_2(X_2) + f_3(X_3) + \epsilon$ mit logit-Link; die Imputation $\tilde{x}_{34}$ für x_{34} erhält man weiterhin als Vorhersage dieser Regression. Analog erfolgt die Imputation von x_{54}. Da x_{41}, x_{34} und x_{54} imputiert worden sind, reduziert sich D_1 auf die leere Menge, $D_s = D_1 = \emptyset$. Daher wird infolge von Schritt 6 s auf $s = 2$ gesetzt und damit all diejenigen Beobachtungen herangezogen, die exakt zwei fehlende Werte aufweisen. Dies betrifft im vorliegenden Beispiel ausschließlich die zweite Zeile. Um einen imputierten Wert für x_{23} zu erhalten, wird – unter Verwendung der log-Linkfunktion aufgrund der Zähldaten – das Modell $\eta(X_3) = \alpha + f_0(y) + f_1(X_1) + f_2(X_2) + \epsilon$ angepasst. Die Variable X_4 wird dabei nicht

verwendet, da x_{24} fehlt. Aus der Vorhersage dieses Modells (dies ist erneut entweder der exakte vorausgesagt Wert $\tilde{x}_{23}$ oder ein zufällig gezogener Wert aus der entsprechenden $Po(\tilde{x}_{23})$-Verteilung) wird erneut eine Imputation gewonnen. Die ständige Aktualisierung von D_s führt an dieser Stelle dazu, dass D_2 leer ist, da der einzige fehlende Wert, x_{24}, nun Teil der Menge D_1 ist. Da s die maximale Anzahl an fehlenden Werten in einer Zeile ($s = 2$) erreicht hat, wird s erneut auf $s = 1$ gesetzt (Schritt 7 und 8) und ein generalisiertes additives Modell mit y, X_1, X_2, X_3 als Kovariablen und der logit-Linkfunktion verwendet, um die letzte Imputation zu erhalten. Der Datensatz ist damit vollständig aufgefüllt.

Die Verwendung des Algorithmus kann als verteilungsbasierte Imputation interpretiert werden, da die GAM-basierte Imputation implizit die bedingte prädiktive Verteilung der fehlenden Werte $p(D^{\text{mis}}|D^{\text{obs}};\theta)$ modelliert. Hierbei wird $\theta = (\alpha, \gamma)$ als fest angenommen und daher die Unsicherheit bei der Schätzung der Regressionsparameter nicht berücksichtigt. Diesem Umstand kann auf verschiedene Art und Weise Rechnung getragen werden: Gemäß (5.1) kann unter Berücksichtigung der – schwer zu bestimmenden – posteriori-Verteilung $p(\theta|D^{\text{obs}})$ die prädiktive a-posteriori-Verteilung $p(D^{\text{mis}}|D^{\text{obs}})$ bestimmt werden. Daraus lassen sich prinzipiell (multiple) Imputationen gewinnen. Praktisch bedeutet dies, dass nicht explizit die geschätzten Regressionsparameter sowie die geschätzte Varianz zur Imputation verwendet werden, sondern zufällige Ziehungen aus deren (posteriori-) Verteilung. Auch kann dem Regressionsmodell aus Schritt 4 b) ein stochastischer Fehlerterm hinzugefügt werden. Einige Beispiele wie sich dies realisieren lässt, finden sich bei Schafer (1997). Generell ist ein solches Vorgehen bei einer Erweiterung der Regressionsimputation auf additive Modelle schwer zu realisieren. Werden in dem vorgestellten Algorithmus jedoch anstelle generalisierter additiver Modelle gewöhnliche GLMs verwendet, so ist eine Umsetzung möglich. In Anlehnung an Schomaker, Wan und Heumann (2010) wird der oben vorgestellte Algorithmus als *Generalized Additive Model based Recursive Imputation* (GAMRI) bezeichnet, derselbe Algorithmus, der nur generalisierte lineare Regressionsmodelle verwendet, als *Generalized Linear Model based Recursive Imputation* (GLMRI).

Die Reihenfolge in der die Werte imputiert werden, die Wahl ob und wann die vollständigen Daten D_*^c aktualisiert werden und auf welche Art und Weise dem Imputationsmodell stochastische Komponenten hinzugefügt werden eröffnet viele Möglichkeiten der Konstruktion alternativer Algorithmen im Stil von GAMRI. Jede Wahl ist in gewisser Weise als ad-hoc anzusehen; dennoch liegt diesem Prozedere ein Grundprinzip zugrunde: die Verwendung der Abhängigkeitsstruktur der Variablen eines Datensatzes, um Impu-

tationen zu gewinnen und die für die Modellselektion und Modellmittelung interessanten Effekte zu erhalten.

Nächste Nachbarn Imputation

Die Kernidee einer Nächste-Nachbarn-Imputation besteht darin, die fehlenden Werte der Datenmatrix $x_{ij} \in D_*$ durch beobachtete Werte einer anderen, „ähnlichen“ Untersuchungseinheit $x_{lj} \in D_*$ zu ersetzen. Dieses sehr allgemeine Prinzip wird in der Literatur aus den unterschiedlichsten Blickwinkeln betrachtet, analytische Betrachtungen sind jedoch häufig restringiert auf den Spezialfall zweier stetiger Variablen von denen die eine vollständige beobachtet wurde und die andere nicht, vergleiche in etwa Chen und Shao (2000) sowie Nittner (2003). Generell ist allen Ansätzen jedoch gemein, dass sie für eine gegebene Metrik $d(\cdot,\cdot)$ den Abstand zwischen der Beobachtung mit dem fehlenden Wert und allen anderen Beobachtungen die keinen fehlenden Wert aufweisen berechnen und denjenigen Wert x_{lj} imputieren, der der Beobachtung $D_l = (y_l, \mathbf{x}_l)$ angehört, die den Abstand zu $D_i = (y_i, \mathbf{x}_i)$ auf Basis von $d(\cdot,\cdot)$ minimiert, $i \neq l$; diese Beobachtung wird auch *nächster Nachbar* genannt. Die Metrik ist dabei häufig die euklidische Distanz, angewandt auf die Beobachtungsvektoren $D_{i_{/j}} = (y_i, x_{i1}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots, x_{ip})$ und $D_{l_{/j}} = (y_l, x_{l1}, \ldots, x_{l(j-1)}, x_{l(j+1)}, \ldots, x_{lp})$ des zugehörigen Subdatensatzes $D_{/j} = \{y, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p\}$. Alternativ können auch die k nächsten Nachbarn für die Imputation verwendet werden: Hierbei werden die k Beobachtungen betrachtet, die bezüglich der Metrik $d(\cdot,\cdot)$ den geringsten Abstand zur Beobachtung $D_{i_{/j}}$ haben und der (gewichtete) Mittelwert dieser k Werte wird dann zur Imputation verwendet. Formell bedeutet dies, dass für einen fehlenden Wert $x_{ij} \in D_*$ die Menge der k nächsten Nachbarn,

$$D^{\mathrm{kNN}} \;=\; \{x_{lj} |\, l = \arg \operatorname*{kmin}_{l \,\neq\, i,\, l \notin \Phi_l} d(D_{i_{/j}}, D_{l_{/j}})\}\,,$$

betrachtet wird, wobei kmin eine Funktion bezeichnet, die die k kleinsten ihrer Argumente zurückgibt und $\Phi_l = \{l : D_{l_{/j}} \text{ enthält fehlende Werte}\}$. Das arithmetische Mittel aller $D_l^{\mathrm{kNN}} \in D^{\mathrm{kNN}}$, $l = 1, \ldots, k$, liefert den Imputationswert

$$\tilde{x}_{ij} \;=\; \frac{1}{k} \sum_{l=1}^{k} D_l^{\mathrm{kNN}}$$

für x_{ij}. Dieses Prinzip ist in dem R-Paket „EMV“ (Estimation of Missing Values, Gottardo (2008)) implementiert und wird im Folgenden als k-Nächste-Nachbarn-Imputation

bezeichnet. Auf Basis einer Faustregel wird dabei in der Voreinstellung stets $k = 0.01n$ verwendet; wenn nicht explizit anders angegeben ist dies auch der Wert, der in den Simulationsstudien in Kapitel 6 verwendet wird.

Der Vorteil bei einem solchen Vorgehen liegt vor allem darin, dass es im Gegensatz zu den anderen vorgestellten Imputationsmöglichkeiten verteilungsfrei ist und die Methodik somit auch für eine Vielzahl komplexer und untypischer Datensituationen programmiert und umgesetzt werden kann. Dies zeigt sich insbesondere auch in dem Anwendungsbeispiel aus Abschnitt 7.3, bei dem aufgrund der sehr geringen Stichprobengröße komplexe Regressionsimputationen, wie auch eine verteilungsbasierte Imputation auf Basis des EM-Algorithmus, keine Imputationen erzeugen können. Eben diese Verteilungsfreiheit führt aber auch dazu, dass keine (sinnvollen) multiplen Imputationen generiert werden können, was – wie zu Beginn dieses Kapitels beschrieben – meist zu einer Unterschätzung der Varianz führt. Für einige wenige, spezielle Situationen existieren jedoch Korrekturverfahren, die die Varianz nicht unterschätzen, vergleiche Chen und Shao (2001).

Problematisch erscheint ferner die Wahl einer geeigneten Metrik $d(\cdot, \cdot)$. Es ist sicher selten, dass ein Datensatz nur metrische Variablen enthält und die Wahl der euklidischen Distanz damit einen Sinn ergibt. Die Sensitivität der Ergebnisse bezüglich der Wahl der Metrik ist damit eine wichtige, jedoch in der Literatur in diesem Zusammenhang selten erwähnte, Problematik, die auch das oben angedeutete EMV-Paket von Gottardo (2008) nicht berücksichtigt. Es stellt als Alternative zur euklidischen Distanz nur die Pseudo-Metrik der Stichprobenkorrelation zur Verfügung.

Bootstrap-Imputation auf Basis des EM-Algorithmus

Die analytische Bestimmung oder eine iterative Approximation der prädiktiven a-posteriori-Verteilung (5.1) ist oft sehr aufwändig und zeitintensiv. Das *R*-Paket „Amelia II“ (Honaker, King und Blackwell (2008)) verfolgt einen neuen, wenig rechenintensiven, Bootstrap-basierten Ansatz um korrekte (multiple) Imputationen zu erhalten. Dabei wird der Datenmatrix D eine multivariate Normalverteilung unterstellt, $D \sim N(\mu, \Sigma)$. Kategoriale Größen mit k Ausprägungen werden dabei in $k - 1$ binäre Dummys umkodiert, binäre Größen bleiben bestehen. Wie zu Beginn dieses Kapitels erwähnt ist die vereinfachte Annahme einer multivariaten Normalverteilung nicht unproblematisch, liefert aber in vielen Situationen vernünftige Ergebnisse. Ausgehend von dieser Annahme werden M Bootstrap-Stichproben der gleichen Größe wie der Originaldaten gezogen und für jede Stichprobe auf Basis des EM-Algorithmus die posteriori Modi $\mu^*_{(m)}$ und $\Sigma^*_{(m)}$,

$m = 1, \ldots, M$, bestimmt. Dieser Schritt soll die Betrachtung der posteriori-Verteilung $p(\theta|D^{\text{obs}})$ aus (5.1) simulieren und ersetzen. Eine ausführliche Rechtfertigung ohne formellen Beweis findet sich bei Honaker und King (2010). Um korrekte, multiple Imputationen $\tilde{x}_{ij}^{(m)}$, $m = 1, \ldots, M$, gemäß der prädiktiven a-posteriori-Verteilung $p(D^{\text{mis}}|D^{\text{obs}})$ zu erhalten, können aufgrund der Annahme der multivariaten Normalverteilung die Werte schlicht über die daraus resultierende Regressionsvorhersage in den Originaldaten

$$\tilde{x}_{ij}^{(m)} \quad = \quad \mathbf{x}_{i_{/j}}\beta_{(m)}^{*} + \epsilon_{(m)}^{*}$$

bestimmt werden. Dabei beschreibt $\mathbf{x}_{i_{/j}}$ die i-te Beobachtung ohne den entsprechenden Wert der Variable X_j und $\beta_{(m)}^{*}$ bzw. $\epsilon_{(m)}^{*}$ sind die aus $\mu_{(m)}^{*}$ und $\Sigma_{(m)}^{*}$ eindeutig resultierenden Parameter. Somit erhält man M imputierte Datensätze, die als sinnvolle Wahl zur korrekten multiplen Imputation auf Basis von (5.1) angesehen werden können.

5.1.3 Weitere Ansätze

Es existieren in der Literatur drei weitere Ansätze zur Berücksichtigung der Problematik fehlender Daten in der Modellselektion, die allesamt Akaikes Informationskriterium für fehlende Daten adjustieren. Die erste Veröffentlichung von Shimodaira (1994), der sein Kriterium als PDIO (Predictive Divergence for Incomplete Observation models) bezeichnet, wie auch die darauf aufbauende Erweiterung von Cavanaugh und Shumway (1998), die ihre Modifikation mit AICcd (AIC complete data) benennen, betrachten ausschließlich den Fall fehlender Werte im Response und vollständig beobachteter Werte in den Kovariablen. Die vorliegende Arbeit diskutiert dagegen im Wesentlichen den Fall unvollständiger Werte in den Kovariablen, weswegen an dieser Stelle nur auf Cavanaugh und Shumway (1998, Abschnitt 4ff.) verwiesen werden soll; hier werden die Kernkonzepte der beiden Kriterien ausführlich motiviert, verglichen und ferner erläutert, wie die aufwändige Implementation von PDIO und AICcd gestaltet werden kann.

Einen etwas anderen Ansatz verfolgt die Arbeit von Claeskens und Consentino (2008). Ihr Kriterium ist eine Adjustierung des AIC und stellt ebenfalls einen unverzerrten Schätzer für die erwartete Kullback-Leibler-Distanz (3.21) dar. Voraussetzung für die Gültigkeit des Kriteriums ist dabei, dass die fehlenden Werte ausschließlich in den Kovariablen auftreten und nicht im Response. Es lautet

$$AIC_{\text{mis}} \quad = \quad -2Q(\hat{\theta}|\hat{\theta}) + 2K\,, \tag{5.15}$$

wobei $\hat{\theta}$ die ML-Schätzung von θ auf Basis des EM-Algorithmus beschreibt und $Q(\hat{\theta}|\hat{\theta})$ die dabei relevante Größe im E-Schritt darstellt, $Q(\hat{\theta}|\hat{\theta}) = \mathbb{E}_{D^{\text{mis}}|D^{\text{obs}},\hat{\theta}}\left[\mathcal{L}(\hat{\theta}|D^{\text{mis}}, D^{\text{obs}})\right]$. Es ist offensichtlich, dass dies einen MAR-Fehlendmechanismus voraussetzt. Zur konkreten Berechnung von $Q(\hat{\theta}|\hat{\theta})$ verwenden die Autoren einen Monte-Carlo EM-Algorithmus, was insbesondere daran liegt, dass das AIC_{mis} nur dann Gültigkeit besitzt, wenn alle Modelle mindestens eine Kovariable mit fehlenden Werten enthalten; ansonsten muss eine Korrektur vorgenommen werden, die eine analytische Berechnung von $Q(\hat{\theta}|\hat{\theta})$ auch für vermeintlich einfache Verteilungsfamilien verhindert. Wie in Abschnitt 5.1.1 angedeutet, basieren die vorwiegend guten Ergebnisse dieses Kriteriums in den Simulationsstudien der Autoren auf der unhaltbaren Annahme, dass überangepasste Modelle als korrekt angesehen werden können.

Ein weiterer, offensichtlich trivialer Ansatz besteht darin, den Selektionsschritt auf dem Datensatz der vollständigen Fälle D_*^c durchzuführen. Für eine Menge an Kandidatenmodellen $\mathcal{M} = \{M_1 \ldots, M_k\}$ wird dann dasjenige Modell $M_\kappa \in \mathcal{M}$ gewählt, das ein Kriterium oder Verfahren Γ, wie in Kapitel 3 beschrieben, auf D_*^c minimiert (bzw. maximiert) und somit zu der Wahl des Parameterschätzers

$$\hat{\theta}_{\text{CC}} \;=\; \hat{\theta}_\kappa : \arg\min_{M_\kappa \in \mathcal{M}} \{\Gamma(M_\kappa; \hat{\theta}_\kappa)|D_*^c\}\,, \qquad \kappa = 1, \ldots, k\,, \tag{5.16}$$

führt. Wie zu Beginn des Kapitels erläutert, hängt die Qualität dieses Schätzers sehr stark von dem zugrundeliegenden Fehlendmechanismus und der Reduzierung des Stichprobenumfangs ab. Die Simulationsergebnisse aus Kapitel 6 unterstreichen, dass Korrekturverfahren, wie etwa die Verwendung des AIC_W oder die „Selektion nach Imputation“-Strategie – wie zu erwarten – in der Regel bessere Schätzungen liefern als (5.16).

5.2 Modellmittelung bei fehlenden Daten

Die im vorangegangenen Abschnitt 5.1 angestellten Überlegungen erlauben es, die Problematik fehlender Daten in der Modellselektion sowohl unter Zuhilfenahme von Gewichtungsansätzen (Abschnitt 5.1.1) als auch unter Verwendung geeigneter Imputationsmethoden (Abschnitt 5.1.2) zu berücksichtigen. Wie in Kapitel 4 ausführlich diskutiert führt die datengestützte Selektion von Modellen jedoch zu einer Unsicherheit, die in der Inferenz eigentlich nicht vernachlässigt werden sollte; dies gilt selbstverständlich auch im Kontext fehlender Daten. Im Folgenden wird daher aufgezeigt wie gängige frequentistische Modellmittelungsansätze für den Kontext fehlender Daten adjustiert werden

können. Die Eigenschaften der vorgeschlagenen Schätzer werden in Kapitel 6 ausführlich untersucht und bewertet.

5.2.1 Mittelung mit adjustierten Kriterien

Ein erster, pragmatischer Ansatz besteht darin, für fehlende Daten adjustierte Kriterien zur Konstruktion von Gewichten eines Modellmittelungsschätzers zu benutzen: Es liegt nahe, anstelle der exponentiellen SBC-Gewichte (4.4) bzw. der exponentiellen AIC-Gewichte (4.6), entsprechende exponentielle Gewichte auf Basis eines in diesem Kontext sinnhaften Kriteriums Γ zu verwenden, das explizit den Umstand fehlender Beobachtungen berücksichtigt. Wie bereits erwähnt, hält die Literatur für den Fall fehlender Responsewerte hierfür bisher das PDIO-Kriterium von Shimodaira (1994) und das AICcd von Cavanaugh und Shumway (1998), für fehlende Werte in den Kovariablen ein EM-basiertes Kriterium von Claeskens und Consentino (2008) sowie im Allgemeinen das gewichtete AIC-Kriterium AIC_W von Hens, Aerts und Molenberghs (2006) bereit. Für letzteres ergibt sich unter Beachtung der Konzepte aus den Abschnitten 4.2.1 und 5.1.1 der Modellmittelungsschätzer

$$\hat{\bar{\theta}}_{\mathrm{AIC}_W} = \sum_{\kappa=1}^{k} w_\kappa^{(8)} \hat{\theta}_{W,\kappa} \tag{5.17}$$

mit

$$w_\kappa^{(8)} = \frac{\exp(-\frac{1}{2} AIC_{W,\kappa})}{\sum_{\kappa=1}^{k} \exp(-\frac{1}{2} AIC_{W,\kappa})}, \tag{5.18}$$

wobei $\hat{\theta}_{W,\kappa}$ die gewichtete ML-Schätzung auf Basis der Gewichte (5.5) und $AIC_{W,\kappa}$ das gewichtete Akaike-Kriterium für das Modell $M_\kappa \in \mathcal{M}$ bezeichnet. Gemäß (4.12) kann die Varianz von $\hat{\bar{\theta}}_{\mathrm{AIC}_W}$ dabei über

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\bar{\theta}}_{\mathrm{AIC}_W}) = \left\{ \sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_{W,\kappa}|M_\kappa) + (\hat{\theta}_{W,\kappa} - \hat{\bar{\theta}}_{\mathrm{AIC}_W})^2} \right\}^2 \tag{5.19}$$

geschätzt werden. Damit wird die durch den Selektionsprozess verursachte Unsicherheit explizit berücksichtigt; wie bereits erwähnt erfordert die Schätzung der Gewichte für das AIC_W jedoch meist das fitten eines GAMs, was die Wahl eines Glättungsparameters beinhaltet und zu einer weiteren Unsicherheit führt, die nicht durch (5.19) abgedeckt wird. Dies verdeutlichen auch die Simulationsergebnisse aus Kapitel 6.

5.2.2 Mittelung nach Imputation

Eine weitere Möglichkeit die Problematik fehlender Daten in der Modellmittelung zu berücksichtigen, besteht darin, den unvollständigen Datensatz D_* aufzufüllen und den daraus resultierenden imputierten Datensatz D^{imp} für die Bestimmung eines Modellmittelungsschätzers zu verwenden. Unabhängig von der gewählten Imputationsmethode ergibt sich für den Fall einer einfachen, nicht-multiplen Imputation der Schätzer

$$\hat{\bar{\theta}}_{\text{imp}} = \sum_{\kappa=1}^{k} \{ w_\kappa \hat{\theta}_\kappa | D^{\text{imp}} \} , \tag{5.20}$$

dessen Varianz gemäß (4.12) über

$$\widehat{\mathbb{V}\text{ar}}(\hat{\bar{\theta}}_{\text{imp}}) = \left\{ \sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\theta}_\kappa | M_\kappa) + (\hat{\theta}_\kappa - \hat{\bar{\theta}}_{\text{imp}})^2} \, | D^{\text{imp}} \right\}^2 \tag{5.21}$$

geschätzt werden kann. In den Simulations- und Anwendungsbeispielen der Kapitel 6 und 7 werden hierfür die in Abschnitt 5.1.2 vorgestellten Imputationsprozeduren der kNN-Imputation, der verallgemeinerten Regressionsimputation (GLMRI, GAMRI) und der Bootstrap-basierten EM-Imputation verwendet. Für die Wahl der Gewichte in (5.20) ergeben sich dabei viele Möglichkeiten: Es können exponentielle Gewichte auf Basis des SBC, AIC oder jedes anderen sinnhaften Kriteriums Γ aus Kapitel 3, also

$$w_\kappa^{(9)} = \frac{\exp(-\frac{1}{2}\Gamma_\kappa)}{\sum_{\kappa=1}^{k} \exp(-\frac{1}{2}\Gamma_\kappa)} , \tag{5.22}$$

verwendet werden; auch Gewichte als Approximation an die a-posteriori-Wahrscheinlichkeit (4.3), die MMA-Gewichte (4.9) oder die OPT-Gewichte (4.11) ergeben einen Sinn. Aus diesen Überlegungen werden auch die großen Vorteile der Verwendung eines Schätzers auf Basis der imputierten Daten klar: Im Gegensatz zu (5.17) muss die Plausibilität eines Modells nicht über informationstheoretisch motivierte Kriterien beschrieben werden; nahezu jeder Ansatz aus Kapitel 3, etwa die Kriterien unter Beachtung der Vorhersagequalität oder auch der MDL-Methodik, eignen sich in diesem Zusammenhang um einen Schätzer zu konstruieren. In einem so weitreichenden Feld wie der Modellselektion und -mittelung erscheint dies von großem Vorteil.

Dennoch beachten weder (5.20) noch (5.21) die Unsicherheit bezüglich der Imputation. Es liegt insofern nahe, die Konzepte multipler Imputation, wie etwa in (5.2) und (5.3) beschrieben, zu adaptieren: Prinzipiell bedeutet dies nichts anderes als

1. M imputierte Datensätze durch Ziehen aus der prädiktiven a-posteriori-Verteilung (5.1) zu generieren,

2. auf jedem dieser M Datensätze den gewünschten Modellmittelungsschätzer zu berechnen und

3. die dadurch erhaltenen M Schätzer für eine finale Schätzung zu kombinieren.

Dadurch erhält man als eine sinnvolle Punktschätzung

$$\hat{\bar{\theta}}^{M}_{\text{imp}} \;=\; \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\theta}}^{(m)}_{\text{imp}}\,, \tag{5.23}$$

wobei $\hat{\bar{\theta}}^{(m)}_{\text{imp}}$ einen beliebigen Modellmittelungsschätzer auf Basis des m-ten imputierten Datensatzes beschreibt, $m = 1, \ldots, M$. Die Varianz kann dabei offensichtlich über

$$\begin{aligned}
\widehat{\mathbb{V}\text{ar}}(\hat{\bar{\theta}}^{M}_{\text{imp}}) \;&=\; \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{V}\text{ar}}(\hat{\bar{\theta}}^{(m)}_{\text{imp}}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\bar{\theta}}^{(m)}_{\text{imp}} - \hat{\bar{\theta}}^{M}_{\text{imp}})^2 \\
&=\; \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\theta}^{(m)}_{\kappa}) + (\hat{\theta}^{(m)}_{\kappa} - \hat{\bar{\theta}}^{(m)})^2}\right\}^2 \\
&\quad + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\bar{\theta}}^{(m)}_{\text{imp}} - \hat{\bar{\theta}}^{M}_{\text{imp}})^2
\end{aligned} \tag{5.24}$$

geschätzt werden. Ein solches Vorgehen berücksichtigt sowohl die Unsicherheit bezüglich der Imputation als auch die Unsicherheit bezüglich des Selektionsschrittes und ist insofern als Weiterentwicklung sowohl von (4.12) als auch von (5.10) anzusehen.

Eine Reflexion über (5.23) und insbesondere (5.24) führt zu der Erkenntnis, dass das oben beschriebene Prozedere eine Rangordnung in der Betrachtung der Unsicherheitskomponenten impliziert: Zuerst wird die Unsicherheit bezüglich der Modellwahl über Bestimmung eines FMA-Schätzers berücksichtigt, anschließend die Unsicherheit bezüglich der Imputation über die Verwendung multipler Imputationen. Auch wenn ein solches Vorgehen gemäß den oben beschriebenen drei Schritten statistischer Praxis entsprechen mag und sicherlich auch plausibel ist, so stellt sich dennoch die Frage, wie sich eine Änderung dieser Reihenfolge auf den Schätzer (5.23) und seine Varianz auswirkt und ob damit sinnvolle Schätzungen generiert werden können oder nicht. Berechnet man also

zuerst einen Schätzer, der die Imputationsunsicherheit berücksichtigt und anschließend die Unsicherheit bezüglich der Modellwahl, so erhält man offensichtlich

$$\hat{\bar{\theta}}^M_{\text{imp,alt}} \;=\; \sum_{\kappa=1}^{k} w_\kappa^M \hat{\theta}_\kappa^M \,, \tag{5.25}$$

wobei

$$w_\kappa^M \;=\; \frac{1}{M}\sum_{m=1}^{M} w_\kappa^{(m)} \,. \tag{5.26}$$

Tatsächlich sind die Punktschätzungen $\hat{\bar{\theta}}^M_{\text{imp,alt}}$ und $\hat{\bar{\theta}}^M_{\text{imp}}$ in der Regel nicht identisch, da

$$\begin{aligned}\hat{\bar{\theta}}^M_{\text{imp,alt}} \;&=\; \sum_{\kappa=1}^{k} w_\kappa^M \hat{\theta}_\kappa^M = \frac{1}{M^2}\sum_{\kappa=1}^{k}\left\{\sum_{m=1}^{M} w_\kappa^{(m)} \sum_{m=1}^{M} \hat{\theta}_\kappa^{(m)}\right\} \\ &\overset{\text{i.A.}}{\neq}\; \frac{1}{M}\sum_{m=1}^{M}\sum_{\kappa=1}^{k} w_\kappa^{(m)}\hat{\theta}_\kappa^{(m)} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\theta}}^{(m)}_{\text{imp}} = \hat{\bar{\theta}}^M_{\text{imp}} \,.\end{aligned}$$

Auch die $\hat{\bar{\theta}}^M_{\text{imp,alt}}$ zugehörige Varianzschätzung, die über

$$\widehat{\mathbb{V}\text{ar}}(\hat{\bar{\theta}}^M_{\text{imp,alt}}) \;=\; \left\{\sum_{\kappa=1}^{k} w_\kappa^M \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\theta}^M_{\kappa,\text{imp}}) + (\hat{\bar{\theta}}^M_{\text{imp}} - \hat{\theta}^M_{\kappa,\text{imp}})^2}\right\}^2 \tag{5.27}$$

erhalten werden kann, wobei

$$\widehat{\mathbb{V}\text{ar}}(\hat{\theta}^M_{\kappa,\text{imp}}) \;=\; \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{V}\text{ar}}(\hat{\theta}^{(m)}_{\kappa,\text{imp}}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\theta}^M_{\kappa,\text{imp}} - \hat{\theta}^{(m)}_{\kappa,\text{imp}})^2 \,,$$

stimmt nicht mit der Varianzschätzung für $\hat{\bar{\theta}}^M_{\text{imp}}$ gemäß (5.24) überein. Der Vergleich von (5.24) und (5.27) ist dabei keineswegs trivial, da die Quadrate größerer, unterschiedlich konzipierter Summen verglichen werden müssen. Entscheidend ist der Unterschied in Bezug auf die Einflussnahme der Gewichte $w_\kappa^{(m)}$: Während bei den „klassischen“ Schätzungen nach (5.23) und (5.24) die Gewichte des m-ten Datensatzes auch nur auf die Schätzungen des entsprechenden m-ten Datensatzes wirken können, so ist dies bei den „alternativen“ Schätzungen nach (5.25) und (5.27) nicht der Fall; Gewichte, die für den m-ten Datensatz berechnet werden, wirken indirekt über (5.26) immer auch

auf die Schätzungen der anderen, imputierten Datensätze. Ein einfaches Beispiel soll dies illustrieren. Betrachtet man $k = 2$ Modelle und $M = 2$ Imputationen, so erhält man als entsprechende Schätzungen für θ:

$$\begin{aligned}\hat{\bar{\theta}}^M_{\text{imp}} &= \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\theta}}^{(m)}_{\text{imp}} = \frac{1}{M}\sum_{m=1}^{M}\sum_{\kappa=1}^{k} w_\kappa^{(m)}\hat{\theta}_\kappa^{(m)} \\ &= \frac{1}{2}w_1^{(1)}\hat{\theta}_1^{(1)} + \frac{1}{2}w_2^{(1)}\hat{\theta}_2^{(1)} + \frac{1}{2}w_1^{(2)}\hat{\theta}_1^{(2)} + \frac{1}{2}w_2^{(2)}\hat{\theta}_2^{(2)}\end{aligned}$$

bzw.

$$\begin{aligned}\hat{\bar{\theta}}^M_{\text{imp,alt}} &= \sum_{\kappa=1}^{k} w_\kappa^M\hat{\theta}_\kappa^M = \frac{1}{M^2}\sum_{\kappa=1}^{k}\left\{\sum_{m=1}^{M} w_\kappa^{(m)}\sum_{m=1}^{M}\hat{\theta}_\kappa^{(m)}\right\} \\ &= \frac{1}{4}(w_1^{(1)} + w_1^{(2)})(\hat{\theta}_1^{(1)} + \hat{\theta}_1^{(2)}) + \frac{1}{4}(w_2^{(1)} + w_2^{(2)})(\hat{\theta}_2^{(1)} + \hat{\theta}_2^{(2)}) \\ &= \frac{1}{4}w_1^{(1)}\hat{\theta}_1^{(1)} + \frac{1}{4}w_1^{(1)}\hat{\theta}_1^{(2)} + \frac{1}{4}w_1^{(2)}\hat{\theta}_1^{(1)} + \frac{1}{4}w_1^{(2)}\hat{\theta}_1^{(2)} \\ &\quad + \frac{1}{4}w_2^{(1)}\hat{\theta}_2^{(1)} + \frac{1}{4}w_2^{(1)}\hat{\theta}_2^{(2)} + \frac{1}{4}w_2^{(2)}\hat{\theta}_2^{(1)} + \frac{1}{4}w_2^{(2)}\hat{\theta}_2^{(2)}\,.\end{aligned}$$

Offensichtlich wirken bei $\hat{\bar{\theta}}^M_{\text{imp,alt}}$ Gewichte des m-ten Datensatzes auch auf Schätzungen des nicht-m-ten Datensatzes, so beispielsweise $w_1^{(1)}$, das auf $\hat{\theta}_1^{(2)}$ wirkt.

Die Simulationsergebnisse aus Abschnitt 6.3 werden bestätigen, dass – wie vermutet – die Verwendung der vorgestellten, alternativen Schätzungen in den betrachteten Situationen tatsächlich zu qualitativ schlechteren Ergebnissen führt als die intuitiven Schätzungen nach (5.23) und (5.24). Die Reihenfolge der Betrachtung der Unsicherheitskomponenten ist bei Gebrauch multipler Imputationen im Kontext der Modellmittelung also entscheidend.

6. Simulationsstudien

Um die Eigenschaften und das Verhalten der vorgeschlagenen Punktschätzungen und deren Varianz zu eruieren, werden in diesem Kapitel ausführliche Monte-Carlo-Simulationen durchgeführt. Dabei sind die folgenden Fragestellungen von besonderem Interesse:

1. Sind Modellmittelungsverfahren der Modellselektion generell überlegen?
2. Ist ein pragmatischer Modellmittelungsansatz unter Verwendung kriteriumsbasierter Gewichte einem Ansatz der unter Optimalitätsbetrachtungen konstruiert wurde, so etwa dem MMA-Schätzer von Hansen, zwingenderweise unterlegen?
3. Welche Erfolgsaussichten versprechen die in Kapitel 5 vorgestellten Korrekturverfahren für fehlende Daten?

In den Abschnitten 6.1 und 6.2 werden hierfür insbesondere die Qualität der entsprechenden Punktschätzungen untersucht und bewertet; Abschnitt 6.3 widmet sich hauptsächlich der Qualität der zugehörigen Varianzschätzungen und den Auswirkungen multipler Imputation. Abschnitt 6.4 fasst die wichtigsten Resultate zusammen.

6.1 Lineare Regression

In diesem Abschnitt werden am Beispiel zweier ausgewählter FMA-Schätzer sowie eines FMS-Schätzers die oben angeführten Fragestellungen im Kontext des linearen Regressionsmodells

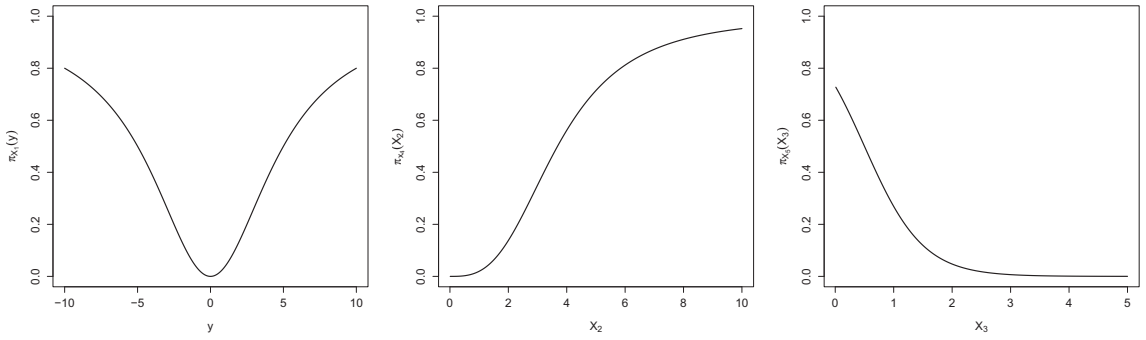
$$y \;=\; X\beta + \epsilon\,, \qquad \epsilon \sim N(0, \sigma^2 I)\,,$$

unter Beachtung der Problematik fehlender Daten erörtert, $\beta = (\alpha, \gamma')'$. Betrachtet werden $\mathcal{R} = 500$ Simulationsläufe, wobei für jeden Lauf je ein $n_{\text{tr}} \times p$ Trainingsdatensatz und ein $n_{\text{test}} \times p$ Testdatensatz erzeugt wird mit $n_{\text{tr}} = 450$, $n_{\text{test}} = 50$ und $p = 6$.

Mit Hilfe eines Clayton-Copulas wird eine multivariate Verteilung für die insgesamt fünf Kovariablen mit den Randverteilungen $X_1 \sim N(0.5, 1)$, $X_2 \sim \log N(0.5, 0.5)$, $X_3 \sim Weibull(1.75, 1.9)$, $X_4 \sim B(1, 0.3)$ und $X_5 \sim Ga(0.25, 2)$ generiert. Die Verteilungen sind so gewählt, dass die Varianz der Kovariablen aus Gründen der Standardisierung immer in etwa Eins beträgt; der Copula-Parameter beträgt durchgehend $\theta_{\text{cop}} = 1$ und resultiert damit in einer mittelstarken Korrelation zwischen allen Kovariablen; für Details bezüglich der Verwendung von Copulas in der statistischen Software *R* sei auf Yan (2007) verwiesen. Der Response y wird aus einer Normalverteilung $N(\mu_0, \sigma_0)$ gezogen mit $\mu_0 = 2.5 - 3X_1 - 0.3X_2 - 2X_4$ und $\sigma_0 = \exp(1)$. Dies bedeutet, dass der Erwartungswert von y maßgeblich von den Variablen X_1, X_2 und X_4 bestimmt wird, daher gilt $\alpha_{\text{wahr}} = 2.5$ und $\gamma_{\text{wahr}} = (-3, -0.3, 0, -2, 0)'$ und damit $\beta_{\text{wahr}} = (\alpha_{\text{wahr}}, \gamma'_{\text{wahr}})'$; der komplette Datensatz wird als $D_{\text{sim1}} = \{y, X_1, X_2, X_3, X_4, X_5\}$ bezeichnet und wird in jedem Simulationslauf neu generiert. Anschließend werden unter Verwendung eines MAR-Fehlendmechanismus Werte von X_1, X_4 und X_5 gemäß der Fehlwahrscheinlichkeitsfunktionen

$$\begin{aligned} \pi_{X_1}(y) &= 1 - \frac{1}{0.04y^2 + 1}\,, \qquad \pi_{X_4}(X_2) = 1 - \frac{1}{1 + 0.02X_2^3}\,, \\ \pi_{X_5}(X_3) &= 1 - \frac{1}{1 + \exp\{1 - 2X_3\}}\,, \end{aligned}$$

als fehlend deklariert. Bei den $\mathcal{R} = 500$ Simulationsläufen fehlen damit im Mittel 31% der Werte von X_1, 15% der Werte von X_4 und 18% der Werte von X_5. Die Gestalt der Fehlwahrscheinlichkeitsfunktionen ist in Abbildung 6.1 illustriert. Um den Zusammenhang



(a) $\pi_{X_1}(y)$ (b) $\pi_{X_4}(X_2)$ (c) $\pi_{X_5}(X_3)$

Abb. 6.1: Gestalt der verwendeten Fehlwahrscheinlichkeitsfunktionen

von y und den Kovariablen zu modellieren, werden fünf konkurrierende Kandidatenmodelle[19] betrachtet:

$$\begin{aligned}
\mathrm{M}_1: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5\,, \\
\mathrm{M}_2: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4\,, \\
\mathrm{M}_3: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_4 X_4 + \gamma_5 X_5\,, \\
\mathrm{M}_4: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_4 X_4\,, \\
\mathrm{M}_5: \quad y &= \alpha + \gamma_1 X_1 + \gamma_4 X_4\,.
\end{aligned}$$

Offensichtlich ist das Modell M_4, das die Kovariablen X_1, X_4 und X_5 enthält, am besten geeignet, um den oben formulierten, datengenerierenden Prozess zu beschreiben. Um nun sowohl die Effekte der Modellselektionsunsicherheit als auch die Problematik fehlender Daten zu berücksichtigen, werden 20 Schätzer, wie in Tabelle 6.1 aufgelistet, analysiert. Alle FMA-Schätzer besitzen dabei die Form

$$\hat{\bar{\beta}}^{\star} \;=\; \sum_{\kappa=1}^{k} w_{\kappa}^{\star} \hat{\beta}_{\kappa}^{\star} \tag{6.1}$$

und unterschieden sich sowohl bezüglich der Datengrundlage zur Berechnung von $\hat{\beta}_{\kappa}^{\star}$ als auch bezüglich der Wahl der Gewichte $w_{\kappa}^{\star}$. Der verwendete FMS-Schätzer kann dabei als Spezialfall des FMA-Schätzers (6.1) angesehen werden, da der durch das Kriterium $\Gamma = \mathrm{AIC}$ ausgewählte Schätzer $\hat{\beta}_{\kappa}^{\mathrm{AIC}}$ das Gewicht 1 zugewiesen bekommt, alle weiteren Schätzer das Gewicht 0.

Die Qualität dieser Schätzer wird anhand vier verschiedener Verlustfunktionen beurteilt: Die Verlustfunktion L_1 beschreibt dabei den MSE der angeführten Schätzer bezüglich β_{wahr},

$$L_1 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \sum_{j=1}^{p} (\hat{\bar{\beta}}_{j,r}^{\star} - \beta_{j,\mathrm{wahr}})^2 \right\}\,, \tag{6.2}$$

[19] Die Auswahl der Kandidatenmodelle bildet die Grundlage zur Berechnung der FMA- und FMS-Schätzer gemäß Tabelle 6.1. Einzige Ausnahme ist der Mallows-Model-Averaging-Schätzer, der gemäß seiner Definition (vgl. Abschnitt 4.2.2) als Kandidatenmodelle *immer* das volle Modell (hier: M_1) und die entsprechenden verschachtelten Submodelle auf Basis der geordneten Regressoren verwendet. Dies gilt es auch für die folgenden Experimente in diesem Abschnitt zu beachten, da an den entsprechenden Stellen auf diesen Sachverhalt nicht noch einmal explizit darauf hingewiesen wird.

Auswahl an FMA- und FMS-Schätzern $\hat{\hat{\beta}}^{\star}$ bei fehlenden Daten	
(a)	**FMA-Akaike-Schätzer**: der FMA-Schätzer auf Basis der exponentiellen AIC-Gewichte (4.6) bzw. (5.18), der den Umstand fehlender Daten wie folgt berücksichtigt: 1) *Original* – der FMA-Akaike-Schätzer unter Verwendung des vollständigen Datensatzes D_{sim1}; dieser Schätzer dient als Referenz. 2) *CC* – der FMA-Akaike-Schätzer unter Verwendung des Subdatensatzes der vollständig beobachteten Fälle D_*^c. 3) *GAMRI, GLMRI, kNN, Amelia* – es wird das Prinzip „Mittelung nach Imputation“ angewendet; dies entspricht dem Schätzer (5.20) unter Verwendung des aufgefüllten Datensatzes D^{imp} gemäß den Imputationsmethoden der in Abschnitt 5.1.2 vorgestellten verallgemeinerten Regressionsimputation (GAMRI, GLMRI), der k-Nächsten-Nachbarn-Imputation (kNN) und der Bootstrap-basierten Imputation des R-Pakets „Amelia II“ (Amelia). 4) *AIC$_W$* – entspricht dem adjustierten FMA-Schätzer (5.17).
(b)	**FMA-Hansen-Schätzer**: der FMA-Schätzer auf Basis der optimalen Gewichte (4.9), der den Umstand fehlender Daten wie in (a) gemäß der Strategien 1), 2) und 3) berücksichtigt; dieser Schätzer wird auch als MMA-Schätzer bezeichnet.
(c)	**FMS-AIC-Schätzer**: der FMS-Schätzer für den $\Gamma = \text{AIC}$ und der den Umstand fehlender Daten wie folgt berücksichtigt: 1) *Original* – der FMS-AIC-Schätzer wird für den vollständigen Datensatz D_{sim1} berechnet. 2) *CC* – entspricht dem FMS-Schätzer (5.16). 3) *GAMRI, GLMRI, kNN, Amelia* – es wird das Prinzip „Selektion nach Imputation“ angewendet; der Schätzer (5.8) basiert dabei auf den Imputationsmethoden wie in (a) beschrieben. 4) *AIC$_W$* – der Schätzer wird durch das Kriterium $\Gamma = \text{AIC}_W$ bestimmt.

Tab. 6.1: Die im Grundszenario verwendeten Modellmittelungsschätzer und Modellselektionsschätzer für verschiedene Strategien zur Berücksichtigung der Problematik fehlender Daten

wobei $\hat{\hat{\beta}}^{\star}_{j,r}$ eine FMA- bzw. FMS-Schätzung des j-ten Elements von β, $j = 1 \ldots, p$, im r-ten Simulationslauf bezeichnet; ferner wird der MSE bezüglich der r-ten Schätzung auf Basis der Originaldaten, $\hat{\hat{\beta}}^{\text{org}}_r$, betrachtet,

$$L_2 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \sum_{j=1}^{p} (\hat{\hat{\beta}}^{\star}_{j,r} - \hat{\hat{\beta}}^{\text{org}}_{j,r})^2 \right\} , \tag{6.3}$$

wobei sich $\hat{\hat{\beta}}^{\text{org}}_r$ immer auf die entsprechende FMA- bzw. FMS-Methodik bezieht. Um die Vorhersagequalität zu beurteilen, wird der MSE der Schätzer bezüglich μ im Testdatensatz berechnet,

$$L_3 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{\tilde{\mu}}^{\star}_{i,r} - \mu_{i,r,\text{wahr}})^2 \right\} , \tag{6.4}$$

mit $\hat{\tilde{\mu}}^{\star}_{i,r} = \hat{\tilde{\alpha}}^{\star}_{r} + \mathbf{x}_{i,r}\hat{\tilde{\gamma}}^{\star}_{r}$ und $\mu_{i,r,\text{wahr}} = \alpha_{\text{wahr}} + \mathbf{x}_{i,r}\gamma_{\text{wahr}}$ sowie der MSE bezüglich y im Testdatensatz,

$$L_4 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{\tilde{y}}^{*}_{i,r} - y_{i,r,\text{wahr}})^2 \right\} , \tag{6.5}$$

wobei $\hat{\tilde{y}}^{*}_{i,r} = \hat{\tilde{\mu}}^{\star}_{i,r}$ den Vorhersagewert für y_i im r-ten Simulationslauf bezeichnet.

Resultate

Die Resultate der Simulation sind in Abbildung 6.2 dargestellt, die detaillierten Ergebnisse befinden sich in Tabelle B.1 in Anhang B. Es lassen sich folgende Ergebnisse konstatieren:

- Unabhängig von der gewählten Methodik zur Berücksichtigung fehlender Werte und unabhängig von der betrachteten Verlustfunktion dominiert der FMA-Akaike-Schätzer sowohl den FMS-AIC-Schätzer als auch den FMA-Hansen-Schätzer. Der FMS-AIC-Schätzer liefert in nahezu allen Situationen geringere Verluste als der FMA-Hansen-Schätzer. Da die Verlustfunktionen jedoch über alle $\mathcal{R} = 500$ Simulationen mitteln, stellt sich die Frage, ob sich diese Aussagen bezüglich der Schätzer auch in allen bzw. den meisten Simulationsläufen beobachten lassen oder nicht. Um dies zu untersuchen, wird in Tabelle 6.2 eine Zusammenfassung der Verteilung der Verluste $\tilde{L}_1$,

  $$\tilde{L}_1 \;=\; \sum_{j=1}^{p} (\hat{\tilde{\beta}}^{\star}_{j} - \beta_{j,\text{wahr}})^2 \,|\, \text{Simulationslauf } r \,, \tag{6.6}$$

  über die 500 Läufe exemplarisch für die Methodik einer kNN-Imputation dargestellt. Es ist offensichtlich, dass der FMA-Akaike-Schätzer durchweg bessere Ergebnisse erzielt als die beiden anderen Schätzer. Die Dominanz des FMS-AIC-Schätzers bezüglich des MMA-Schätzers fällt jedoch nicht ganz so eindeutig aus

wie zunächst vermutet: Es scheint als hinge der Erfolg der Methodologie von Hansen in gewissem Maße von der konkreten Datensituation ab, weswegen die beiden Schätzer bei Betrachtung des ersten Quartils Verluste in ähnlicher Größenordung aufweisen, für den Median und das dritte Quartil der FMS-AIC-Schätzer dem MMA-Schätzer aber überlegen ist.

- Unabhängig von der betrachteten Verlustfunktion werden die Schätzer auf Basis einer Complete Case Analyse sowohl von den Schätzern, die Imputationsmethoden verwenden, als auch von denen auf Basis des AIC_W dominiert; Die Verluste L_1, L_2, L_3 und L_4 sind bei den Imputationsverfahren jeweils am geringsten. Die EM-basierte Imputation unter Verwendung des R-Pakets „Amelia II“ schneidet dabei am besten ab, es folgt die kNN-Methode, gefolgt von GAMRI und GLMRI. Um auch hier zu prüfen, ob sich diese Aussagen konsequent über alle $\mathcal{R} = 500$ Simulationsläufe erkennen lassen, wird in Tabelle 6.3 die Zusammenfassung der Verteilung des Verlustes $\tilde{L}_1$ über alle 500 Läufe für den FMA-Akaike-Schätzer unter Verwendung ausgewählter Methoden zur Berücksichtigung fehlender Werte aufgelistet. Es bestätigt sich das Bild, dass eine Complete Case Analyse durchweg am schlechtesten abschneidet, gefolgt vom Gewichtungsansatz unter Verwendung des AIC_W und den Imputationsmethoden.

Methodik	Minimum	1. Quartil	Median	arith. Mittel	3. Quartil	Maximum
FMA-Akaike	0.0294	0.3181	0.4957	0.6021	0.7614	2.7710
FMA-Hansen	0.0258	0.3452	0.5590	0.6755	0.8463	3.1820
FMS-AIC	0.0261	0.3487	0.5333	0.6434	0.7971	2.9170

Tab. 6.2: Zusammenfassung der Resultate im Grundszenario (Verlust $\tilde{L}_1$) in den 500 Simulationsläufen für den FMA-Akaike-, FMA-Hansen- und FMS-AIC-Schätzer nach Imputation mit der kNN-Methode

Methodik	Minimum	1. Quartil	Median	arith. Mittel	3. Quartil	Maximum
Complete Cases	0.2182	0.9787	1.4100	1.5220	1.9050	4.9580
Imputation (GAMRI)	0.0393	0.4254	0.6813	0.8403	1.1170	4.0270
Imputation (Amelia)	0.0110	0.1953	0.3520	0.4644	0.6205	3.5580
Gewichtung (AIC_W)	0.0376	0.4031	0.7524	1.0710	1.4220	6.4180

Tab. 6.3: Auswahl an Resultaten im Grundszenario (Verlust $\tilde{L}_1$) in den 500 Simulationsläufen für den FMA-Akaike-Schätzer abhängig von der Behandlung der fehlenden Werte

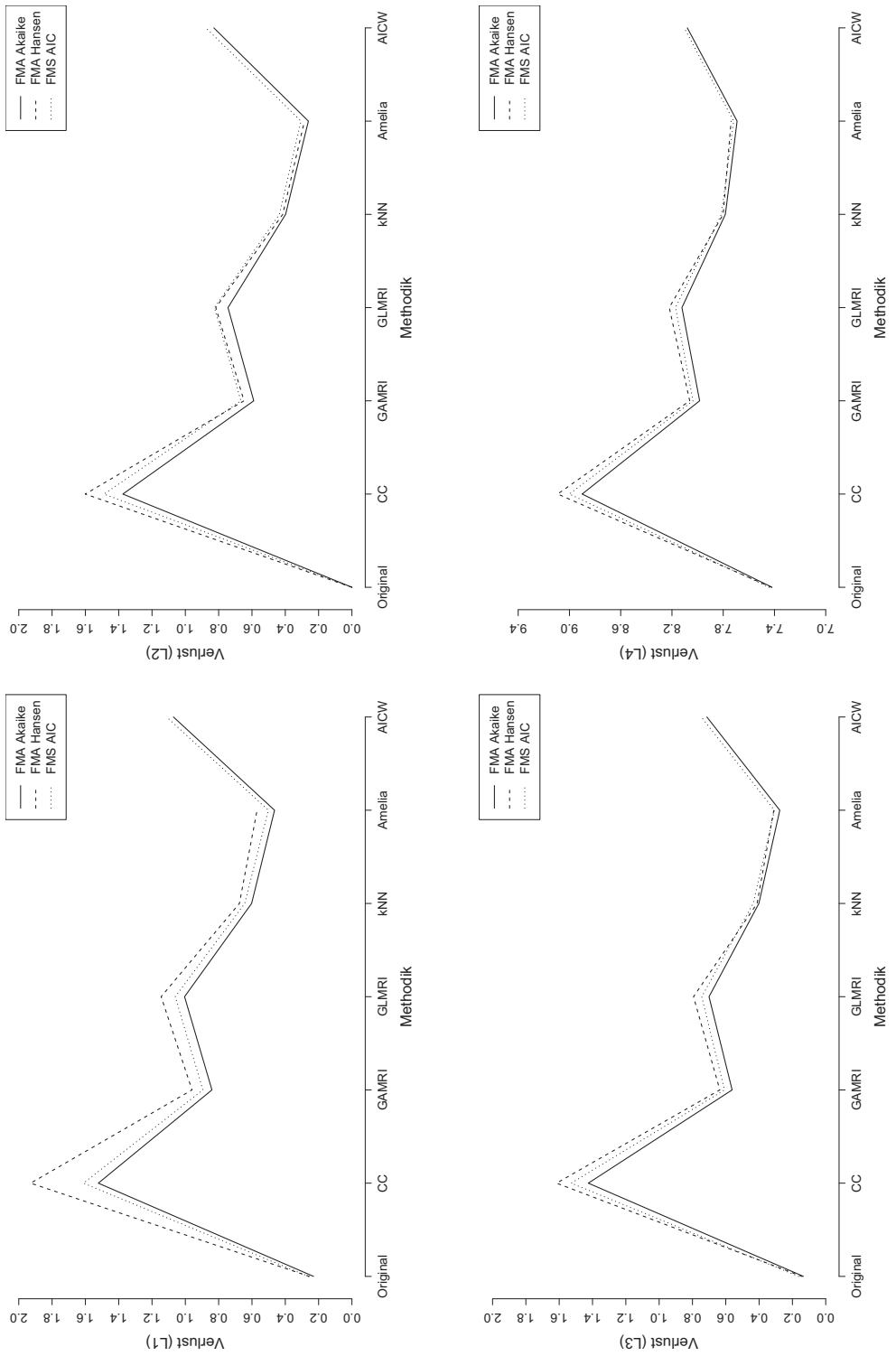


Abb. 6.2: Resultate im Grundszenario bezüglich der Verlustfunktionen L_1, L_2, L_3 und L_4

Um nun auch die Qualität der Varianzschätzungen der FMA- bzw. FMS-Schätzer zu beurteilen, wird der empirische Standardfehler

$$\widehat{\mathrm{se}}_{\beta_{j,\mathrm{emp}}} = \sqrt{\frac{1}{\mathcal{R}-1}\sum_{r=1}^{\mathcal{R}}(\hat{\tilde{\beta}}^{\star}_{j,r} - \hat{\bar{\tilde{\beta}}}^{\star}_{j})^2} \tag{6.7}$$

für jedes β_j, $j = 0, \ldots, 5$, berechnet, wobei $\hat{\bar{\tilde{\beta}}}^{\star}_{j}$ den mittleren FMS- bzw. FMA-Schätzer bezüglich aller Simulationsläufe bezeichnet. Dieser wird mit der mittleren Schätzung des Standardfehlers nach Buckland, Burnham und Anderson (1997) gemäß

$$\widehat{\mathrm{se}}_{\beta_j} = \frac{1}{\mathcal{R}}\sum_{r=1}^{\mathcal{R}}\left\{\sum_{\kappa=1}^{k}\hat{w}_{\kappa,r}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^{\star}_{j,\kappa,r}|M_\kappa) + (\hat{\beta}^{\star}_{j,\kappa,r} - \hat{\bar{\tilde{\beta}}}^{\star}_{j,r})^2}\right\} \tag{6.8}$$

verglichen. Liegen die Werte von (6.7) und (6.8) nahe beisammen, deutet dies auf eine gute Qualität der Varianzschätzung (6.8) hin, andernfalls kann man davon ausgehen, dass eine oder mehrere Unsicherheitskomponenten nicht erfasst werden konnten. Die Resultate befinden sich in Tabelle B.2 im Anhang. Es lassen sich folgende Ergebnisse konstatieren:

- Die beiden Modellmittelungsschätzer führen in der Regel zu höheren Varianzschätzungen gemäß (6.8) als der Selektionsschätzer. Dies war zu erwarten, da durch das Kombinieren mehrerer Modelle explizit die Unsicherheit bezüglich der Selektion modelliert wird.

- Betrachtet man den FMA-Akaike-Schätzer für die Originaldaten (ohne fehlende Werte), so ist zu erkennen, dass die Verwendung von (6.8) für alle β_j fast exakt dem zugehörigen empirischen Standardfehler entspricht. Dies legt nahe, dass (zumindest in dieser Simulation) der Ansatz von Buckland, Burnham und Anderson (1997) geeignet ist, um die Varianz des FMA-Akaike-Schätzers adäquat zu erfassen. Die Varianzschätzungen für den MMA-Schätzer liefern ein differenzierteres Ergebnis: Die Standardfehler (6.7) und (6.8) liegen nicht immer sehr nahe beieinander, insgesamt scheint die Schätzung etwas instabiler zu sein. Für den FMS-AIC-Schätzer fallen die Schätzungen der entsprechenden Standardfehler geringer aus und liegen ausnahmslos unter dem empirischen Standardfehler. Auch dies war zu erwarten, da hier – wie bereits erwähnt – die Unsicherheit bezüglich der Selektion nicht beachtet wird.

- Die Betrachtung der Schätzer der Standardfehler auf Basis der vollständigen Fälle D_*^c führt zu denselben Aussagen wie die Betrachtung der Schätzer der Standardfehler auf Basis der Originaldaten.

- Die Verwendung von Imputationen, und damit der Strategien „Selektion nach Imputation“ und „Mittelung nach Imputation“ führen zu einer Unterschätzung der Varianz; der geschätzte Standardfehler (6.7) ist ausnahmslos kleiner als der empirische Standardfehler (6.8). Wie in den Abschnitten 5.1.2 und 5.2.2 bereits diskutiert, mag dies in erster Linie daran liegen, dass die Unsicherheit bezüglich der Imputation in (6.8) nicht berücksichtigt wird. Eine geeignete Adjustierung mit Hilfe multipler Imputationen wird in Abschnitt 6.3 diskutiert.

- Die Verwendung des AIC_W zur Selektion bzw. zur Konstruktion von FMA-Gewichten führt ebenfalls zu einer Unterschätzung der Varianz. Potentiell ist dies auf die Unsicherheit in der Wahl des Glättungsparameters bzw. der Kovariablen für das GAM zur Schätzung der Gewichte (5.5) zurückzuführen.

Es stellt sich die Frage, wie sensitiv die in diesem Abschnitt getroffenen Aussagen bezüglich des Simulationssettings sind; hierfür werden im Folgenden kritische Annahmen der Simulation variiert und darauf aufbauend die bisherigen Resultate überprüft.

Experiment 2: Veränderung des Fehlendmechanismus

In der oben betrachteten Simulation wird von einem MAR-Fehlendmechanismus ausgegangen. Es stellt sich die Frage, ob 1) bei komplett zufälligem Fehlen der Daten, also einem MCAR-Fehlendmechanismus, die Aussagen bezüglich der FMA- und des FMS-Schätzers bestehen bleiben und ob 2) die Verwendung von Korrekturverfahren zur Berücksichtigung der Problematik fehlender Daten bezüglich einer CC-Analyse einen Gewinn erbringt oder nicht. Hierfür wird erneut das Grundszenario betrachtet, jedoch unter Verwendung der konstanten Fehlwahrscheinlichkeitsfunktionen $\pi_{X_1} = \pi_{X_4} = \pi_{X_5} = 0.1$. Die Resultate befinden sich in Tabelle B.3 im Anhang. Bezüglich der ersten Frage lässt sich festhalten, dass der FMA-Akaike-Schätzer weiterhin bezüglich aller Verlustfunktionen und aller Strategien zur Berücksichtigung der fehlenden Werte den beiden anderen Schätzern überlegen ist. Im Gegensatz zum Grundszenario scheint der der MMA-Schätzer häufig etwas bessere Resultate zu erzielen als der die FMS-AIC-Schätzer, diese sind jedoch meist marginal. Insgesamt liegt das Abschneiden dieser beiden Schätzer wohl in einer ähnlichen Größenordnung. Für die zweite Frage lässt sich konstatieren,

dass bei einem MCAR-Fehlendmechanismus die Imputationsverfahren bzw. die Verwendung des adjustierten Kriteriums AIC_W oft nur geringfügig besser abschneiden als eine simple Complete Case Analyse. Dies war zu erwarten, da wie eingangs in Kapitel 5 bemerkt, für einen MCAR-Fehlendmechanismus und unter Verwendung der vollständigen Fälle D_*^c in der linearen und logistischen Regressionsanalyse weiterhin von konsistenten Schätzungen der Regressionsparameter ausgegangen werden kann.

Experiment 3: Geringere Anzahl an fehlenden Werten

Das oben angeführte Grundszenario betrachtet fünf Kovariablen. Zwei dieser Kovariablen, X_2 und X_3, sind vollständig beobachtet, die weiteren drei Kovariablen, X_1, X_4 und X_5, weisen fehlende Werte auf, im Mittel bei 31%, 15% bzw. 18% der Fälle. In diesem Experiment werden nur für zwei Variablen, X_4 und X_5, Werte als fehlend deklariert und zwar gemäß der Fehlwahrscheinlichkeitsfunktion

$$\tilde{\pi}_{X_4,X_5}(y) \;=\; 1 - \frac{1}{0.01y^2 + 1}\,,$$

die für die $\mathcal{R} = 500$ Simulationsläufe im Mittel in 14% fehlender Fälle für X_4 und 13% fehlender Fälle für X_5 resultiert. Insgesamt liegt also die gleiche Situation wie im Grundszenario vor, jedoch mit einer geringeren Anzahl an fehlenden Werten; die Resultate dieses Experiments sind in Tabelle B.4 dargestellt. Es lässt sich erkennen, dass die beiden FMA-Schätzer den FMS-Schätzer für alle Verlustfunktionen und die meisten Strategien zur Berücksichtigung fehlender Werte dominieren. Der MMA-Schätzer nach Hansen schneidet in diesem Experiment etwas besser ab als in den ersten beiden Experimenten. Weiterhin führt die Verwendung von Imputationsmethoden bzw. Strategien auf Basis des AIC_W zu deutlich besseren Resultaten als eine Complete Case Analyse. Im Unterschied zum Grundszenario ist die Imputation mit dem *R*-Paket „Amelia II“ nicht mehr einheitlich besser als die kNN-Methode oder die verallgemeinerten Regressionsimputationen. Insbesondere bei Verwendung des MMA-Schätzers lässt sich dies erkennen.

Experiment 4: Berücksichtigung zufälliger Effekte

Dieses Experiment unterscheidet sich vom Grundszenario in der Annahme, dass $y \sim N(\mu_0 + \epsilon_0, \sigma_0)$, $\epsilon_0 \sim N(0, 2)$, also zufällige Effekte vorliegen. Die Resultate für dieses Szenario befinden sich in Tabelle B.5. Im Wesentlichen werden die Resultate des

Grundszenarios bestätigt, nahezu alle gefundenen Strukturen finden sich auch hier wieder und sind in der Regel etwas stärker ausgeprägt. Interessant ist erneut, dass der MMA-Schätzer in den meisten Fällen keine Verbesserung im Vergleich zu dem Selektionsschätzer auf Basis des AIC erzielen kann.

Experiment 5: Betrachtung höherer Abhängigkeitsstrukturen

Im Grundszenario wird angenommen, dass alle Kovariablen mittelstark korreliert sind; der Copula-Parameter $\theta_{\text{cop}} = 1$ führt dort für die metrischen Variablen zu Korrelationen nach Bravais-Pearson in der Größenordnung zwischen 0.2 und 0.4. Ein interessanter Fall liegt dann vor, wenn die Variablen sehr stark korrelieren, jedoch noch nicht wirklich linear abhängig sind, also im Grenzgebiet zur Kollinearität liegen. In diesem Experiment wird das Grundszenario dahingehen adjustiert, als dass $\theta_{\text{cop}} = 3$ gesetzt wird, was für die metrischen Kovariablen in Korrelationen nach Bravais-Pearson zwischen 0.5 und 0.8 resultiert. Die Größe des Trainingsdatensatzes wird ebenfalls erhöht, $n_{\text{train}} = 700$. Die Ergebnisse dieses Experiments befinden sich in Tabelle B.6. Die Aussagen bezüglich des FMA-Akaike-, des FMA-Hansen- und des FMS-AIC-Schätzers bleiben gegenüber dem Grundszenario in etwa dieselben: Der erstgenannte dominiert in der Regel die beiden anderen und der Modellmittelungsschätzer nach Hansen kann im Vergleich zu einem einfachen AIC-Selektionsschätzer in der Regel keine Verbesserungen erwirken. Interessant sind jedoch die Resultate bezüglich der Methoden zur Berücksichtigung fehlender Daten: Die verallgemeinerte Regressionsimputation, die anstelle der generalisierten additiven Modelle schlichte GLMs verwendet (GLMRI), erbringt in diesem Szenario – für alle drei Schätzer und alle Verlustfunktionen – keine Verbesserung bezüglich einer einfachen Complete Case Analyse. Ferner führt die GAMRI-Imputation zwar noch zu Verbesserungen bezüglich einer CC-Methodik, ist jedoch unabhängig von der gewählten Verlustfunktion bzw. dem gewählten Schätzer schlechter als die Strategien, die das AIC_W verwenden. Im Unterschied zum ursprünglichen Experiment kann nicht mehr festgehalten werden, dass die Imputationsmethoden generell besser abschneiden als AIC_W-Methodologien.

Experiment 6: Höhere Komplexität I (Interaktion)

In diesem Experiment wird zusätzlich ein Interaktionseffekt betrachtet. Hierfür erfolgt die Modifikation von $y \sim N(\mu_1, \sigma_0)$ mit $\mu_1 = 2.5 - 3X_1 - 0.3X_2 - 2X_4 + X_1 \cdot X_4$. Die Fehlwahrscheinlichkeitsfunktionen $\pi_{X_4}(X_2)$ und $\pi_{X_5}(X_3)$ sind dieselben wie im Grundszenario, $\pi_{X_1}(y)$ wird dagegen geringfügig adjustiert,

$$\pi_{X_1}(y) \;=\; 1 - \frac{1}{0.0225y^2 + 1}\,.$$

Die Kandidatenmodelle sind nun

$$\begin{aligned}
\mathrm{M}_1: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_1 X_4\,,\\
\mathrm{M}_2: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_4 X_4 + \gamma_6 X_1 X_4\,,\\
\mathrm{M}_3: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_4 X_4\,,\\
\mathrm{M}_4: \quad y &= \alpha + \gamma_1 X_1 + \gamma_4 X_4 + \gamma_6 X_1 X_4\,,\\
\mathrm{M}_5: \quad y &= \alpha + \gamma_1 X_1 + \gamma_4 X_4\,.
\end{aligned}$$

Offensichtlich beschreibt das Modell M_2 den datengenerierenden Prozess am besten; für die wahren Koeffizienten ergibt sich in diesem Experiment $\alpha_{\text{wahr}} = 2.5$ und $\gamma_{\text{wahr}} = (-3, -0.3, 0, -2, 0, 1)'$. Die Resultate sind in Tabelle B.7 zu finden. Auch in diesem Szenario werden die Grundaussagen bezüglich der Schätzer und der Strategien zur Berücksichtigung fehlender Daten weitgehend bestätigt. Einzig der Vergleich der Qualität zwischen den Imputationsmethoden stellt sich etwas anders dar: Die Imputationen auf Basis des Amelia II-Pakets führen nicht durchgängig zu besseren Schätzungen als die kNN-Imputationen; es scheint als würde sich speziell in dieser komplexeren Situation ein verteilungsfreies Verfahren bewähren, da auch die Verluste bei den beiden Regressionsimputationen im Größenbereich der Amelia II-Methodik liegen.

Experiment 7: Höhere Komplexität II (kubischer Effekt)

Um die Komplexität weiter zu erhöhen, werden in diesem Experiment quadratische und kubische Effekte betrachtet. Hierfür erfolgt die Modifikation von $y \sim N(\mu_2, \sigma_0)$ mit $\mu_2 = 3 - 2X_4 - 2X_5 + 1.75X_5^2 - 0.2X_5^3$. Es werden sieben Kandidatenmodelle betrachtet:

$$\begin{aligned}
\mathrm{M}_1: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2 + \gamma_7 X_5^3\,,\\
\mathrm{M}_2: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2\,,\\
\mathrm{M}_3: \quad y &= \alpha + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5\,,\\
\mathrm{M}_4: \quad y &= \alpha + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2 + \gamma_7 X_5^3\,,\\
\mathrm{M}_5: \quad y &= \alpha + \gamma_2 X_2 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2 + \gamma_7 X_5^3\,,\\
\mathrm{M}_6: \quad y &= \alpha + \gamma_1 X_1 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2 + \gamma_7 X_5^3\,,\\
\mathrm{M}_7: \quad y &= \alpha + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_5^2 + \gamma_7 X_5^3\,.
\end{aligned}$$

Offensichtlich beschreibt das Modell M_7 den datengenerierenden Prozess am besten; für die wahren Koeffizienten ergibt sich in diesem Experiment $\alpha_{\text{wahr}} = 3$ und $\gamma_{\text{wahr}} = (0, 0, 0, -2, -2, 1.75, -0.2)'$. Die Resultate sind in Abbildung 6.3 illustriert und im Detail in Tabelle B.8 zu finden. In diesem Experiment sind die Resultate im Vergleich zum Grundszenario in vielen Punkten verschieden: Die grundlegenden Aussagen bezüglich der drei betrachteten FMA- bzw. FMS-Schätzer entsprechen in weiten Teilen nicht mehr denen der vorangegangenen sechs Experimente; kein Schätzer dominiert den anderen, die Qualität der Schätzungen hängt sowohl von der betrachteten Verlustfunktion als auch der gewählten Methodik zur Berücksichtigung der Problematik fehlender Daten ab. Abgesehen von der Complete Case Analyse für die Verlustfunktionen L_3 und L_4 liefert der FMA-Akaike-Schätzer meist, jedoch nicht immer, etwas bessere Ergebnisse als die anderen beiden Schätzer. Diese Unterschiede scheinen in vielen Fällen marginal zu sein, dies liegt jedoch vor allem an der starken Streuung der Ergebnisse und der damit verbundenen Skalierung in Abbildung 6.3. Eine genauere Aufsplittung der Ergebnisse am Beispiel der kNN-Imputation für die Verlustfunktion L_2 soll helfen, diesen Sachverhalt zu verstehen. Betrachtet wird dabei die Verteilung des Verlusts

$$\tilde{L}_2 \quad = \quad \sum_{j=1}^{p} (\hat{\tilde{\beta}}_j^{\star} - \hat{\tilde{\beta}}_j^{\text{org}})^2 \,|\, \text{Simulationslauf } r \tag{6.9}$$

über alle $\mathcal{R} = 500$ Simulationen. Tabelle 6.4 präsentiert die Ergebnisse. Betrachtet man

Methodik	Minimum	1. Quartil	Median	arith. Mittel	3. Quartil	Maximum
FMA-Akaike	0.0088	0.0843	0.1642	0.8719	0.3217	141.9000
FMA-Hansen	0.0244	0.1469	0.2682	1.3190	0.4867	186.4000
FMS-AIC	0.0042	0.1082	0.2124	0.9227	0.3954	156.5000

Tab. 6.4: Zusammenfassung der Resultate in Experiment 7 (Verlust $\tilde{L}_2$) in den 500 Simulationsläufen für den FMA-Akaike-, FMA-Hansen- und FMS-AIC-Schätzer nach Imputation mit der kNN-Methode

$\tilde{L}_2$ ab dem ersten Quartil bis hin zum Maximum, so ist klar, dass der FMA-Akaike-Schätzer in dieser Situation durchweg besser abschneidet als der FMS-AIC-Schätzer, der wiederum besser abschneidet als der FMA-Hansen-Schätzer. Das Durchführen paarweiser, einseitiger Tests nach Wilcoxon für verbundene Stichproben zeigt, dass sich die Verteilung der $\tilde{L}_2$-Werte dieser drei Schätzer zu einem Signifikanzniveau von 0.01 auch signifikant unterscheiden. In Abbildung 6.3 dagegen scheinen alle Ergebnisse nahezu identisch zu sein.

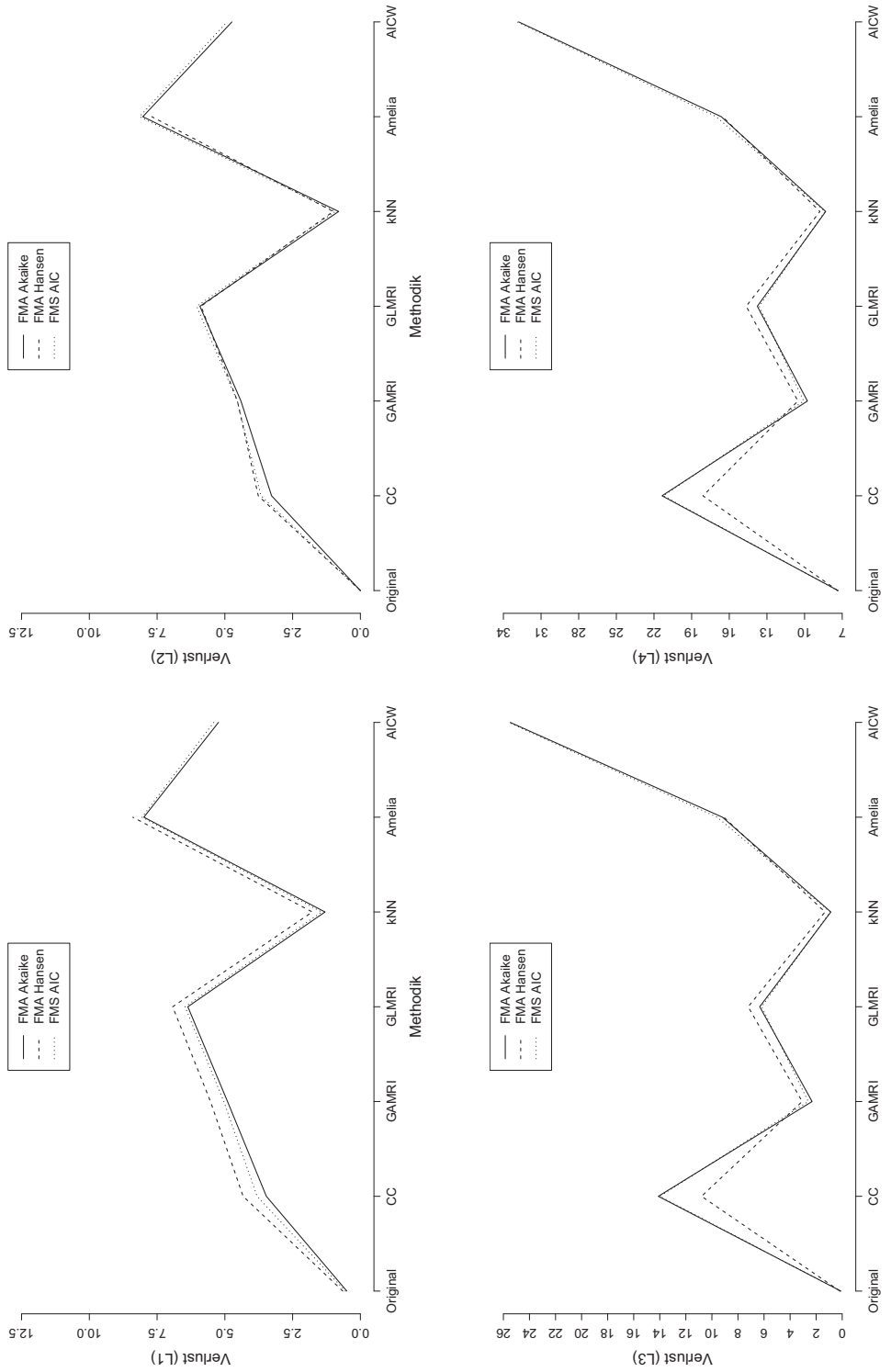


Abb. 6.3: Resultate in Experiment 7 bezüglich der Verlustfunktionen L_1, L_2, L_3 und L_4

Auch die Aussage aus den vorhergehenden Experimenten, dass ein Complete Case Analyse in der Regel zu den schlechtesten Ergebnissen führt, Imputationsansätze in der Regel zu den besten, kann hier nicht bestätigt werden. Für die Verlustfunktionen L_1 und L_2 kann einzig eine kNN-Imputation bessere Ergebnisse erzielen als die CC-Methodik. Die Imputationen des *R*-Pakets „Amelia II“ führen hier zu den schlechtesten Resultaten. Die Ansätze auf Basis des AIC_W führen zu keinen guten, jedoch vergleichsweise akzeptablen Ergebnissen. Betrachtet man dagegen die Verlustfunktionen L_3 und L_4, die die Vorhersagequalität der Schätzer beurteilen, so stellt man fest, dass alle Imputationsansätze eine Verbesserung bezüglich eine Complete Case Analyse erwirken. Nicht nur die kNN-Methodik, sondern auch die GAMRI-Methodik liefert dabei die besten Ergebnisse. Die Schätzer, die das gewichtete Akaike-Kriterium verwenden, führen zu vergleichsweise schlechten Ergebnissen.

Weitere Experimente und Ergebnisse

Die Ergebnisse weiterer Experimente, die in etwa die Varianz σ_0^2 variieren, andere Verteilungen für die Kovariablen annehmen bzw. über die Copula-Parameter und andere Copulas die Abhängigkeitsstruktur verändern, führen zu keinen relevanten Veränderungen bezüglich des Grundszenarios; die Resultate werden in dieser Arbeit daher nicht näher aufgeführt. Die Aussagen bezüglich der Varianzschätzungen nach Tabelle B.2 bestätigen sich auch für die Experimente 2-7, werden jedoch ebenfalls nicht mehr genauer erläutert, da dieser Thematik in Abschnitt 6.3 besondere Aufmerksamkeit zuteil wird.

6.2 Logistische Regression

In diesem Abschnitt werden am Beispiel des FMA-Akaike-Schätzers sowie des FMS-AIC-Schätzers die zu Beginn dieses Kapitels angeführten Fragestellungen im Kontext des logistischen Regressionsmodells

$$p_i = P(y_i = 1|\mathbf{x}_i) \quad = \quad F(\mu_i)$$

erörtert, wobei y einen binären Responsevektor repräsentiert, $\mu_i = \alpha + \mathbf{x}_i\gamma$, $F(.) = 1/\{1+\exp(-.)\}$, $\beta = (\alpha, \gamma')'$. Der MMA-Schätzer von Hansen (2007) wird im Gegensatz zum vorangegangenen Abschnitt nicht näher betrachtet, da seine Optimalitätseigenschaften – wie in Abschnitt 4.2.2 beschrieben – nur für das lineare Modell gelten. Betrachtet

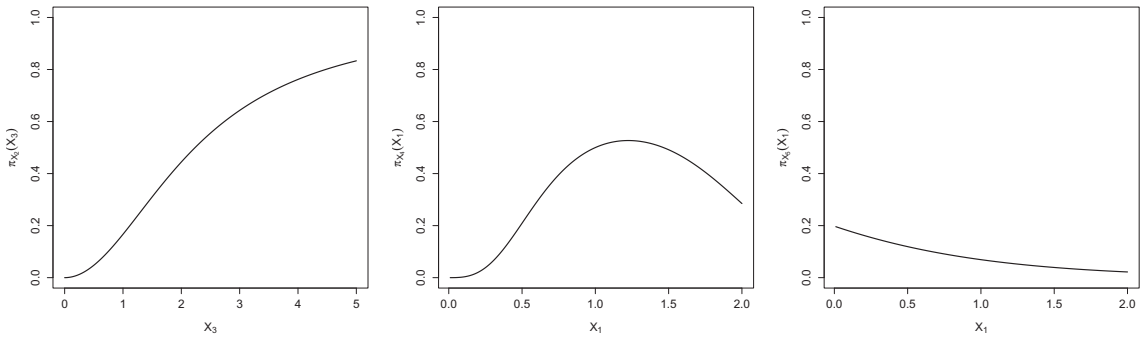
werden erneut $\mathcal{R} = 500$ Simulationsläufe, wobei für jeden Lauf je ein $n_{\text{tr}} \times p$ Trainingsdatensatz und ein $n_{\text{test}} \times p$ Testdatensatz erzeugt wird mit $n_{\text{tr}} = 450$, $n_{\text{test}} = 50$ und $p = 6$. Mit Hilfe eines Clayton-Copulas wird eine multivariate Verteilung für die insgesamt fünf Kovariablen mit den Randverteilungen $X_1 \sim Ga(0.25, 5)$, $X_2 \sim \log N(0.5, 0.5)$, $X_3 \sim Exp(1)$, $X_4 \sim \log N(0.5, 0.5)$ und $X_5 \sim B(1, 0.65)$ generiert. Die Verteilungen sind so gewählt, dass die Varianz der Kovariablen aus Gründen der Standardisierung immer in etwa Eins beträgt; der Copula-Parameter beträgt durchgehend $\theta_{\text{cop}} = 1.25$ und resultiert damit in einer mittelstarken Korrelation zwischen allen Kovariablen. Der Response y wird aus einer Binomialverteilung $B(1, p_0)$ gezogen mit $p_0 = 1/(1 + \exp(-\mu_0))$, wobei $\mu_0 = 3 - 2X_1 + 0.25X_2 - 3X_5$. Dies bedeutet, dass der Erwartungswert von y maßgeblich von den Variablen X_1, X_2 und X_5 bestimmt wird. Es gilt $\alpha_{\text{wahr}} = 3$ und $\gamma_{\text{wahr}} = (-2, 0.25, 0, 0, -3)'$ und damit $\beta_{\text{wahr}} = (\alpha_{\text{wahr}}, \gamma'_{\text{wahr}})'$; der komplette Datensatz wird als $D_{\text{sim2}} = \{y, X_1, X_2, X_3, X_4, X_5\}$ bezeichnet und wird in jedem Simulationslauf neu generiert. Anschließend werden unter Verwendung eines MAR-Fehlendmechanismus Werte von X_2, X_4 und X_5 gemäß der Fehlwahrscheinlichkeitsfunktionen

$$\begin{aligned}
\pi_{X_2}(X_3) &= 1 - \frac{1}{0.2X_3^2 + 1}\,, \qquad \pi_{X_4}(X_1) = 1 - \frac{1}{1 + X_1^3 \cdot \exp(1 - X_1^2)}\,, \\
\pi_{X_5}(X_1) &= 1 - \frac{1}{1 + \exp\{1 - 1.2(X_2 + 2)\}}\,,
\end{aligned}$$

als fehlend deklariert. Bei den $\mathcal{R} = 500$ Simulationsläufen fehlen damit im Mittel 18% der Beobachtungen bei X_2, 9% der Beobachtungen bei X_4 und 15% der Beobachtungen bei X_5. Die Gestalt der Fehlwahrscheinlichkeitsfunktionen ist in Abbildung 6.4 dargestellt. Um den Zusammenhang des Response mit den Kovariablen zu modellieren, werden fünf konkurrierende Kandidatenmodelle betrachtet:

$$\begin{aligned}
\text{M}_1: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5}\,, \\
\text{M}_2: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_5 x_{i5}\,, \\
\text{M}_3: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_4 x_{i4} + \gamma_5 x_{i5}\,, \\
\text{M}_4: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_5 x_{i5}\,, \\
\text{M}_5: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_5 x_{i5}\,.
\end{aligned}$$

Offensichtlich ist das Modell M_4, das die Kovariablen X_1, X_2 und X_5 enthält, am besten geeignet um den oben formulierten, datengenerierenden Prozess zu beschreiben. Um nun sowohl die Effekte der Modellselektionsunsicherheit als auch die Problematik fehlender



(a) $\pi_{X_2}(X_3)$ (b) $\pi_{X_4}(X_1)$ (c) $\pi_{X_5}(X_1)$

Abb. 6.4: Gestalt der verwendeten Fehlwahrscheinlichkeitsfunktionen

Daten zu berücksichtigen, werden im Folgenden die 14 FMA-Akaike- bzw. FMS-AIC-Schätzer, wie im vorhergehenden Abschnitt in Tabelle 6.1 erläutert, betrachtet. Die Qualität dieser Schätzer wird anhand sieben verschiedener Verlustfunktionen beurteilt: Dies sind die bereits vorgestellten Verlustfunktionen L_1, L_2 und L_3; ferner die geschätzte, mittlere Fehlklassifikationsrate

$$L_5 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} I(\hat{\tilde{y}}^{\star}_{i,r} \neq y_{i,r,\text{wahr}}) \right\}, \tag{6.10}$$

wobei $I(\cdot)$ die Indikatorfunktion und $\hat{\tilde{y}}^{\star}_{i,r}$ den Vorhersagewert für y_i im r-ten Simulationslauf bezeichnet. Dieser basiert auf der Regel, dass $\hat{\tilde{y}}^{\star}_{i,r} = 1$, wenn $\hat{\tilde{\alpha}}^{\star}_r + \mathbf{x}_{i,r}\hat{\tilde{\gamma}}^{\star}_r \geq 0$ und $\hat{\tilde{y}}^{\star}_{i,r} = 0$, wenn $\hat{\tilde{\alpha}}^{\star}_r + \mathbf{x}_{i,r}\hat{\tilde{\gamma}}^{\star}_r < 0$. Betrachtet wird auch der MSE bezüglich p im Testdatensatz,

$$L_6 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{\tilde{p}}^{\star}_{i,r} - p_{i,r,\text{wahr}})^2 \right\}, \tag{6.11}$$

wobei $\hat{\tilde{p}}^{\star}_{i,r} = 1/\{1+\exp(-\hat{\tilde{\mu}}^{\star}_{i,r})\}$, $\hat{\tilde{\mu}}^{\star}_{i,r} = \hat{\tilde{\alpha}}^{\star}_r + \mathbf{x}_{i,r}\hat{\tilde{\gamma}}^{\star}_r$, und $p_{i,r,\text{wahr}} = 1/\{1+\exp(-\mu_{i,r,\text{wahr}})\}$, $\mu_{i,r,\text{wahr}} = \alpha_{\text{wahr}} + \mathbf{x}_{i,r}\gamma_{\text{wahr}}$, sowie die mittlere Kullback-Leibler-Distanz,

$$L_7 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \text{KL}\left(p_{i,r,\text{wahr}}; \hat{\tilde{p}}^{\star}_{i,r}\right) \right\}, \tag{6.12}$$

mit $\mathrm{KL}(p_{i,r,\mathrm{wahr}}; \hat{\bar{p}}^{\star}_{i,r}) = p_{i,r,\mathrm{wahr}} \cdot \ln\{p_{i,r,\mathrm{wahr}}/\hat{\bar{p}}^{\star}_{i,r}\} + (1-p_{i,r,\mathrm{wahr}}) \cdot \ln\{(1-p_{i,r,\mathrm{wahr}})/(1-\hat{\bar{p}}^{\star}_{i,r})\}$. Zur Beurteilung der Vorhersagequalität wird ferner die mittlere, geschätzte Fläche unter der entsprechenden ROC-Kurve, also die mittlere geschätzte *Area Under Curve* (AUC) im Testdatensatz betrachtet:

$$L_8 \;=\; \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \widehat{\mathrm{AUC}}_r \,|\, n_{\mathrm{test}}, \hat{\bar{p}}^{\star}_r \right\} . \tag{6.13}$$

Für die Schätzung der AUC-Werte wird das *R*-Paket „ROCR“ verwendet, Details dazu finden sich bei Sing et al. (2005).

Resultate

Eine Auswahl an Resultaten ist in Abbildung 6.5 dargestellt, die detaillierten Ergebnisse befinden sich in Tabelle B.9 in Anhang B. Es lassen sich folgende Ergebnisse konstatieren:

- Unabhängig von der gewählten Methodik zur Berücksichtigung fehlender Werte und unabhängig von der betrachteten Verlustfunktion liefert der FMA-Akaike-Schätzer fast ausschließlich bessere Resultate als der FMS-AIC-Schätzer. Diese Aussage war aufgrund der Erkenntnisse aus den Simulationen in Abschnitt 6.1 zu erwarten.

- Die Strategien zur Berücksichtigung fehlender Werte versprechen unterschiedliche Erfolgsaussichten: Bezüglich der Verlustfunktionen L_1, L_3, L_6 und L_7 führen alle vier Imputationsansätze sowohl für den Selektions- als auch den Mittelungsschätzer zu deutlich besseren Ergebnissen als eine entsprechende Complete Case Analyse. Diese Aussage gilt bei L_2 nicht für die Imputation auf Basis des *R*-Pakets „Amelia II“ bzw. bei L_5 nicht für die verallgemeinerten Regressionsimputationen (GAMRI, GLMRI) und die kNN-Methodik. Generell scheinen die Imputationen von Amelia II am besten abzuschneiden; die anderen drei Imputationsmethoden liegen in einer ähnlichen Größenordnung, wobei die k-Nächste-Nachbarn-Imputation oft geringfügig bessere Resultate erzielt als der GAMRI-Algorithmus, der wiederum etwas bessere Resultate als der GLMRI-Algorithmus liefert. Die Gewichtungsansätze auf Basis des AIC_W können gegenüber einer Complete Case Analyse durchweg keine Verbesserung erzielen.

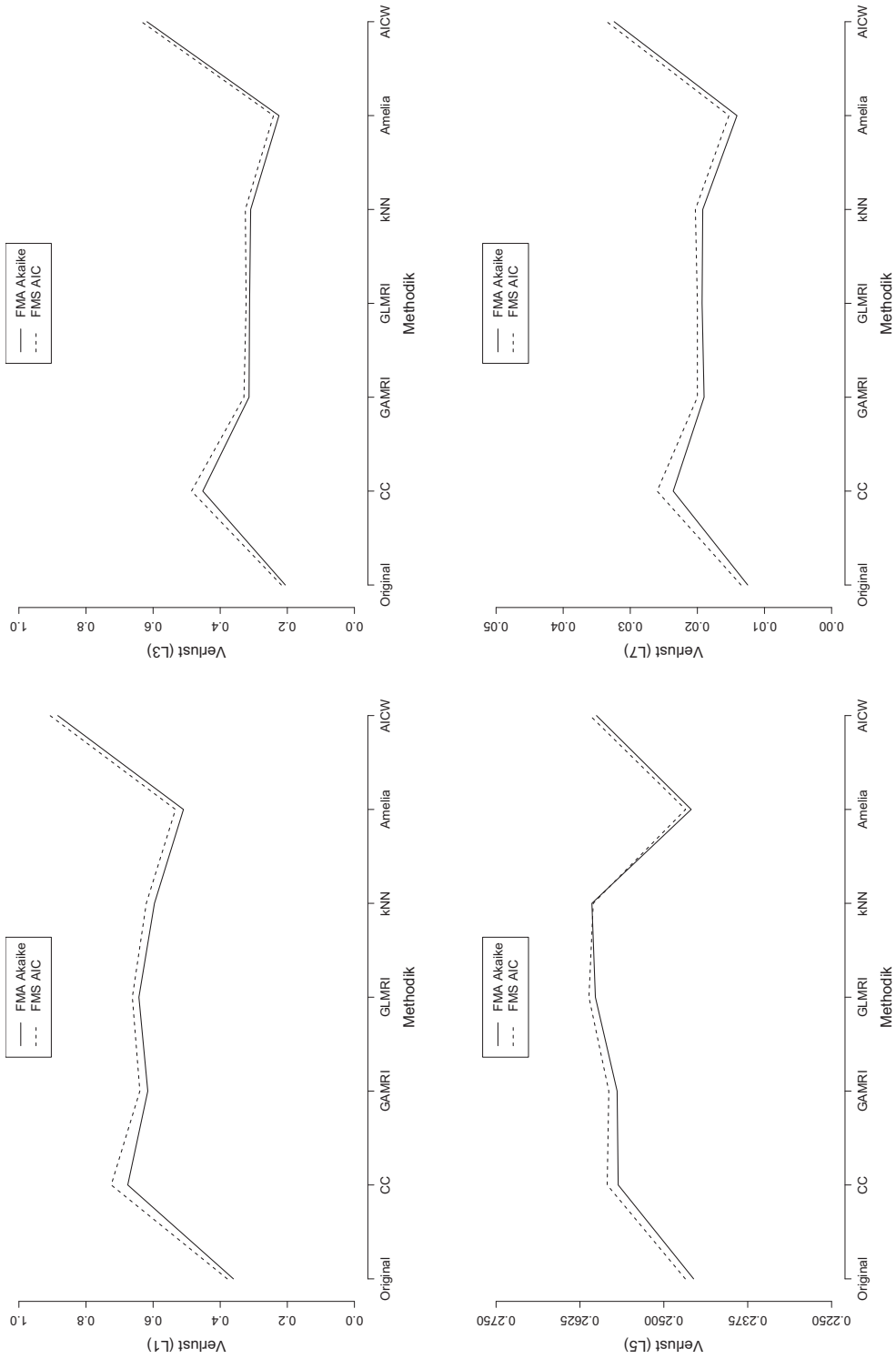


Abb. 6.5: Resultate im Grundszenario bezüglich der Verlustfunktionen L_1, L_3, L_5 und L_7

- Die Verlustfunktion L_8, die für die 500 Simulationsläufe die mittlere Fläche unter der ROC-Kurve bestimmt, nimmt eine Sonderstellung ein, da im Gegensatz zu den anderen Verlustfunktionen offensichtlich höhere und nicht niedrigere Werte eine bessere Qualität der zugehörigen Schätzung implizieren. In Tabelle B.9 erkennt man hierfür die Dominanz des FMA-Akaike-Schätzers gegenüber des FMS-AIC-Schätzers und die Vorteile der Imputationsmethoden gegenüber der CC- bzw. AIC_W-Methodik. Die Unterschiede sind jedoch meist gering; dies verdeutlicht auch Abbildung 6.6, in der die Verteilung der ROC-Kurven über die $\mathcal{R} = 500$ Simulationsläufe exemplarisch für die CC- und GAMRI-Methodik beim FMA-Akaike- sowie FMS-AIC-Schätzer betrachtet werden: Es ist fast kein Unterschied zwischen den Abbildungen 6.6a, 6.6b, 6.6c und 6.6d zu erkennen; die Relevanz dieser Ergebnisse wird auch untenstehend diskutiert.

- Wie in der vorhergehenden Simulation in Abschnitt 6.1 stellt sich auch hier die Frage, ob die auf Basis der Verlustfunktionen getroffenen Aussagen über die meisten der Simulationsläufe beobachtet werden können oder nicht. Zur Klärung dieses Sachverhalts soll daher im Folgenden die Verteilung der Verluste

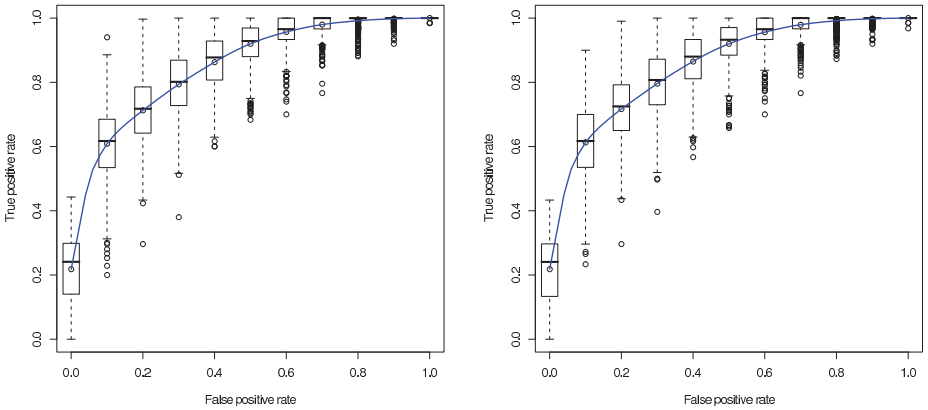
$$\tilde{L}_3 \;=\; \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{\tilde{\mu}}_i^{\star} - \mu_{i,\text{wahr}})^2 \,|\, \text{Simulationslauf } r \tag{6.14}$$

und

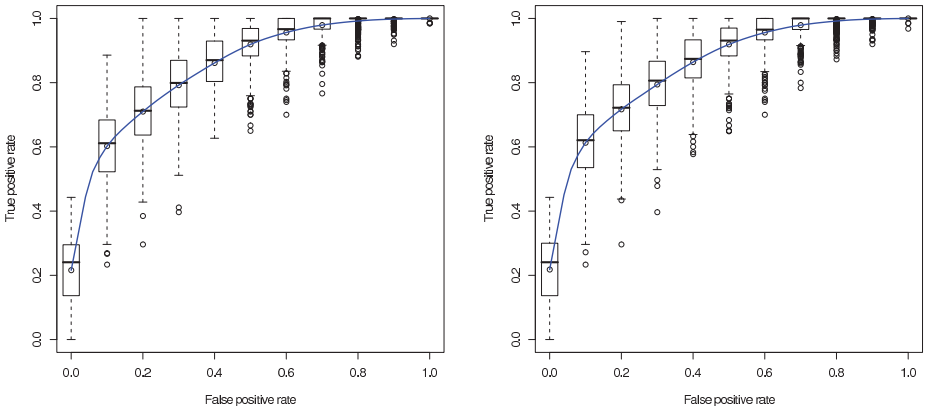
$$\tilde{L}_8 \;=\; \widehat{\text{AUC}} \,|\, \text{Simulationslauf } r \tag{6.15}$$

exemplarisch für die CC- und GAMRI-Methodik beim FMA-Akaike- sowie FMS-AIC-Schätzer über die $\mathcal{R} = 500$ Simulationsläufe betrachtet werden, vergleiche Tabelle 6.5. Ein Blick vom ersten Quartil bis hin zum dritten Quartil bestätigt für $\tilde{L}_3$ sowohl die Überlegenheit der Modellmittelung gegenüber der Modellselektion als auch der GLMRI-Methodik bezüglich einer CC-Analyse. Die Verteilung von $\tilde{L}_8$ ist dagegen weniger klar; eine eindeutige Aussage ist auf den ersten Blick nicht möglich. Um die Relevanz dieser Zahlen besser einordnen zu können, werden mehrere einseitige Wilcoxon-Vorzeichen-Rangtests für verbundene Stichproben durchgeführt: Dabei wird stets geprüft, ob sich die Lage der Verteilungsfunktionen der Verluste für die entsprechenden FMA- und FMS-Schätzer bzw. CC- und GAMRI-Schätzer bezüglich eines Signifikanzniveaus von 5% signifikant unterscheiden. Die Resultate befinden sich in Tabelle 6.6; die Alternativhypothese ist dabei

stets $H_1 : F_X < F_Y$. Es zeigt sich, dass – auch wenn die Unterschiede gering sind – die Verluste $\tilde{L}_3$ für den FMA-Akaike-Schätzer signifikant unter denen des FMS-AIC-Schätzers liegen bzw. für die GLMRI-Methodik signifikant geringer sind als für die Complete Case Analyse. Für die Verluste $\tilde{L}_8$ gilt dasselbe mit Ausnahme der FMS-AIC- und FMA-Akaike-Schätzungen für die GAMRI-Methodik, hier wird die Nullhypothese beibehalten. Die oben getroffenen Grundaussagen können also in weiten Teilen über alle Simulationsläufe hinweg bestätigt werden.



(a) FMA-Akaike (CC) (b) FMA-Akaike (GLMRI)



(c) FMS-AIC (CC) (d) FMS-AIC (GLMRI)

Abb. 6.6: Veranschaulichung des Verlustes $\tilde{L}_8$ für einige ausgewählte Schätzer über die entsprechenden ROC-Kurven

Um nun die Qualität der Varianzschätzungen der FMA- bzw. FMS-Schätzer zu beurteilen, wird erneut der empirische Standardfehler (6.7) für jedes β_j, $j = 0, \ldots, 5$, berechnet

	Methodik		Min.	1. Quartil	Median	arith. Mittel	3. Quartil	Max.
$\tilde{L}_3$	FMA-Akaike	(CC)	0.0062	0.1083	0.2294	0.4520	0.4945	7.4040
	FMA-Akaike	(GLMRI)	0.0101	0.0926	0.2002	0.3125	0.3885	3.2790
	FMS-AIC	(CC)	0.0014	0.1335	0.2673	0.4855	0.5435	7.4920
	FMS-AIC	(GLMRI)	0.0035	0.0967	0.2000	0.3228	0.4098	3.2160
$\tilde{L}_8$	FMA-Akaike	(CC)	0.6635	0.8235	0.8589	0.8549	0.8914	0.9813
	FMA-Akaike	(GLMRI)	0.6827	0.8229	0.8629	0.8560	0.8945	0.9796
	FMS-AIC	(CC)	0.6651	0.8211	0.8571	0.8533	0.8896	0.9643
	FMS-AIC	(GLMRI)	0.6667	0.8251	0.8631	0.8557	0.8926	0.9779

Tab. 6.5: Auswahl an Resultaten für das Grundszenario (Verlust $\tilde{L}_3$ und $\tilde{L}_8$) in den 500 Simulationsläufen bezüglich des FMA-Akaike-Schätzers und FMS-AIC-Schätzers abhängig von der Behandlung der fehlenden Werte

F_X			F_Y			p-Wert
FMA-Akaike	(GLMRI)	(L_3)	FMA-Akaike	(CC)	(L_3)	< 0.001
FMA-Akaike	(CC)	(L_3)	FMS-AIC	(CC)	(L_3)	< 0.001
FMS-AIC	(GLMRI)	(L_3)	FMS-AIC	(CC)	(L_3)	< 0.001
FMA-Akaike	(GLMRI)	(L_3)	FMS-AIC	(GLMRI)	(L_3)	< 0.001
FMA-Akaike	(CC)	(L_8)	FMA-Akaike	(GLMRI)	(L_8)	0.017
FMS-AIC	(CC)	(L_8)	FMA-Akaike	(CC)	(L_8)	0.008
FMS-AIC	(CC)	(L_8)	FMS-AIC	(GLMRI)	(L_8)	0.002
FMS-AIC	(GLMRI)	(L_8)	FMA-Akaike	(GLMRI)	(L_8)	0.163

Tab. 6.6: Ergebnisse der Wilcoxon-Vorzeichen-Rangtests für verbundene Stichproben zum Vergleich der Verteilungen der Verluste $\tilde{L}_3$ und $\tilde{L}_8$ für ausgewählte Schätzer

und mit der mittleren Schätzung des Standardfehlers nach Buckland, Burnham und Anderson (1997) gemäß (6.8) verglichen. Tabelle B.10 präsentiert die Resultate. Es lassen sich folgende Ergebnisse konstatieren:

- Die Verwendung des FMA-Akaike-Schätzers führt generell zu höheren Varianzschätzungen gemäß (6.8) als die Verwendung des FMS-AIC-Schätzers. Dies war zu erwarten, da durch das Kombinieren mehrerer Modelle explizit die Unsicherheit bezüglich der Selektion beachtet wird.

- Betrachtet man den FMA-Akaike-Schätzer für die Originaldaten (ohne fehlende Werte) bzw. für die Complete Case Analyse, so ist zu erkennen, dass die Schätzung des Standardfehlers gemäß (6.8) dem zugehörigen empirischen Standardfehler für alle β_j meist sehr nahe liegt, in einigen Fällen jedoch etwas geringer ist. Für den FMS-AIC-Schätzer fallen die Schätzungen der Standardfehler geringer aus und liegen ausnahmslos unter dem empirischen Standardfehler.

- Die Verwendung von Imputationen und damit der Strategien „Selektion nach Imputation“ und „Mittelung nach Imputation“ führen zu einer Unterschätzung der Varianz; der geschätzte Standardfehler (6.7) ist ausnahmslos kleiner als der empirische Standardfehler (6.8). Wie in den Abschnitten 5.1.2, 5.2.2 und 6.1 bereits diskutiert, mag dies in erster Linie daran liegen, dass die Unsicherheit bezüglich der Imputation in (6.8) nicht berücksichtigt wird. Eine mögliche Adjustierung mit Hilfe multipler Imputationen wird in Abschnitt 6.3 diskutiert.
- Die Verwendung des AIC_W zur Selektion bzw. zur Konstruktion von FMA-Gewichten führt ebenfalls zu einer Unterschätzung der Varianz. Potentiell ist dies auf die Unsicherheit in der Wahl des Glättungsparameters bzw. der Kovariablen für das GAM zur Schätzung der Gewichte (5.5) zurückzuführen.
- Diese Aussagen decken sich fast ausnahmslos mit denen des vorangegangenen Abschnitts 6.1.

Es stellt sich die Frage, wie sensitiv die in diesem Abschnitt getroffenen Aussagen bezüglich des Simulationssettings sind; hierfür werden im Folgenden kritische Annahmen der Simulation variiert und darauf aufbauend die bisherigen Resultate überprüft.

Experiment 2: Veränderung des Fehlendmechanismus

In der oben betrachteten Simulation wird von einem MAR-Fehlendmechanismus ausgegangen. Es stellt sich die Frage, ob bei komplett zufälligem Fehlen der Daten, also einem MCAR-Fehlendmechanismus, die Aussagen bezüglich des FMA- und des FMS-Schätzers bestehen bleiben und ob die Verwendung von Korrekturverfahren zur Berücksichtigung der Problematik fehlender Daten bezüglich einer CC-Analyse einen Gewinn erbringt oder nicht. Hierfür wird erneut das Grundszenario betrachtet, jedoch unter Verwendung der konstanten Fehlwahrscheinlichkeitsfunktionen $\pi_{X_2} = \pi_{X_4} = \pi_{X_5} = 0.1$. Die Resultate befinden sich in Tabelle B.11 im Anhang. Die Kernaussagen werden dabei weitgehend bestätigt: Der FMA-Akaike-Schätzer liefert in der Regel bessere Ergebnisse als der FMS-AIC-Schätzer, alle Imputationsansätze erbringen eine Verbesserung gegenüber einer CC-Analyse und innerhalb der Imputationsmethoden versprechen diejenigen auf Basis des Amelia II-Pakets die besten Resultate; einzig die Aussage bezüglich der Schätzungen unter Verwendung des AIC_W muss revidiert werden. Die Qualität dieser Schätzungen ist nicht mehr durchgehend schlechter als die Qualität der Schätzungen auf Basis einer Complete Case Analyse, meist liegen sie in einer ähnlichen Größenordnung.

Experiment 3: Geringere Anzahl an fehlenden Werten

Um zu sehen, ob die Kernaussagen auch für eine geringere Anzahl an fehlenden Werten bestätigt werden können, werden die Fehlwahrscheinlichkeitsfunktionen des Grundszenarios teilweise modifiziert: Im Folgenden wird angenommen, dass die Kovariable X_4 vollständig beobachtet wird, also $\pi_{X_4}(X_1) = 0$ gilt und die Fehlwahrscheinlichkeit für X_2 gemäß

$$\tilde{\pi}_{X_2}(X_3) \;=\; 1 - \frac{1}{0.1X_3^2 + 1}$$

bestimmt wird. Die Fehlwahrscheinlichkeitsfunktion $\pi_{X_5}(X_1)$ zur Generierung fehlender Werte für X_5 bleibt bestehen. Für die $\mathcal{R} = 500$ Simulationsläufe werden dabei im Mittel 12% fehlende Werte für X_2 und 15% fehlende Werte für X_5 beobachtet. Die Resultate dieses Experiments befinden sich in Tabelle B.12. Auch hier werden die Aussagen des Grundszenarios weitgehend bestätigt. Man erkennt erneut die besten Resultate für die „Mittelung nach Imputation"-Strategie; In Analogie zu den beiden vorangegangenen Experimenten schneiden dabei die Amelia II-Imputationen erneut am besten ab, mit einer Ausnahme: Für die Verlustfunktion L_2, die die FMA- bzw. FMS-Schätzer mit den entsprechenden Schätzern unter Verwendung der Originaldaten vergleicht, können (wie auch in einigen anderen Experimenten bereits zu beobachten) die verallgemeinerten Regressionsimputationen GAMRI und GLMRI respektable Ergebnisse vorweisen. Es scheint, als führe eine solche Imputationsstrategie – wie beabsichtigt – zu einer adäquaten Wiedergabe der vorliegenden Datenstruktur.

Experiment 4: Berücksichtigung zufälliger Effekte

Dieses Experiment unterscheidet sich vom Grundszenario in der Annahme, dass $y \sim B(1, p_0 + \epsilon_0)$, $\epsilon_0 \sim N(0, 1)$, also zufällige Effekte vorliegen. Die Resultate für dieses Szenario befinden sich in Tabelle B.13. Weiterhin ist der FMA-Akaike-Schätzer dem FMS-AIC-Schätzer bezüglich aller Verlustfunktionen deutlich überlegen. Bei Betrachtung der verschiedenen Strategien zur Berücksichtigung der fehlenden Werte ergibt sich sowohl für die Modellmittelung als auch die Modellselektion ein wesentlicher Unterschied im Vergleich zu den ersten drei Experimenten: Die Qualität der vier Imputationsansätze ist genau gegenläufig zu beurteilen; die besten Resultate erhält man bei Betrachtung der Verlustfunktion L_1 unter Verwendung einer verallgemeinerten Regressionsimputation mit generalisierten linearen Modellen (GLMRI), gefolgt von der GAMRI-Methodik, der

kNN-Methodologie und den Amelia II-Imputationen. Das letztgenannte Vorgehen verspricht dabei nicht einmal Verbesserungen bezüglich einer Complete Case Analyse. Für die anderen Verlustfunktionen gelten ähnliche Aussagen: Die Imputationen von Amelia II sind stets am schlechtesten, die der Regressionsimputationen meist am besten. Diese Ergebnisse verdeutlichen noch einmal die Herausforderung der Wahl geeigneter Imputationen und die Notwendigkeit von Sensitivitätsanalysen.

Experiment 5: Höhere Komplexität I (logarithmischer Effekt)

Zur Modellierung höherer Komplexität wird in diesem Experiment zusätzlich ein logarithmischer Effekt betrachtet. Hierfür erfolgt die Modifikation von $y \sim B(1, p_1)$, $p_1 = 1/(1 + \exp(-\mu_1))$ mit $\mu_1 = 3 - 2X_1 + 0.25X_2 - 3X_5 + 0.35\ln(X_1)$. Die Fehlwahrscheinlichkeitsfunktionen sind dieselben wie im Grundszenario; der Stichprobenumfang im Trainingsdatensatz wird zu $n_{\text{train}} = 750$ verändert, im Testdatensatz beträgt er weiterhin $n_{\text{test}} = 50$. Es werden die folgenden Kandidatenmodelle betrachtet:

$$
\begin{aligned}
\mathrm{M}_1: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 \ln x_{i1}\,, \\
\mathrm{M}_2: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_5 x_{i5} + \gamma_6 \ln x_{i1}\,, \\
\mathrm{M}_3: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_5 x_{i5}\,, \\
\mathrm{M}_4: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_5 x_{i5} + \gamma_6 \ln x_{i1}\,, \\
\mathrm{M}_5: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_5 x_{i5}\,.
\end{aligned}
$$

Die Resultate für dieses Experiment befinden sich in Tabelle B.14. Offensichtlich weisen die Ergebnisse mehr Ähnlichkeiten mit Experiment 4 als mit den ersten drei Experimenten auf: Die Schätzer auf Basis des AIC_W, wie auch die Schätzer unter Verwendung der Amelia II-Imputationen, können die Resultate einer Complete Case Analyse meist nicht verbessern; sie liegen oft in einer ähnlichen Größenordnung und modellieren die Datensituation nicht adäquat. Die anderen drei Imputationsmethoden liefern durchgängig bessere Resultate als die CC-Analyse, mit Ausnahme für die Verlustfunktion L_5. Dies unterstreicht erneut, dass die verallgemeinerten Regressionsimputationen GLMRI und GAMRI sowie die k-Nächste-Nachbarn-Methodologie sehr häufig zu qualitativ ähnlichen Schätzern für die Strategien „Selektion nach Imputation“ und „Mittelung nach Imputation“ führen und sich – teils positiv, teils negativ – von den Amelia II-Imputationen unterschieden.

Experiment 6: Höhere Komplexität II (kubischer Effekt)

In Analogie zu Experiment 7 in Abschnitt 6.1 wird in diesem Experiment eine höhere Komplexität über die Betrachtung quadratischer und kubischer Effekte modelliert. Dazu wird der Response gemäß $y \sim B(1, p_2)$, $p_2 = 1/(1 + \exp(-\mu_2))$ mit $\mu_2 = -2 + 0.3X_4 + 0.7X_4^2 - 0.2X_4^3 + 2X_5$ generiert. Die Fehlwahrscheinlichkeitsfunktionen sind dieselben wie im Grundszenario; der Stichprobenumfang im Trainingsdatensatz wird zu $n_{\text{train}} = 950$ verändert, im Testdatensatz beträgt er weiterhin $n_{\text{test}} = 50$. Es werden die folgenden Kandidatenmodelle betrachtet:

$$
\begin{aligned}
\mathrm{M}_1: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2 + \gamma_7 x_{i4}^3\,, \\
\mathrm{M}_2: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2 + \gamma_7 x_{i4}^3\,, \\
\mathrm{M}_3: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_2 x_{i2} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2 + \gamma_7 x_{i4}^3\,, \\
\mathrm{M}_4: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 x_{i1} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2 + \gamma_7 x_{i4}^3\,, \\
\mathrm{M}_5: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2 + \gamma_7 x_{i4}^3\,, \\
\mathrm{M}_6: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i4}^2, \\
\mathrm{M}_7: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_4 x_{i4} + \gamma_5 x_{i5}\,.
\end{aligned}
$$

Die Resultate für dieses Experiment befinden sich in Tabelle B.15. Die Qualität der Strategien zur Berücksichtigung fehlender Werte variiert in diesem Experiment sehr stark mit der betrachteten Verlustfunktion. Einen ersten interessanten Aspekt liefert dabei die FMA- bzw. FMS-Schätzung auf Basis der Originaldaten, die im Gegensatz zu allen vorher durchgeführten Experimenten – abhängig von der betrachteten Verlustfunktion – nicht mehr die besten Ergebnisse liefert. Auch wenn diese Schätzungen nur als Referenz dienen und deshalb bisher nicht näher kommentiert worden sind, so zeigt sich schon hier die höhere Komplexität des Szenarios und die damit verbundene stärkere Instabilität der Schätzungen. Wie schon häufig beobachtet und in Experiment 3 auch näher kommentiert, lässt sich unter den Imputationsstrategien eine starke Diskrepanz in den Resultaten bezüglich L_1 und L_2 beobachten: Für L_1 liefert die Amelia II-Methodik die besten Ergebnisse, die kNN-Methodik die schlechtesten; für die Verlustfunktion L_2, die die FMA- bzw. FMS-Schätzer mit den entsprechenden Schätzern unter Verwendung der Originaldaten vergleicht, resultiert die Verwendung der Regressionsimputationen bzw. der kNN-Imputation in deutlich besseren Ergebnissen als die Amelia II-Imputation, die in diesem Kontext nicht einmal die Ergebnisse einer Complete Case Analyse verbessern kann. Bezüglich der Verlustfunktionen L_3-L_8, die die Vorhersagequalität der Schätzungen

beurteilen, existieren zahlreiche marginale Unterschiede; in der Regel führen jedoch alle Imputationsstrategien zu einer Verbesserung gegenüber einer CC-Analyse und meist liefern die Amelia II-Imputationen geringfügig bessere Resultate als die drei anderen Imputationsstrategien; die Schätzer auf Basis des AIC_W können die Resultate einer Complete Case Analyse erneut bezüglich keiner Verlustfunktion verbessern.

Weitere Experimente

Die Ergebnisse weiterer Experimente, die in etwa über die Copula-Parameter bzw. andere Copulas die Abhängigkeitsstruktur verändern oder andere Verteilungen für die Kovariablen annehmen, führen zu keinen relevanten Veränderungen bezüglich des Grundszenarios; die Resultate werden in dieser Arbeit daher nicht näher aufgeführt. Die Aussagen bezüglich der Varianzschätzungen nach Tabelle B.10 bestätigen sich auch für die Experimente 2-6, werden jedoch ebenfalls nicht mehr genauer erläutert, da dieser Thematik im folgendem Abschnitt 6.3 besondere Aufmerksamkeit zuteil wird.

6.3 Die Auswirkungen multipler Imputation

Die beiden vorangegangenen Abschnitte haben gezeigt, dass die Verwendung von Imputationen die Qualität der betrachteten FMA- wie auch FMS-Schätzer im Vergleich zu einer einfachen Complete Case Analyse oder Adjustierungen auf Basis des AIC_W in der Regel verbessert. Es zeigt sich jedoch auch, dass eine solche Strategie zu einer Unterschätzung der Varianz führen kann. Dies liegt offensichtlich an der Imputationsunsicherheit, die bei der Verwendung nicht-multipler Imputationen vernachlässigt wird. Es stellt sich die Frage, ob korrekte multiple Imputationen (etwa unter Verwendung des Amelia II-Pakets für die statistische Software R) diese Problematik beheben können. Die in diesem Abschnitt durchgeführten Monte-Carlo-Simulationen orientieren sich sehr stark an Experiment 6 aus Abschnitt 6.1. Im Fokus steht damit das lineare Regressionsmodell

$$y \;=\; X\beta + \epsilon\,, \qquad \epsilon \sim N(0, \sigma^2 I)\,,$$

mit $\beta = (\alpha, \gamma')'$. Betrachtet werden erneut $\mathcal{R} = 500$ Simulationsläufe, wobei für jeden Lauf je ein $n_{\mathrm{tr}} \times p$ Trainingsdatensatz und ein $n_{\mathrm{test}} \times p$ Testdatensatz erzeugt

wird mit $n_{\text{tr}} = 450$, $n_{\text{test}} = 50$ und $p = 6$. Mit Hilfe eines Clayton-Copulas wird eine multivariate Verteilung für die insgesamt fünf Kovariablen mit den Randverteilungen $X_1 \sim N(0.5, 1)$, $X_2 \sim \log N(0.5, 0.5)$, $X_3 \sim Weibull(1.75, 1.9)$, $X_4 \sim B(1, 0.3)$ und $X_5 \sim Ga(0.25, 2)$ generiert. Die Verteilungen sind so gewählt, dass die Varianz der Kovariablen aus Gründen der Standardisierung immer in etwa Eins beträgt; der Copula-Parameter beträgt durchgehend $\theta_{\text{cop}} = 1$ und resultiert damit in einer mittelstarken Korrelation zwischen allen Kovariablen. Der Response y wird aus einer Normalverteilung $N(\mu_1, \sigma_0)$ gezogen mit $\mu_1 = 2.5 - 3X_1 - 0.3X_2 - 2X_4 + X_1X_4$ und $\sigma_0 = \exp(1)$. Dies bedeutet, dass der Erwartungswert von y maßgeblich von den Variablen X_1, X_2 und X_4, insbesondere auch über die Interaktion X_1X_4, bestimmt wird. Daher gilt $\alpha_{\text{wahr}} = 2.5$ und $\gamma_{\text{wahr}} = (-3, -0.3, 0, -2, 0, 1)'$ und damit $\beta_{\text{wahr}} = (\alpha_{\text{wahr}}, \gamma'_{\text{wahr}})'$; der komplette Datensatz wird als $D_{\text{sim3}} = \{y, X_1, X_2, X_3, X_4, X_5\}$ bezeichnet und wird in jedem Simulationslauf neu generiert. Anschließend werden unter Verwendung eines MAR-Fehlendmechanismus Werte von X_1, X_4 und X_5 gemäß der Fehlwahrscheinlichkeitsfunktionen

$$\begin{aligned} \pi_{X_1}(y) &= 1 - \frac{1}{0.0225y^2 + 1}\,, \qquad \pi_{X_4}(X_2) = 1 - \frac{1}{1 + 0.02X_2^3}\,, \\ \pi_{X_5}(X_3) &= 1 - \frac{1}{1 + \exp\{1 - 2X_3\}}\,, \end{aligned}$$

als fehlend deklariert. Bei den $\mathcal{R} = 500$ Simulationsläufen fehlen damit im Mittel 21% der Werte bei X_1, 15% der Werte bei X_4 und 18% der Werte bei X_5. Um den Zusammenhang von y und den Kovariablen zu modellieren, werden sieben konkurrierende Kandidatenmodelle betrachtet:

$$\begin{aligned} \text{M}_1: \quad y &= \alpha + \gamma_1X_1 + \gamma_2X_2 + \gamma_3X_3 + \gamma_4X_4 + \gamma_5X_5 + \gamma_6X_1X_4\,, \\ \text{M}_2: \quad y &= \alpha + \gamma_1X_1 + \gamma_2X_2 + \gamma_3X_3 + \gamma_4X_4 + \gamma_6X_1X_4\,, \\ \text{M}_3: \quad y &= \alpha + \gamma_1X_1 + \gamma_2X_2 + \gamma_4X_4 + \gamma_5X_5 + \gamma_6X_1X_4\,, \\ \text{M}_4: \quad y &= \alpha + \gamma_1X_1 + \gamma_2X_2 + \gamma_4X_4 + \gamma_6X_1X_4\,, \\ \text{M}_5: \quad y &= \alpha + \gamma_1X_1 + \gamma_2X_2 + \gamma_4X_4\,, \\ \text{M}_6: \quad y &= \alpha + \gamma_1X_1 + \gamma_4X_4 + \gamma_6X_1X_4\,, \\ \text{M}_7: \quad y &= \alpha + \gamma_1X_1 + \gamma_4X_4\,. \end{aligned}$$

Offensichtlich ist das Modell M_4 am besten geeignet, um den datengenerierenden Prozess zu beschreiben. Um nun sowohl die Effekte der Modellselektionsunsicherheit als auch die Problematik fehlender Daten zu berücksichtigen, werden analog zu den vorangegangenen

beiden Abschnitten sowohl die 20 Schätzer, wie in Tabelle 6.1 aufgelistet, analysiert als auch die 3 passenden Schätzer unter Verwendung multipler Imputationen, wie in Tabelle 6.7 näher erläutert.

Erweiterte Auswahl an FMA- und FMS-Schätzern $\hat{\bar{\beta}}^{\star}$ bei fehlenden Daten	
(a)	**FMA-Akaike-Schätzer**: der FMA-Schätzer auf Basis der exponentiellen AIC-Gewichte (4.6), der den Umstand fehlender Daten wie folgt berücksichtigt: *MI* – es wird das Prinzip „Mittelung nach (multipler) Imputation“ angewendet; dies entspricht dem Schätzer (5.23) unter Verwendung des aufgefüllten Datensatzes D^{imp} gemäß den Bootstrap-basierten multiplen Imputationen des R-Pakets „Amelia II“.
(b)	**FMA-Hansen-Schätzer**: der FMA-Schätzer auf Basis der Gewichte (4.9), der den Umstand fehlender Daten wie in (a) gemäß der MI-Strategie berücksichtigt.
(c)	**FMS-AIC-Schätzer**: der FMS-Schätzer für den $\Gamma = \text{AIC}$ und der den Umstand fehlender Daten wie folgt berücksichtigt: *MI* – es wird das Prinzip „Selektion nach (multipler) Imputation“ angewendet; der Schätzer (5.9) basiert dabei auf den Bootstrap-basierten multiplen Imputationen wie in (a) beschrieben.

Tab. 6.7: Erweiterung von Tabelle 6.1; drei weitere Modellmittelungsschätzer und Modellselektionsschätzer unter Verwendung multipler Imputationen

Wie bereits angedeutet sind die Varianzschätzungen in diesem Abschnitt von besonderem Interesse. Um den Effekt multipler Imputationen für die Simulationsläufe adäquat zu modellieren, werden in Anlehnung an (5.24) die mittleren Standardfehler gemäß

$$\begin{aligned}\widehat{\text{se}}_{\beta_j} \;=\; \frac{1}{\mathcal{R}}\sum_{r=1}^{\mathcal{R}}\Bigg\{\frac{1}{M}\sum_{m=1}^{M}\Bigg(\sum_{\kappa=1}^{k}\hat{w}_{\kappa,r}\sqrt{\widehat{\text{Var}}(\hat{\beta}_{j,\kappa,r}^{\star,(m)}|M_\kappa)-(\hat{\beta}_{j,\kappa,r}^{\star,(m)}-\hat{\bar{\beta}}_{j,r}^{\star,(m)})^2}\Bigg)^2 \\ +\frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\bar{\beta}}_{j,r}^{\star,(m)}-\hat{\bar{\beta}}_{j,r}^{\star,M})^2\Bigg\}^{\frac{1}{2}} \qquad (6.16)\end{aligned}$$

geschätzt und mit dem empirischen Standardfehler (6.7) verglichen. Dabei beschreibt $\hat{\beta}_{j,\kappa,r}^{\star,(m)}$ den Schätzer von β_j in Modell M_κ und $\hat{\bar{\beta}}_{j,r}^{\star,(m)}$ den gemittelten Schätzer von β_j für alle $M_\kappa \in \mathcal{M}$ jeweils im r-ten Simulationslauf für die m-te Imputation; außerdem ist $\hat{\bar{\beta}}_{j,r}^{\star,M} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\beta}}_{j,r}^{\star,(m)}$. Offensichtlich kann (6.16) auch für die FMA-Schätzer mit einfachen, nicht-multiplen Imputationen, wie auch für den FMS-AIC-Schätzer verwendet

werden. Im ersten Fall ist schlicht $M = 1$ und (6.16) entspricht (6.8); im zweiten Fall erhält unabhängig von der Methodik zur Berücksichtigung fehlender Werte das durch $\Gamma = \text{AIC}$ gewählte Siegermodell das Gewicht $w_\kappa = 1$, alle anderen Modelle das Gewicht Null.

Resultate

Die Resultate bezüglich der geschätzten Standardfehler befinden sich in Tabelle 6.8. Die Ergebnisse für die zugehörigen Punktschätzungen, exemplarisch dargestellt über den Verlust L_1, werden in Tabelle B.16 im Anhang aufgeführt; sie entsprechen qualitativ denen von Experiment 6 aus Abschnitt 6.1, wobei die Verwendung multipler Imputationen erwartungsgemäß zu marginal besseren Resultaten führt als die entsprechenden einfachen Imputationen. Es lassen sich für die Varianzschätzungen folgende Ergebnisse konstatieren:

- Betrachtet man die Originaldaten, so wird für den FMA-Akaike-Schätzer der empirische Standardfehler fast perfekt erfasst und nur an einigen Stellen marginal unterschätzt. Im Gegensatz dazu unterschätzt der FMS-AIC-Schätzer durch die nicht berücksichtigte, hier durchaus vorhandene, Modellselektionsunsicherheit den Standardfehler relativ deutlich. Einen interessanten Effekt erkennt man bei Betrachtung des FMA-Hansen-Schätzers: Generell entspricht der durch (6.16) geschätzte Standardfehler in etwa dem empirischen, mit Ausnahme von $\widehat{\text{se}}(\beta_4)$ und $\widehat{\text{se}}(\beta_6)$. Hier wird die Varianz deutlich überschätzt. Tatsächlich hängt dies mit der bereits häufig angesprochenen Einschränkung der Kandidatenmodelle zusammen: Das einzige Modell, das den wichtigen und relevanten Interaktionseffekt enthält, ist offensichtlich das volle Modell; genau dieses Modell wird von der Methodik sehr häufig, sehr stark gewichtet, enthält aber zu hohe Schätzungen für die Standardfehler, die sich eben auch im finalen Modellmittelungsschätzer wiederfinden.
- Generell lassen sich die bisher getroffenen Aussagen auch bei Durchführung einer Complete Case Analyse bestätigen: Für den FMA-Akaike-Schätzer wird der Standardfehler prinzipiell adäquat modelliert. Im Gegensatz zu den Originaldaten ist die Unterschätzung der Varianz jedoch etwas deutlicher ausgeprägt, wohl insgesamt aber noch im tolerierbaren Bereich. Erneut werden aus den oben genannten Gründen beim FMA-Hansen-Schätzer die Standardfehler $\text{se}(\beta_4)$ und $\text{se}(\beta_6)$ überschätzt. Der Selektionsschätzer auf Basis des AIC unterschätzt die Standardfehler deutlich.

	se(β_0)		se(β_1)		se(β_2)		se(β_3)		se(β_4)		se(β_5)		se(β_6)	
FMA-Akaike-Schätzer														
1) Original	0.29	(0.30)	0.16	(0.16)	0.14	(0.15)	0.07	(0.08)	0.39	(0.41)	0.07	(0.08)	0.31	(0.33)
2) CC	0.39	(0.39)	0.23	(0.24)	0.19	(0.20)	0.07	(0.09)	0.49	(0.51)	0.07	(0.09)	0.42	(0.46)
3) GAMRI	0.28	(0.34)	0.17	(0.22)	0.13	(0.17)	0.09	(0.12)	0.41	(0.55)	0.08	(0.15)	0.33	(0.45)
GLMRI	0.28	(0.32)	0.17	(0.21)	0.13	(0.16)	0.08	(0.12)	0.41	(0.53)	0.08	(0.14)	0.33	(0.40)
kNN	0.30	(0.35)	0.18	(0.21)	0.14	(0.17)	0.08	(0.10)	0.43	(0.50)	0.08	(0.11)	0.36	(0.45)
Amelia	0.30	(0.36)	0.17	(0.19)	0.14	(0.18)	0.08	(0.10)	0.40	(0.47)	0.08	(0.13)	0.33	(0.35)
MI	0.34	(0.34)	0.19	(0.18)	0.16	(0.16)	0.09	(0.10)	0.46	(0.43)	0.10	(0.11)	0.38	(0.32)
4) AIC_W	0.41	(0.57)	0.23	(0.29)	0.20	(0.32)	0.12	(0.20)	0.53	(0.63)	0.11	(0.19)	0.44	(0.54)
FMA-Hansen-Schätzer														
1) Original	0.35	(0.30)	0.17	(0.17)	0.15	(0.14)	0.14	(0.13)	0.61	(0.44)	0.12	(0.11)	0.41	(0.35)
2) CC	0.49	(0.41)	0.24	(0.23)	0.22	(0.20)	0.16	(0.14)	0.67	(0.55)	0.11	(0.12)	0.40	(0.42)
3) GAMRI	0.33	(0.35)	0.18	(0.22)	0.15	(0.16)	0.14	(0.15)	0.60	(0.59)	0.12	(0.16)	0.37	(0.44)
GLMRI	0.32	(0.35)	0.18	(0.21)	0.15	(0.16)	0.14	(0.15)	0.61	(0.57)	0.12	(0.16)	0.37	(0.39)
kNN	0.37	(0.36)	0.19	(0.21)	0.16	(0.15)	0.14	(0.14)	0.65	(0.54)	0.12	(0.14)	0.43	(0.45)
Amelia	0.37	(0.37)	0.18	(0.20)	0.16	(0.17)	0.14	(0.14)	0.60	(0.50)	0.12	(0.15)	0.38	(0.35)
MI	0.40	(0.35)	0.20	(0.19)	0.17	(0.16)	0.15	(0.14)	0.65	(0.46)	0.14	(0.13)	0.43	(0.31)
FMS-AIC-Schätzer														
1) Original	0.25	(0.32)	0.15	(0.17)	0.11	(0.16)	0.02	(0.11)	0.37	(0.41)	0.02	(0.10)	0.30	(0.33)
2) CC	0.30	(0.44)	0.22	(0.25)	0.10	(0.24)	0.01	(0.11)	0.45	(0.54)	0.02	(0.12)	0.32	(0.51)
3) GAMRI	0.24	(0.36)	0.16	(0.23)	0.09	(0.18)	0.04	(0.15)	0.39	(0.55)	0.03	(0.17)	0.29	(0.46)
GLMRI	0.23	(0.35)	0.16	(0.21)	0.08	(0.17)	0.03	(0.15)	0.39	(0.53)	0.04	(0.17)	0.29	(0.40)
kNN	0.26	(0.39)	0.17	(0.22)	0.11	(0.18)	0.03	(0.13)	0.41	(0.51)	0.03	(0.14)	0.32	(0.45)
Amelia	0.26	(0.40)	0.16	(0.20)	0.11	(0.19)	0.03	(0.13)	0.38	(0.47)	0.03	(0.16)	0.29	(0.36)
MI	0.32	(0.36)	0.19	(0.18)	0.15	(0.17)	0.06	(0.12)	0.45	(0.43)	0.08	(0.13)	0.36	(0.32)
4) AIC_W	0.38	(0.59)	0.22	(0.29)	0.17	(0.32)	0.07	(0.22)	0.52	(0.64)	0.06	(0.21)	0.41	(0.55)

Tab. 6.8: Mittlere geschätzte Standardfehler unter Verwendung von (6.16); in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

- Wie in der vorliegenden Arbeit bereits vielfach diskutiert, führt die Verwendung des AIC_W, unabhängig ob für eine einfache Selektion oder zur Konstruktion von Gewichten verwendet, zu einer Unterschätzung der Varianz. Dies ist erneut klar zu erkennen. Die Unsicherheit bei der Wahl der für das GAM notwendigen Kovariablen, wie auch die Unsicherheit bei der Wahl des Glättungsparameters scheinen insgesamt nicht vernachlässigbar zu sein.

- Unabhängig davon, ob für einfache, nicht-multiple Imputationen die GAMRI-, GLMRI-, kNN- oder Amelia-Methode gewählt wird, ist eine Unterschätzung der Varianz sowohl für die beiden FMA- als auch den FMS-Schätzer zu erkennen. Einzige Ausnahme bilden die Schätzer der Standardfehler $\widehat{\text{se}}(\beta_4)$ und $\widehat{\text{se}}(\beta_6)$ des FMA-Hansen-Schätzers. Wie oben diskutiert, ist dies jedoch ausnahmslos auf die restriktive Wahl der Kandidatenmodelle innerhalb der MMA-Methodik zurückzuführen.

- Für die Schätzer auf Basis multipler Imputationen ist prinzipiell zu erwarten, dass die oben angesprochene Imputationsunsicherheit erfasst wird und damit eine Unterschätzung der Standardfehler verhindert werden kann. Ein Blick auf die Ergebnisse zeigt, dass für viele Effekte, etwa bei den Variablen X_1, X_2, X_3 und X_5, die geschätzten Standardfehler der FMA-Schätzer gemäß (6.16) den empirischen Standardfehler fast perfekt erfassen. Tatsächlich ist jedoch auch eine klare Überschätzung der Varianz für $\text{se}(\beta_4)$ und $\text{se}(\beta_6)$ festzustellen – und zwar für alle drei betrachteten Schätzer. Für den FMA-Hansen-Schätzer ist hierfür bereits eine ausreichende Erklärung gefunden worden: die restriktive Auswahl der Kandidatenmodelle. Für den FMA-Akaike- und den FMS-AIC-Schätzer kann diese Erklärung offensichtlich nicht herangezogen werden. Das folgende Experiment 2 wird zeigen, dass die Überschätzung der Standardfehler auf eine sehr große Imputationsunsicherheit für die binomialverteilte Variable X_4 zurückzuführen ist: Die multiplen Imputationen für X_4 sind insgesamt sehr verschieden, es herrscht eine große Imputationsunsicherheit, so dass der letzte Term von (6.16) für die relevanten β_j, hier β_4 und β_6, sehr groß wird und damit für den komplexen Teil des postulierten Zusammenhangs, also den Einfluss von X_4 auch über den Interaktionseffekt, sehr unsichere Schätzungen entstehen. Experiment 2 wird zeigen, dass für einen alternativen Fehlendmechanismus, der nicht auf den Interaktionseffekt wirkt, multiple Imputationen zu einer durchgehend adäquaten Modellierung der Standardfehler führen.

Experiment 2: Die Rolle des Fehlendmechanismus

Das vorliegende Experiment entspricht dem vorhergehenden, mit einer entscheidenden Modifikation: Der Fehlendmechanismus wirkt nicht mehr auf die Variablen X_1 und X_4, die den Interaktionseffekt prägen. Betrachtet wird nur

$$\pi_{X_2,X_5}(y) \;=\; 1 - \frac{1}{0.0225y^2 + 1}\,,$$

wodurch ausschließlich Werte von X_2 und X_5 als fehlend deklariert werden. In den $\mathcal{R} = 500$ Simulationsläufen resultiert dies im Mittel in je 21% fehlender Werte für die beiden Variablen.

Resultate

Die Resultate befinden sich in Tabelle 6.9 und Tabelle B.17. Offensichtlich sind die Ergebnisse für die Punktschätzungen qualitativ dieselben wie im vorhergehenden Abschnitt, die der zugehörigen Varianzschätzungen eindeutig interessanter:

- Die Aussagen für die Originaldaten sowie die CC-, AIC_W- , GAMRI-, GLMRI-, kNN- und Amelia-Methodik bleiben dieselben: Generell ist für eine Complete Case Analyse eine marginale Unterschätzung, für die AIC_W- bzw. Imputationsverfahren eine deutliche Unterschätzung der Varianz zu erkennen. Weiterhin überschätzt der FMA-Hansen-Schätzer aus den oben genannten Gründen die Standardfehler für X_4 und den Interaktionseffekt.
- Betrachtet man die Schätzungen der Standardfehler basierend auf multiplen Imputationen, so ist tatsächlich ein klare Änderung festzustellen: Der FMA-Akaike-Schätzer erfasst hier die Standardfehler fast perfekt; geschätzte und empirische Standardfehler sind für alle Effekte nahezu identisch. Eine Überschätzung findet im Gegensatz zum vorhergehenden Experiment nicht mehr statt. Dies zeigt erneut, dass der Gewichtungsansatz unter Verwendung exponentieller AIC-Gewichte kombiniert mit multiplen Imputationen in vielen Situationen ein guter Ansatz sein kann. Für den FMA-Hansen-Schätzer sind nur geringe Änderungen im Vergleich zu den einfachen, nicht-multiplen Imputationen festzustellen. Insgesamt bleibt auf die bestehenden Instabilitäten dieses Schätzers hinzuweisen. Der FMS-Schätzer unterschätzt für einige Effekte weiterhin die Standardfehler aufgrund der nicht berücksichtigten Modellselektionsunsicherheit.

	se(β_0)		se(β_1)		se(β_2)		se(β_3)		se(β_4)		se(β_5)		se(β_6)	
FMA-Akaike-Schätzer														
1) Original	0.29	(0.28)	0.16	(0.16)	0.14	(0.14)	0.07	(0.07)	0.39	(0.38)	0.07	(0.08)	0.31	(0.32)
2) CC	0.30	(0.29)	0.20	(0.21)	0.14	(0.15)	0.06	(0.08)	0.40	(0.42)	0.06	(0.08)	0.35	(0.40)
3) GAMRI	0.32	(0.39)	0.17	(0.19)	0.16	(0.23)	0.08	(0.08)	0.41	(0.41)	0.10	(0.20)	0.33	(0.33)
GLMRI	0.32	(0.39)	0.17	(0.18)	0.16	(0.24)	0.08	(0.08)	0.41	(0.41)	0.10	(0.19)	0.33	(0.33)
kNN	0.32	(0.39)	0.17	(0.18)	0.16	(0.23)	0.08	(0.08)	0.41	(0.41)	0.10	(0.17)	0.33	(0.33)
Amelia	0.30	(0.36)	0.17	(0.18)	0.14	(0.21)	0.08	(0.08)	0.41	(0.41)	0.09	(0.15)	0.33	(0.33)
MI	0.34	(0.34)	0.17	(0.18)	0.18	(0.19)	0.08	(0.08)	0.41	(0.41)	0.12	(0.14)	0.33	(0.33)
4) AIC_W	0.36	(0.49)	0.20	(0.28)	0.17	(0.25)	0.11	(0.20)	0.48	(0.66)	0.11	(0.20)	0.39	(0.55)
FMA-Hansen-Schätzer														
1) Original	0.35	(0.29)	0.17	(0.17)	0.15	(0.13)	0.14	(0.12)	0.61	(0.41)	0.12	(0.11)	0.41	(0.33)
2) CC	0.38	(0.31)	0.21	(0.21)	0.15	(0.14)	0.13	(0.12)	0.55	(0.45)	0.10	(0.10)	0.34	(0.36)
3) GAMRI	0.38	(0.38)	0.18	(0.19)	0.18	(0.23)	0.15	(0.13)	0.63	(0.44)	0.15	(0.21)	0.42	(0.34)
GLMRI	0.39	(0.38)	0.18	(0.18)	0.18	(0.23)	0.15	(0.13)	0.63	(0.44)	0.15	(0.21)	0.42	(0.34)
kNN	0.38	(0.38)	0.18	(0.19)	0.18	(0.22)	0.14	(0.13)	0.63	(0.44)	0.15	(0.19)	0.43	(0.34)
Amelia	0.37	(0.36)	0.18	(0.18)	0.16	(0.20)	0.14	(0.13)	0.64	(0.44)	0.13	(0.17)	0.43	(0.34)
MI	0.40	(0.34)	0.19	(0.18)	0.20	(0.18)	0.15	(0.13)	0.64	(0.44)	0.16	(0.16)	0.43	(0.34)
FMS-AIC-Schätzer														
1) Original	0.25	(0.31)	0.15	(0.17)	0.11	(0.16)	0.02	(0.10)	0.37	(0.38)	0.02	(0.10)	0.30	(0.31)
2) CC	0.25	(0.34)	0.18	(0.22)	0.08	(0.17)	0.02	(0.11)	0.37	(0.45)	0.02	(0.11)	0.27	(0.44)
3) GAMRI	0.27	(0.41)	0.16	(0.19)	0.13	(0.25)	0.02	(0.11)	0.39	(0.41)	0.05	(0.22)	0.31	(0.33)
GLMRI	0.28	(0.41)	0.16	(0.19)	0.13	(0.25)	0.02	(0.11)	0.39	(0.41)	0.05	(0.22)	0.31	(0.33)
kNN	0.28	(0.41)	0.16	(0.19)	0.13	(0.24)	0.02	(0.10)	0.39	(0.41)	0.05	(0.20)	0.31	(0.32)
Amelia	0.26	(0.38)	0.16	(0.18)	0.11	(0.22)	0.02	(0.10)	0.39	(0.40)	0.04	(0.17)	0.31	(0.32)
MI	0.32	(0.36)	0.17	(0.18)	0.18	(0.20)	0.03	(0.09)	0.40	(0.40)	0.11	(0.16)	0.32	(0.32)
4) AIC_W	0.32	(0.51)	0.20	(0.28)	0.14	(0.26)	0.06	(0.23)	0.46	(0.66)	0.06	(0.22)	0.36	(0.56)

Tab. 6.9: Mittlere geschätzte Standardfehler unter Verwendung von (6.16); in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

Alternative Varianzschätzungen

An dieser Stelle soll noch einmal die Diskussion aus Abschnitt 5.2.2 aufgenommen werden. Es stellt sich die Frage, ob die Reihenfolge der Berücksichtigung der Unsicherheiten eine Rolle spielt oder nicht. In den vorangegangenen beiden Experimenten wurde gemäß gängiger statistischer Konvention zuerst jede Form von Inferenz auf den M imputierten Datensätzen durchgeführt, anschließend wurden diese Schätzungen kombiniert. Dies impliziert, dass zuerst die Modellselektionsunsicherheit berücksichtigt wird, danach die Imputationsunsicherheit. Die Verwendung der Ideen von (5.27) führt dagegen für die vorliegenden beiden Experimente zu folgender, alternativen Schätzung des mittleren Standardfehlers:

$$\widehat{\mathrm{se}}_{\beta_j} = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \sum_{\kappa=1}^{k} w_\kappa^M \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_{j,\kappa,r}^{\star,M}) + (\hat{\bar{\beta}}_{j,r}^{\star,M} - \hat{\beta}_{j,\kappa,r}^{\star,M})^2} \right\} . \tag{6.17}$$

Dabei ist

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_{j,\kappa,r}^{\star,M}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_{j,\kappa,r}^{\star,(m)}) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\beta}_{j,\kappa,r}^{\star,M} - \hat{\beta}_{j,\kappa,r}^{\star,(m)})^2 , \tag{6.18}$$

wobei $\hat{\beta}_{j,\kappa,r}^{\star,M} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_{j,\kappa,r}^{\star,(m)}$ und $w_\kappa^M = \frac{1}{M} \sum_{m=1}^{M} w_\kappa^{(m)}$. Offensichtlich wird dadurch zuerst für jedes Modell ein Schätzer für den Standardfehler berechnet, der die Unsicherheit bezüglich der Imputationen berücksichtigt und dann anschließend über die Modelle gewichtet mittelt.

Im Folgenden werden nun erneut die beiden vorangegangenen Experimente durchgeführt und die mittleren geschätzten Standardfehler nach (6.16) den alternativen Schätzungen gemäß (6.17) gegenübergestellt. Um die Schätzungen weiter zu stabilisieren, werden die Simulationsläufe erhöht: $\mathcal{R} = 1000$. Betrachtet wird ausschließlich der FMA-Akaike-Schätzer. Die wesentlichen Resultate dafür sind in den Tabellen 6.10 und 6.11 zusammengefasst.

Es ist zu erkennen, dass für das erste Experiment die beiden Schätzungen nicht allzu weit voneinander entfernt liegen. Im Trend scheint die alternative Schätzung nach (6.17) die Standardfehler jedoch etwas stärker zu unter- bzw. überschätzen. Für die Resultate bezüglich des zweiten Experiments ist dies noch deutlicher zu erkennen: Während die geschätzten Standardfehler für die Effekte X_1 und X_2 noch relativ ähnlich eingeschätzt

	se(β_0)		se(β_1)		se(β_2)		se(β_3)	
MI	0.34	(0.34)	0.19	(0.19)	0.16	(0.17)	0.09	(0.10)
MI alt.	0.36	(0.34)	0.20	(0.19)	0.15	(0.17)	0.07	(0.10)

	se(β_4)		se(β_5)		se(β_6)	
MI	0.47	(0.41)	0.10	(0.11)	0.38	(0.32)
MI alt.	0.49	(0.41)	0.08	(0.11)	0.39	(0.32)

Tab. 6.10: Ergebnisse für die mittlere klassische Varianzschätzung (6.16) und die mittlere alternative Varianzschätzung (6.17) des FMA-Akaike-Schätzers im Grundszenario; in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

	se(β_0)		se(β_1)		se(β_2)		se(β_3)	
MI	0.34	(0.34)	0.17	(0.17)	0.18	(0.19)	0.08	(0.09)
MI alt.	0.37	(0.34)	0.18	(0.17)	0.18	(0.19)	0.06	(0.09)

	se(β_4)		se(β_5)		se(β_6)	
MI	0.41	(0.42)	0.12	(0.13)	0.33	(0.35)
MI alt.	0.46	(0.42)	0.09	(0.12)	0.42	(0.35)

Tab. 6.11: Ergebnisse für die mittlere klassische Varianzschätzung (6.16) und die mittlere alternative Varianzschätzung (6.17) des FMA-Akaike-Schätzers in Experiment 2; in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

werden, so werden sie für X_3 und X_5 von (6.17) im Vergleich zu (6.16) merklich unterschätzt. Noch größer gestaltet sich Unterschied für die bereits als kritisch eingeordneten Effekte X_4 und X_1X_4; die alternative Schätzung überschätzt die Standardfehler hier deutlich.

Während (6.16) für das zweite Experiment eine adäquate Modellierung der Varianz bietet, ist dies bei (6.17) nicht zu erkennen. Insgesamt scheint das klassische Vorgehen, also zuerst die Modellselektionsunsicherheit zu beachten, dann Imputationsunsicherheit, wie erwartet die bessere Wahl zu sein.

6.4 Zusammenfassung

Die in diesem Kapitel gewonnenen Erkenntnisse sind äußerst vielfältig und in vielerlei Hinsicht aufschlussreich für die gewählten Modellselektions- und Modellmittelungsverfahren im Kontext fehlender Daten. Auch wenn Simulationsergebnisse grundsätzlich immer nur mit Vorsicht auf Situationen außerhalb der betrachteten übertragen werden können und die Resultate zu einem gewissen Teil von den gewählten Einstellungen abhängen, so ergeben sich für die aufwändig und intensiv durchgeführten Studien dieses Kapitels dennoch einige entscheidende Feststellungen:

- Unabhängig von der gewählten Methodik zur Berücksichtigung fehlender Werte sind für den FMA-Akaike-Schätzer fast durchgängig bessere Punktschätzungen als beim FMS-AIC- bzw. FMA-Hansen-Schätzer zu erkennen. Letzterer wirkt insgesamt sehr instabil; die Qualität seiner Schätzungen hängt stark von der betrachteten Situation ab. In einigen Fällen, etwa bei den Experimenten 2 und 3 in Abschnitt 6.1, liefert er zwar insbesondere für die Verlustfunktionen L_2, L_3 und L_4 bessere Schätzungen als der FMS-AIC-Schätzer, insgesamt kann jedoch nicht konstatiert werden, dass er dem Selektionsschätzer überlegen ist. Die Qualität der entsprechenden Varianzschätzungen scheint beim FMA-Akaike-Schätzer ebenfalls am besten zu sein.

- Die Verwendung der vollständigen Fälle (CC) führt bei allen FMA- und FMS-Schätzern nahezu ausschließlich zu sehr unzureichenden Punktschätzungen – sowohl bei Betrachtung der linearen Regression als auch bei Betrachtung der logistischen Regression. Dennoch sind die zugehörigen Varianzschätzungen in einem akzeptablen Bereich; sie unterschätzen die empirische Varianz nur in etwas komplexeren Situationen und auch dann nur marginal.

- Die Verwendung des AIC_W, entweder direkt zur Selektion des Modells oder alternativ zur Konstruktion von Modellmittelungsgewichten, führt in Abschnitt 6.1 (mit Ausnahme von Experiment 7) in der Regel zu besseren Schätzungen als eine Complete Case Analyse und zu etwas schlechteren Ergebnissen als die Imputationsverfahren. In den betrachteten Situationen der logistischen Regressionsanalyse aus Abschnitt 6.2, kann die AIC_W-Methodologie meist keinen Gewinn erbringen. Sie schneidet häufig am schlechtesten ab. Die Varianz der entsprechenden Schätzer wird in allen Simulationen der drei vorangegangenen Abschnitte 6.1-6.3 unterschätzt. Korrekterweise müsste eigentlich noch die Unsicherheit bei der Wahl

des Glättungsparameters und den Kovariablen für das zur Schätzung der Gewichte (5.5) notwendige GAM berücksichtigt werden.

- Die Verwendung von einfachen, nicht-multiplen Imputationen führt in der Regel zu sehr guten Schätzungen für alle FMA- und FMS-Schätzer. Häufig liefert die Amelia-Methode, also die Verwendung des *R*-Pakets „Amelia II“, sehr gute und stabile Ergebnisse. In einigen Situationen, etwa den Experimenten 3 und 7 aus Abschnitt 6.1 und den Experimenten 4,5 und 6 aus Abschnitt 6.2, kann jedoch auch die Verwendung der GAMRI-,GLMRI- bzw. kNN-Methodik zu den besten Resultaten führen. Prinzipiell kann nicht konstatiert werden, dass bei Gebrauch des in Abschnitt 5.1.2 vorgestellten verallgemeinerten Regressionsimputationsalgorithmus die Verwendung von generalisierten additiven Modellen (GAMRI) der Verwendung simpler generalisierter linearer Regressionsmodelle (GLMRI) vorzuziehen ist. Der Erfolg dieser beiden Methoden variiert sehr stark mit dem gewählten Setting. Generell wird bei allen vier betrachteten Imputationsmethoden die Varianz unterschätzt, da die Imputationsunsicherheit nicht erfasst wird.

- Die Verwendung von korrekten multiplen Imputationen unter Verwendung des Amelia II-Pakets der statistischen Software *R* führt – wie in Abschnitt 6.3 diskutiert – zu guten Schätzungen, insbesondere auch in Bezug auf die Varianz, die speziell für den FMA-Akaike-Schätzer nicht unterschätzt wird. Wirkt der Fehlendmechanismus auf eine Variable, die einen komplexeren Effekt, etwa einen Interaktionseffekt, mitgestaltet und ist die Imputationsunsicherheit groß, so wird diese Unsicherheit explizit mitmodelliert und kann zu einer Überschätzung der Varianz führen. Dies verdeutlicht noch einmal die großen Herausforderungen, die in der Modellbildung unter Berücksichtigung fehlender Daten liegen. Ferner werden in den betrachteten Situationen die Standardfehler besser geschätzt, wenn zuerst die Modellselektionsunsicherheit und dann die Imputationsunsicherheit berücksichtigt wird und nicht umgekehrt.

Die zu Beginn dieses Kapitel aufgeworfenen Fragen lassen sich unter Berücksichtigung dieser Erkenntnisse wie folgt beantworten:

1. Modellmittelungsverfahren sind der Modellselektion nicht „generell“ überlegen, dies zeigen insbesondere die Ergebnisse der Mallows-Model-Averaging-Methodologie von Hansen (2007). Bei Betrachtung zweier spezifischer Verfahren, wie in diesem Kapitel etwa der FMA-Akaike- und FMS-AIC-Schätzer, kann jedoch konsta-

tiert werden, dass die Modellmittelung in der Regel deutlich bessere Schätzungen liefert als die Modellselektion.

2. Schätzer die Optimalitätseigenschaften besitzen sind solchen die eher pragmatisch motiviert sind nicht zwingenderweise überlegen. Dies folgt eindeutig aus dem Vergleich des FMA-Akaike- und FMA-Hansen-Schätzers.

3. Um auf den Sachverhalt fehlender Werte einzugehen, bieten sich viele Möglichkeiten an. In der Regel versprechen Imputationsverfahren die besten Ergebnisse. Einfache Imputationen liefern respektable Punktschätzungen, unterschätzen jedoch die zugehörige Varianz; Multiple Imputationen führen ebenfalls zu guten Punktschätzungen, können in einigen Situationen jedoch zu einer Überschätzung der entsprechenden Varianz führen. Der finale Erfolg der einzelnen Imputationsmethoden hängt von der gewählten Situation ab.

7. Anwendungsbeispiele

Die folgenden drei Abschnitte illustrieren die Verwendung der vorgestellten FMS- und FMA-Schätzer unter spezieller Berücksichtigung der Problematik fehlender Daten. Abschnitt 7.1 betrachtet dabei das lineare Regressionsmodell, Abschnitt 7.2 das logistische Regressionsmodell und Abschnitt 7.3 diskutiert die Chancen und Risiken der vorgestellten Methoden im Kontext der Faktorenanalyse.

7.1 Phasenangepasste Führung von Wachstumsunternehmen

Die folgenden beiden Abschnitte betrachten und erweitern die Analysen von Klaußner (2007), der den Erfolg von Unternehmen in der Life-Science-Branche abhängig vom Führungsverhalten und abhängig von der Unternehmenslebensphase untersucht. Die Methoden, die dabei von Klaußner (2007) verwendet werden, sind vorwiegend deskriptiver Natur, enthalten aber auch lineare Regressionsanalysen, die unter Berücksichtigung der Erkenntnisse der ersten sechs Kapitel überprüft, adjustiert und in den Kontext fehlender Daten gebracht werden. Die verwendeten Daten stammen dabei aus $n = 101$ Unternehmen und messen die interessierenden Größen wie folgt:

- Als „Erfolg“ wird jede Art von Führungserfolg verstanden: Dies meint explizit 1) die *Zufriedenheit* der Mitarbeiter mit ihrer Führung, 2) die *Leistungsbereitschaft* eines jeden Einzelnen, die über das grundsätzliche Maß einer Führer-Geführten-Beziehung hinausgeht und 3) die *Effektivität* der Führung, eine Maßzahl für die wirksame Erfüllung der Aufgaben durch den Geführten. Diese drei Größen werden über den Multifactor Leadership Questionnaire (MLQ, vgl. Bass, Aviolo und Active (2003)) anhand eines Scores bestimmt. Der Score ist dabei der Mittelwert von vier Fragen auf einer Skala zwischen null und vier. Abschnitt 7.1.1 betrachtet eine lineare Regressionsanalyse für die Zufriedenheit, Abschnitt 7.1.2 eine lineare Regressionsanalyse für die Effektivität; die Leistungsbereitschaft als Indikator für Erfolg wird in dieser Arbeit nicht näher untersucht.

- Das „Führungsverhalten“ wird durch eine Einteilung von Bass und Steyrer (1995) charakterisiert, die versuchen das gesamte bekannte Führungsspektrum über neun verschiedene Führungsstile abzubilden. Die Intensität jedes einzelnen Führungsstils wird dabei erneut über einen Score zwischen null und vier mit Hilfe des MLQ erfasst. Bass und Steyrer (1995) ordnen dabei – wie in der einschlägigen Literatur üblich – die entsprechenden Führungsstile sowohl transaktionalem als auch transformationalem Führungsverhalten zu:

  - Transaktionales Führungsverhalten beschreibt dabei eine einfache Austauschbeziehung zwischen Führungskraft und Geführtem; dies bedeutet, dass das Verhältnis zwischen beiden Seiten als eine Transaktion aufgefasst wird: Der Geführte leistet genau das, was von ihm erwartet wird und bekommt dafür als Gegenleistung von der Führungspersönlichkeit das, was vorher vereinbart wurde, beispielsweise ein höheres Gehalt oder eine Beförderung. Ein solcher Führungsstil beinhaltet Verhaltensweisen, die mit dem Multifactor Leadership Questionnaire über die Variablen *Contingent Reward (CR), Management by Exception Active (MEA)* und *Management by Exception Passive (MEP)* abgebildet werden. Für eine detaillierte Motivation und eine ausführliche Unterscheidung dieser Größen wird auf Klaußner (2007, Seite 88ff.) sowie die folgenden Analysen verwiesen.
  - Transformationale Führung meint jene Art von Führung, die es vermag, Leistungen freizusetzen, die über dem liegen, was von Mitarbeitern unter „normalen Bedingungen“ erwartet wird. Dies beinhaltet insbesondere charismatische Führung und Inspiration, intellektuelle Stimulation und eine individualisierte Betrachtung der Mitarbeiter. Der MLQ bildet dieses Führungsverhalten über die Variablen *Idealized Influence Attributed (IIA), Idealized Influence Behavioral (IIB), Inspirational Motivation (IM), Intellectual Stimulation (IS)* und *Individual Consideration (IC)* ab.
  - Betrachtet wird ferner auch ein *Laisser-faire-Führungsstil*, der als Sonderfall zu verstehen ist; er ist weder transaktional noch transformational und zeichnet sich durch ein Höchstmaß an Ineffektivität aus, da bei zu bewältigenden Aufgaben nicht entschieden gehandelt wird und dadurch möglicherweise erst Probleme entstehen.

- Um die „Lebensphase“ eines Unternehmens zu charakterisieren, wird das *Lebenszyklus-Modell* von Pümpin und Prange (1991) betrachtet: Die Autoren gehen davon aus, dass jedes Unternehmen die Phasen Entstehung (=1), Wachstum (=2), Reife (=3) und Niedergang (=4) durchlebt – eventuell wiederholen sich diese Phasen nach vielen Jahren und auf die Niedergangsphase erfolgt eine Bereinigung und neues Entstehen und Wachstum. Jedes der 101 untersuchten Unternehmen wird auf Basis eines Fragebogens in eine dieser Phasen eingeteilt.

Ziel der Untersuchung ist es, den Erfolg eines Unternehmens durch ein phasenspezifisches Führungsverhalten zu charakterisieren. Dies beinhaltet die folgenden Fragestellungen: Welches Führungsverhalten steigert bzw. vermindert den (Führungs-)Erfolg in großem Ausmaß? Sind in verschiedenen Unternehmenslebensphasen verschiedene Führungsstile notwendig?

7.1.1 Analyse der Zufriedenheit

In diesem Abschnitt wird der Erfolgsindikator Zufriedenheit betrachtet. Es stellt sich die Frage, durch welches Führungsverhalten sie besonders gesteigert bzw. vermindert werden kann und ob dies abhängig von der Unternehmenslebensphase ist oder nicht. Tabelle 7.1.1 zeigt die Korrelation nach Spearman zwischen der Zufriedenheit und allen interessierenden Größen. Es fällt sofort auf, dass die transformationalen Führungsstile

	IIA	IIB	IM	IS	IC	CR	MEA	MEP	L	Phase
Zufriedenheit	0.65	0.40	0.43	0.53	0.67	0.55	0.01	-0.37	-0.37	-0.45

Tab. 7.1: Korrelation nach Spearman zwischen der Zufriedenheit und den potentiellen Einflussgrößen

IIA, IIB, IM, IS und IC allesamt positiv mit der Zufriedenheit korrelieren. Transaktionales Führungsverhalten, erfasst durch CR, MEA und MEP, wirkt positiv, neutral sowie negativ auf den Erfolg, was dafür spricht, dass diese Größen unterschiedlich erfasst und interpretiert werden müssen. Ein Laisser-faire-Führungsstil korreliert – wie zu erwarten – negativ mit der Zufriedenheit. Die negative Korrelation für die Unternehmenslebensphase impliziert, dass für spätere Phasen (Reife, Niedergang) eine geringere Zufriedenheit mit dem Führungspersonal beobachtet werden kann.

Um nun die relevanten Variablen zu klassifizieren und die Stärke ihres Einflusses zu quantifizieren, wird eine lineare Regressionsanalyse mit der Zufriedenheit als Response und allen oben angegebenen Größen als potentiellen Kovariablen durchgeführt; die Unternehmenslebensphase wird dabei dummykodiert mit Phase 1 (=Entstehung) als Referenz. Um die vorgestellten Methoden zur Berücksichtigung fehlender Daten zu veranschaulichen, wird ein MCAR-Fehlendmechanismus eingeführt[20]; für alle Kovariablen wird die konstante Fehlwahrscheinlichkeit

$$\pi_{\mathrm{IIA}} \;=\; \ldots = \pi_{\mathrm{L}} = \pi_{\mathrm{Phase}} = 0.1$$

angenommen. Dies resultiert aufgrund der vielen betrachteten Variablen zu 68.32% an Beobachtungen, bei denen mindestens ein Wert fehlt. Eine Complete Case Analyse kann also nur auf etwas mehr als 30% der Fälle zurückgreifen. Die Kandidatenmodelle sind dieselben wie bei Klaußner (2007, Seite 235), der aufgrund verschiedener Selektionsprozeduren und einigen ad-hoc Betrachtungen sieben plausible Modelle vorstellt, deren AIC-Werte jeweils sehr nahe beisammen liegen[21]:

$$\begin{aligned}
\mathrm{M}_1: \quad y = \;& \alpha + \gamma_1\mathrm{IIA} + \gamma_2\mathrm{IIB} + \gamma_3\mathrm{IM} + \gamma_4\mathrm{IS} + \gamma_5\mathrm{IC} + \gamma_6\mathrm{CR} + \gamma_7\mathrm{MEA} + \gamma_8\mathrm{MEP} \\
& + \gamma_9\mathrm{L} + \gamma_{10}\mathrm{P2} + \gamma_{11}\mathrm{P3} + \gamma_{12}\mathrm{P4} + \gamma_{13}\mathrm{LP2} + \gamma_{14}\mathrm{LP3} + \gamma_{15}\mathrm{LP4}\,, \\
\mathrm{M}_2: \quad y = \;& \alpha + \gamma_1\mathrm{IIA} + \gamma_5\mathrm{IC} + \gamma_7\mathrm{MEA} + \gamma_9\mathrm{L} + \gamma_{10}\mathrm{P2} + \gamma_{11}\mathrm{P3} + \gamma_{12}\mathrm{P4} + \gamma_{13}\mathrm{LP2} \\
& + \gamma_{14}\mathrm{LP3} + \gamma_{15}\mathrm{LP4}\,, \\
\mathrm{M}_3: \quad y = \;& \alpha + \gamma_1\mathrm{IIA} + \gamma_5\mathrm{IC} + \gamma_9\mathrm{L} + \gamma_{10}\mathrm{P2} + \gamma_{11}\mathrm{P3} + \gamma_{12}\mathrm{P4} + \gamma_{13}\mathrm{LP2} \\
& + \gamma_{14}\mathrm{LP3} + \gamma_{15}\mathrm{LP4}\,,
\end{aligned}$$

[20] Die Motivation für dieses Anwendungsbeispiel Daten künstlich zu verwerfen, liegt in der von Klaußner (2007, Seite 184 ff.) angesprochenen Problematik der Datenerhebung. Ursprünglich angedacht war eine günstige und schnelle Online-Befragung. Aufgrund der geringen Rücklaufquote und einer Vielzahl fehlender Werte sah sich der Autor dazu gezwungen, durch den Besuch mehrerer Fachmessen die Umfragen persönlich durchzuführen. Da sich ein solches Prozedere nicht nur als zeit-, sondern auch als kostenintensiv darstellt, ist die Sensitivität der von Klaußner (2007) gefundenen Resultate bezüglich fehlender Daten auch für zukünftige, eventuell zeitlich eingeschränkte Forschungsarbeiten relevant. Ein MCAR-Fehlendmechanismus dient dabei einer einfachen Illustration der Problematik; MAR- und MNAR-Fehlendmechanismen werden im nächsten Abschnitt 7.1.2 betrachtet und motiviert.

[21] Für fünf der sieben betrachteten Modelle ist die Differenz der entsprechenden AIC-Werte zum geringsten AIC-Wert kleiner als 5: $\mathrm{AIC}(\mathrm{M}_2)$=174.38, $\mathrm{AIC}(\mathrm{M}_3)$=173.28, $\mathrm{AIC}(\mathrm{M}_4)$=175.19, $\mathrm{AIC}(\mathrm{M}_5)$=175.93, $\mathrm{AIC}(\mathrm{M}_6)$=178.21. Die AIC-Werte unterschieden sich geringfügig von denen von Klaußner (2007, Abbildung 87), da dieser teilweise gerundete Scores für die Analyse verwendet.

$$
\begin{aligned}
\mathrm{M}_4: \quad y &= \alpha + \gamma_1 \mathrm{IIA} + \gamma_5 \mathrm{IC} + \gamma_7 \mathrm{MEA} + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_5: \quad y &= \alpha + \gamma_1 \mathrm{IIA} + \gamma_5 \mathrm{IC} + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_6: \quad y &= \alpha + \gamma_1 \mathrm{IIA} + \gamma_5 \mathrm{IC} + \gamma_9 \mathrm{L}\,, \\
\mathrm{M}_7: \quad y &= \alpha + \gamma_1 \mathrm{IIA} + \gamma_5 \mathrm{IC}\,.
\end{aligned}
$$

Diese Auswahl deutet darauf hin, dass die transformationalen Führungsstile IIA und IC, der transaktionale Führungsstil MEA und ein Laisser-faire Verhalten (teilweise in Interaktion mit der Unternehmenslebensphase) von besonderem Interesse sind. Um die Daten zu analysieren und die vorgestellten Methoden geeignet zu illustrieren, werden dieselben 20 Schätzer wie in Tabelle 6.1 betrachtet und ferner der FMA-Akaike-, FMA-Hansen- bzw. FMS-AIC-Schätzer nach multipler Imputation gemäß Tabelle 6.7. Es ist klar, dass alle FMA-Hansen-Schätzer das Modell M_1 und die entsprechenden 15 Submodelle betrachten. Die Ergebnisse der Analyse befinden sich in Tabelle 7.2.

Resultate

Es lassen sich folgende Resultate bezüglich der inhaltlichen Interpretation, der drei betrachteten Schätzer und der verschiedenen Methoden zur Berücksichtigung der fehlenden Werte konstatieren:

- Betrachtet man die Schätzungen der Regressionsparameter sowie deren geschätzte Standardfehler für die Originaldaten, so ist zu erkennen, dass alle Schätzer ein IIA-Führungsverhalten als besonders relevant für die Zufriedenheit einstufen; die Parameterschätzungen sind dabei stets mindestens dreimal so groß wie der zugehörige Standardfehler. Insbesondere der FMA-Akaike-, aber auch der FMS-AIC-Schätzer ordnet darüber hinaus auch IC eine größere Bedeutung zu. Dabei wirken die Führungsstile IIA und IC positiv auf die Zufriedenheit. Dies war aufgrund der betrachteten Korrelationen zu erwarten. Alle anderen Größen scheinen keinen relevanten Effekt auf die Zufriedenheit zu besitzen.

- Generell erfassen die Schätzer auf Basis der betrachteten Imputationsansätze die Struktur ähnlich wie die entsprechenden Schätzer auf Basis der Originaldaten. Der Hauptunterschied liegt im Zusammenspiel von Phase und Laisser-faire: Während die Verwendung multipler Imputationen (MI) bzw. des GAMRI-Algorithmus ähnlich den Originaldaten sowohl Phase als auch Laisser-faire als wenig relevant einstufen, so gesteht die kNN- bzw. Amelia-Methodik Laisser-faire an sich und das

	I.		IIA		IIB		IM		IS		IC		CR		MEA		
FMA-Akaike-Schätzer																	
1) Original	0.91	(0.43)	0.38	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.37	(0.10)	0.00	(0.00)	-0.04	(0.07)	···
2) CC	1.68	(2.22)	0.08	(0.31)	-0.07	(0.30)	-0.24	(0.35)	-0.04	(0.39)	0.32	(0.31)	0.51	(0.44)	-0.24	(0.23)	···
3) GAMRI	0.66	(0.44)	0.38	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.37	(0.10)	0.00	(0.00)	-0.01	(0.03)	···
GLMRI	0.47	(0.48)	0.32	(0.11)	-0.01	(0.03)	0.00	(0.02)	0.02	(0.04)	0.45	(0.11)	0.05	(0.09)	-0.07	(0.09)	···
kNN	1.23	(0.47)	0.31	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.42	(0.10)	0.00	(0.01)	-0.09	(0.10)	···
Amelia	1.49	(0.46)	0.31	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.41	(0.10)	0.00	(0.01)	-0.13	(0.10)	···
MI	1.14	(0.63)	0.36	(0.10)	0.00	(0.01)	0.01	(0.03)	0.00	(0.02)	0.39	(0.11)	0.01	(0.05)	-0.08	(0.10)	···
4) AIC_W	1.95	(1.31)	0.04	(0.26)	-0.07	(0.18)	-0.07	(0.21)	-0.04	(0.21)	0.42	(0.26)	0.24	(0.36)	-0.24	(0.19)	···
FMA-Hansen-Schätzer																	
1) Original	0.71	(0.59)	0.38	(0.14)	-0.07	(0.09)	0.02	(0.10)	0.11	(0.13)	0.26	(0.16)	0.07	(0.11)	-0.06	(0.08)	···
2) CC	1.74	(0.72)	0.33	(0.26)	-0.08	(0.16)	-0.08	(0.18)	-0.04	(0.15)	0.12	(0.20)	0.17	(0.25)	-0.01	(0.02)	···
3) GAMRI	0.55	(0.60)	0.36	(0.15)	-0.04	(0.09)	0.06	(0.12)	0.07	(0.13)	0.29	(0.17)	0.04	(0.08)	-0.01	(0.04)	···
GLMRI	0.44	(0.60)	0.28	(0.14)	-0.08	(0.10)	0.01	(0.12)	0.11	(0.13)	0.33	(0.15)	0.22	(0.17)	-0.11	(0.10)	···
kNN	0.95	(0.60)	0.28	(0.14)	-0.08	(0.09)	0.04	(0.10)	0.08	(0.13)	0.29	(0.14)	0.17	(0.15)	-0.11	(0.10)	···
Amelia	1.23	(0.53)	0.29	(0.14)	-0.04	(0.09)	0.03	(0.10)	0.03	(0.12)	0.28	(0.15)	0.19	(0.15)	-0.15	(0.11)	···
MI	1.08	(0.79)	0.34	(0.15)	-0.03	(0.10)	-0.03	(0.16)	0.04	(0.15)	0.26	(0.16)	0.19	(0.18)	-0.13	(0.12)	···
FMS-AIC-Schätzer																	
1) Original	0.73	(0.38)	0.39	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.36	(0.10)	0.00	(0.00)	0.00	(0.00)	···
2) CC	1.68	(2.22)	0.08	(0.31)	-0.07	(0.30)	-0.24	(0.35)	-0.04	(0.39)	0.32	(0.31)	0.51	(0.44)	-0.24	(0.23)	···
3) GAMRI	0.59	(0.41)	0.38	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.36	(0.10)	0.00	(0.00)	0.00	(0.00)	···
GLMRI	0.30	(0.38)	0.36	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.45	(0.10)	0.00	(0.00)	0.00	(0.00)	···
kNN	1.42	(0.42)	0.29	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.42	(0.09)	0.00	(0.00)	-0.15	(0.08)	···
Amelia	1.62	(0.39)	0.29	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.42	(0.10)	0.00	(0.00)	-0.18	(0.08)	···
MI	1.25	(0.61)	0.34	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.39	(0.10)	0.00	(0.00)	-0.10	(0.12)	···
4) AIC_W	1.76	(1.83)	0.06	(0.29)	-0.14	(0.29)	-0.14	(0.34)	-0.09	(0.37)	0.36	(0.30)	0.48	(0.41)	-0.24	(0.22)	···

Tab. 7.2: Schätzung der Regressionsparameter (Teil I), in Klammern die zugehörigen Standardfehler gemäß (5.24)

	MEP		L		P2		P3		P4		LP2		LP3		LP4	
FMA-Akaike-Schätzer																
1) Original	...	0.00	(0.00)	-0.07	(0.23)	0.27	(0.28)	-0.24	(0.46)	0.06	(0.43)	-0.14	(0.23)	0.11	(0.27)	-0.36	(0.34)
2) CC	...	-0.02	(0.21)	0.43	(1.23)	0.41	(0.97)	0.36	(1.46)	0.05	(1.26)	-0.77	(1.24)	-0.45	(1.38)	-0.68	(1.30)
3) GAMRI	...	0.00	(0.00)	0.04	(0.26)	0.41	(0.32)	-0.16	(0.48)	0.38	(0.46)	-0.19	(0.27)	0.06	(0.31)	-0.56	(0.35)
GLMRI	...	0.00	(0.02)	0.26	(0.25)	0.52	(0.31)	0.00	(0.46)	0.47	(0.43)	-0.46	(0.28)	-0.17	(0.32)	-0.70	(0.34)
kNN	...	0.00	(0.00)	-0.28	(0.12)	0.12	(0.21)	0.05	(0.25)	-0.32	(0.29)	0.01	(0.05)	0.02	(0.07)	-0.02	(0.06)
Amelia	...	0.00	(0.00)	-0.29	(0.12)	-0.03	(0.22)	-0.13	(0.30)	-0.46	(0.30)	0.02	(0.08)	0.05	(0.12)	-0.03	(0.08)
MI	...	0.00	(0.01)	-0.26	(0.25)	0.10	(0.35)	-0.13	(0.45)	-0.17	(0.44)	0.02	(0.23)	0.12	(0.28)	-0.09	(0.22)
4) AIC_W	...	-0.01	(0.10)	0.02	(0.80)	0.11	(0.28)	0.07	(0.31)	-0.06	(0.19)	-0.20	(0.40)	-0.03	(0.27)	-0.07	(0.20)
FMA-Hansen-Schätzer																
1) Original	...	0.01	(0.05)	0.03	(0.16)	0.30	(0.28)	-0.12	(0.39)	-0.02	(0.47)	-0.14	(0.23)	0.01	(0.19)	-0.30	(0.36)
2) CC	...	0.00	(0.00)	0.00	(0.00)	-0.03	(0.36)	-0.03	(0.41)	-0.52	(0.49)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
3) GAMRI	...	-0.01	(0.04)	0.04	(0.14)	0.34	(0.28)	-0.07	(0.38)	0.06	(0.50)	-0.09	(0.19)	0.02	(0.17)	-0.28	(0.37)
GLMRI	...	0.01	(0.06)	0.16	(0.23)	0.40	(0.34)	0.07	(0.41)	0.19	(0.52)	-0.32	(0.33)	-0.16	(0.28)	-0.45	(0.42)
kNN	...	0.00	(0.06)	-0.20	(0.15)	0.16	(0.21)	0.10	(0.23)	-0.33	(0.27)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
Amelia	...	0.01	(0.07)	-0.22	(0.14)	0.03	(0.22)	-0.04	(0.23)	-0.47	(0.26)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
MI	...	0.01	(0.06)	-0.20	(0.27)	0.12	(0.36)	-0.04	(0.39)	-0.23	(0.41)	0.01	(0.20)	0.06	(0.22)	-0.06	(0.18)
FMS-AIC-Schätzer																
1) Original	...	0.00	(0.00)	0.01	(0.23)	0.36	(0.29)	-0.38	(0.49)	0.29	(0.36)	-0.20	(0.26)	0.17	(0.34)	-0.55	(0.28)
2) CC	...	-0.02	(0.21)	0.43	(1.23)	0.41	(0.97)	0.36	(1.46)	0.05	(1.26)	-0.77	(1.24)	-0.45	(1.38)	-0.68	(1.30)
3) GAMRI	...	0.00	(0.00)	0.09	(0.26)	0.46	(0.31)	-0.21	(0.51)	0.51	(0.39)	-0.23	(0.28)	0.08	(0.36)	-0.66	(0.29)
GLMRI	...	0.00	(0.00)	0.34	(0.22)	0.62	(0.27)	-0.06	(0.49)	0.64	(0.37)	-0.52	(0.25)	-0.16	(0.33)	-0.83	(0.27)
kNN	...	0.00	(0.00)	-0.29	(0.09)	0.10	(0.19)	0.07	(0.22)	-0.42	(0.24)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
Amelia	...	0.00	(0.00)	-0.29	(0.08)	-0.04	(0.19)	-0.07	(0.22)	-0.55	(0.24)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
MI	...	0.00	(0.00)	-0.29	(0.20)	0.06	(0.32)	-0.10	(0.48)	-0.29	(0.37)	0.07	(0.20)	0.11	(0.32)	-0.02	(0.14)
4) AIC_W	...	-0.02	(0.18)	0.33	(1.00)	0.16	(0.34)	0.08	(0.38)	-0.03	(0.23)	-0.34	(0.51)	-0.12	(0.38)	-0.14	(0.25)

Tab. 7.2: Schätzung der Regressionsparameter (Teil II), in Klammern die zugehörigen Standardfehler gemäß (5.24)

GLMRI-Verfahren insbesondere der Interaktion von Phase und Laisser-faire eine gewisse, wenn auch nicht besonders große, Bedeutung zu.

- Wie bereits in den Simulationen aus Kapitel 6 häufig zu beobachten, liegen die Ergebnisse einer Complete Case Analyse und einer AIC_W-Methodik – unabhängig vom gewählten Schätzer – in einer ähnlichen Größenordnung. Durchgehend sind sehr hohe Varianzschätzungen zu erkennen; insgesamt wird keine Größe als wirklich relevant eingestuft. Dies ist sicherlich nicht adäquat und liegt wohl dem Umstand zugrunde, dass beiden Verfahren nur etwas mehr als 30 Fälle zur Verfügung stehen.

- Offensichtlich sind die Ergebnisse des FMA-Hansen-Schätzers generell etwas differenzierter zu betrachten als die des FMA-Akaike- bzw. FMS-AIC-Schätzers: Die Schätzungen für die relevanten Größen IIA und IC sind im Allgemeinen etwas konservativer, die Schätzungen für alle anderen Größen etwas weniger stabil, sie hängen in gewissem Maße von der gewählten Methodik zur Berücksichtigung der fehlenden Werte ab. Diese größere Volatilität, diese weniger klaren Bekenntnisse für oder gegen eine Variable ist nicht überraschend: Die Simulationen aus den Abschnitten 6.1 und 6.3 haben dies bereits erahnen lassen.

- Wie oben bereits erwähnt, liegen die AIC-Werte der Modelle sehr nahe beisammen; trotzdem liegen die Schätzungen der Standardfehler für den FMA-Akaike- und den FMS-AIC-Schätzer in einer ähnlichen Größenordung. Dies deutet darauf hin, dass die verschiedenen konkurrierenden Modelle die Effekte der Variablen in etwa gleich stark einschätzen.

- Insgesamt lässt sich konstatieren, dass die Zufriedenheit der Mitarbeiter maßgeblich von den (transformationalen) Führungsstilen IIA und IC beeinflusst wird und weitgehend unabhängig von der entsprechenden Unternehmenslebensphase ist. IIA bezeichnet eine charismatische Führung, bei welcher die Führungspersönlichkeit ungeachtet ihrer Kompetenzen für den Erfolg und die Besonderheit eines Ergebnisses mitverantwortlich gemacht wird. Eine Führungspersönlichkeit mit hohem IIA-Wert steht sinnbildlich für eine Identifikationsfigur, die motivierend und begeisternd wirken kann. IC, Individualized Consideration, ist ein Führungsverständnis, welches die individuellen Charakteristika und Wesenszüge der Mitarbeiter erkennt und fördert. Höhere IC-Werte stehen für Führungskräfte, die gezielt die Potentiale ihrer Mitarbeiter entdecken und unterstützen. Sowohl ein IIA- als auch ein IC-Führungsverhalten wirkt positiv auf die Zufriedenheit.

Um nun die Qualität der verschiedenen Methoden zur Berücksichtigung der Problematik fehlender Daten genauer zu quantifizieren, werden die folgenden beiden Verlustfunktionen

$$L_9 = \sum_{j=0}^{p} (\hat{\tilde{\beta}}_j^{\star} - \hat{\tilde{\beta}}_j^{\text{org}})^2 , \qquad L_{10} = \sum_{j=1}^{p} (\hat{\tilde{\beta}}_j^{\star} - \hat{\tilde{\beta}}_j^{\text{org}})^2 ,$$

betrachtet, die die FMA- bzw. FMS-Schätzer $\hat{\tilde{\beta}}^{\star}$ und die entsprechenden Schätzer auf Basis der Originaldaten ohne fehlende Werte $\hat{\tilde{\beta}}^{\text{org}}$ über den MSE vergleichen – einmal unter Berücksichtigung (L_9) und einmal ohne Berücksichtigung (L_{10}) des Intercept. Die Resultate sind in den Tabellen 7.3 und 7.4 abgebildet. Es ist zu erkennen, dass un-

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	1.9128	0.1888	0.7141	0.4539	0.5735	0.2332	0.4832
FMA-Hansen	0.5366	0.0231	0.2328	0.3501	0.4852	0.2477	–
FMS-AIC	2.0032	0.1201	0.7032	1.2646	1.4561	0.9666	1.1731

Tab. 7.3: Übersicht über den Verlust L_9

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	2.5042	0.2503	0.9069	0.5556	0.9134	0.2903	1.5627
FMA-Hansen	1.5955	0.0514	0.3079	0.4058	0.7535	0.3810	–
FMS-AIC	2.8905	0.1400	0.8933	1.7407	2.2429	1.2381	2.2320

Tab. 7.4: Übersicht über den Verlust L_{10}

abhängig vom betrachteten Schätzer und unabhängig von der betrachten Verlustfunktion alle Imputations- und AIC_W-Methoden geringere Verluste aufweisen als eine Complete Case Analyse. Die besten Ergebnisse liefert dabei die GAMRI-Methodik. Für die Modellmittelungsschätzer lassen sich auch sehr gute Ergebnisse der multiplen Imputationen des Amelia II-Pakets erkennen. Insgesamt bestätigt sich das Bild der Simulationen, dass die Imputationsverfahren die Originaldaten deutlich besser erfassen können als eine ein Complete Case Analyse und dass die Schätzer auf Basis des AIC_W einen guten Korrekturansatz im linearen Modell liefern können, in der Regel aber etwas schlechter abschneiden als die Imputationsmethoden.

Da die Zielgröße einen Score beschreibt, stellt sich abschließend noch die Frage wie sehr die Annahmen, die an ein lineares Modell gestellt werden, an dieser Stelle gerechtfertigt sind. Die Annahme der Normalverteilung der Residuen, wie auch die Homoskedaszität der Varianzen für das als plausibel eingestufte Modell M_3 werden in Abbildung 7.1 überprüft. Der QQ-Plot für die Residuen zeigt, dass die Annahme der Normalverteilung

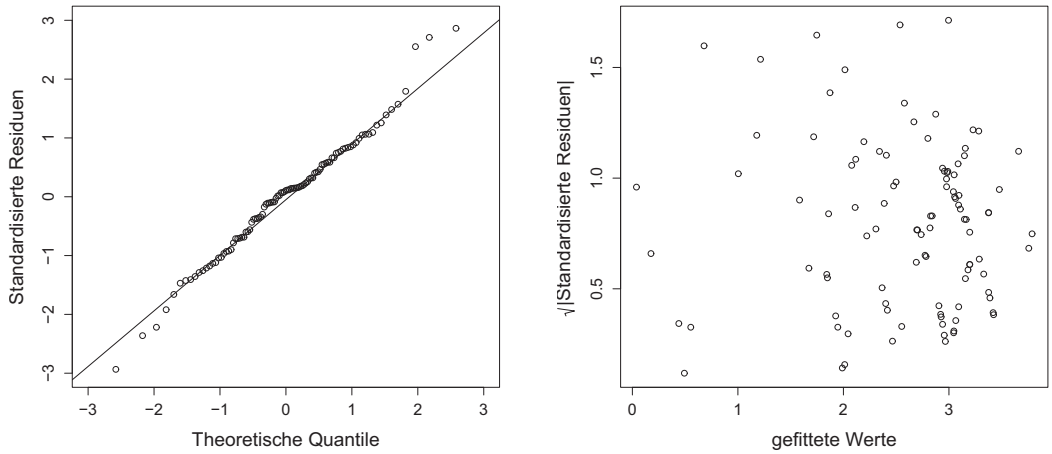


Abb. 7.1: QQ-Plot und Plot der gefitteten Werte gegen die Wurzel der Beträge der standardisierten Residuen für Modell 3 unter Verwendung der Originaldaten

an den Rändern etwas problematisch ist, insgesamt aber wohl noch im tolerierbaren Bereich liegt; der Plot der gefitteten Werte $\hat{y}_i$ gegen die Wurzel der Beträge der standardisierten Residuen ist ein Chaos-Plot; es deutet nichts auf eine Heteroskedaszität der Varianzen hin.

7.1.2 Analyse der Effektivität

In diesem Abschnitt wird der Erfolgsindikator Effektivität betrachtet. Es wird untersucht, durch welches Führungsverhalten sie besonders gesteigert bzw. vermindert werden kann und ob dies abhängig von der Unternehmenslebensphase ist oder nicht. Tabelle 7.5 zeigt die Korrelation nach Spearman zwischen der Effektivität und allen interessierenden Größen. Genau wie im vorhergehenden Abschnitt 7.1.1 ist zu erkennen, dass die transformationalen Führungsstile IIA, IIB, IM, IS und IC allesamt positiv wirken – hier nur eben im Bezug auf die Effektivität. Die transaktionalen Führungsstile CR, MEA und MEP

	IIA	IIB	IM	IS	IC	CR	MEA	MEP	L	Phase
Effektivität	0.67	0.39	0.45	0.50	0.55	0.56	0.18	-0.39	-0.45	-0.46

Tab. 7.5: Korrelation nach Spearman zwischen der Effektivität und den potentiellen Einflussgrößen

wirken sich erneut relativ unterschiedlich, sowohl positiv als auch negativ auf den Erfolg aus. Ein Laisser-faire-Führungsstil korreliert, ebenso wie die Unternehmenslebensphase, negativ mit der Effektivität.

Betrachtet wird eine lineare Regressionsanalyse mit der Effektivität als Response und allen anderen angeführten Größen als Kovariablen; die Unternehmenslebensphase wird dabei dummykodiert mit Phase 1 (=Entstehung) als Referenz. Um die vorgestellten Methoden zur Berücksichtigung fehlender Daten zu veranschaulichen, wird in diesem Beispiel sowohl ein MAR- als auch ein MNAR-Fehlendmechanismus eingeführt; für die Variablen IIA, IC, Laisser-faire und Phase werden gemäß der Fehlwahrscheinlichkeitsfunktionen

$$\begin{aligned} \pi_{\text{Phase}}(\text{Phase}) &= 1/(\text{Phase}^2 + 1) \\ \pi_{\text{IIA,IC,L}}(\text{Effektivität}) &= 1 - 1/(1 + 0.25 \cdot |\log(\text{Effektivität})|) \end{aligned}$$

Werte als fehlend erklärt.[22] Dies resultiert in 19.8 % fehlender Werte für IIA, 15.84 % fehlender Werte für IC, 14.85 % fehlender Werte für Laisser-faire und 18.81 % fehlender Werte für Phase; insgesamt fehlt bei 55.45% der Beobachtungen mindestens ein Wert, eine Complete Case Analyse kann also nur auf 45 Fälle zurückgreifen.

Betrachtet werden acht Kandidatenmodelle, von denen sechs dieselben sind wie bei Klaußner (2007, Seite 238) sowie zwei weitere Modelle (M_2 und M_7), die aufgrund ihres AIC-Werts als plausibel eingestuft werden können:

[22] Da die Wahrscheinlichkeit, dass ein Wert der Variable Phase fehlt, von der Phase selbst abhängt, liegt ein MNAR-Fehlendmechanismus vor. Die Motivation dafür folgt aus der Feststellung, dass einige Items des Fragebogens, so etwa die Frage nach der Größe des Rückgangs des Umsatzes oder nach der Intensität zunehmender Bürokratisierung, in jungen Unternehmen teilweise schwer zu beantworten ist und in einer früheren Phase deshalb eine höhere Fehlwahrscheinlichkeit angenommen werden könnte. Der MAR-Fehlendmechanismus für IIA, IC und L soll die Problematik nicht-konsistenter Parameterschätzungen aufgreifen, da die Fehlwahrscheinlichkeit vom Response – hier der Effektivität – abhängt.

$$
\begin{aligned}
\mathrm{M}_1: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_2 \mathrm{IIB} + \gamma_3 \mathrm{IM} + \gamma_4 \mathrm{IS} + \gamma_5 \mathrm{IC} + \gamma_6 \mathrm{CR} + \gamma_7 \mathrm{MEA} + \gamma_8 \mathrm{MEP} \\
& + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_2: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_6 \mathrm{CR} + \gamma_7 \mathrm{MEA} + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_3: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_7 \mathrm{MEA} + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_4: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_6 \mathrm{CR} + \gamma_7 \mathrm{MEA} + \gamma_9 \mathrm{L}\,, \\
\mathrm{M}_5: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_9 \mathrm{L} + \gamma_{10} \mathrm{P2} + \gamma_{11} \mathrm{P3} + \gamma_{12} \mathrm{P4}\,, \\
\mathrm{M}_6: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_7 \mathrm{MEA} + \gamma_9 \mathrm{L}\,, \\
\mathrm{M}_7: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_6 \mathrm{CR} + \gamma_9 \mathrm{L}\,, \\
\mathrm{M}_8: \quad y = \; & \alpha + \gamma_1 \mathrm{IIA} + \gamma_9 \mathrm{L}\,.
\end{aligned}
$$

Um die Daten zu analysieren und die in dieser Arbeit vorgestellten Methoden geeignet zu illustrieren, werden dieselben Schätzer wie im vorhergehenden Abschnitt betrachtet. Es ist klar, dass alle FMA-Hansen-Schätzer das Modell M_1 und die entsprechenden 12 Submodelle betrachten. Die Ergebnisse der Analyse befinden sich in Tabelle 7.6.

Resultate

Es lassen sich folgende Resultate bezüglich der inhaltlichen Interpretation, der drei betrachteten Schätzer und der verschiedenen Methoden zur Berücksichtigung der fehlenden Werte konstatieren:

- Betrachtet man die Schätzung der Regressionsparameter sowie deren geschätzte Standardfehler für die Originaldaten, so ist zu erkennen, dass alle Schätzer ein IIA-Führungsverhalten als besonders relevant für die Effektivität einstufen; die Parameterschätzungen sind dabei stets mindestens zweimal so groß wie der zugehörige Standardfehler. Der FMA-Akaike-, wie auch der FMS-AIC-Schätzer ordnen darüber hinaus einem Laisser-faire-Führungsstil eine größere Bedeutung zu. Dabei wirkt IIA positiv, der Laisser-fair-Stil negativ auf die Effektivität. Dies war aufgrund der betrachteten Korrelationen und den vorangegangenen Auswertungen zu erwarten. Alle anderen Größen scheinen keinen relevanten Effekt auf die Effektivität zu besitzen.

- Generell führen die Schätzungen auf Basis aller Imputationsmethoden zu in etwa den gleichen Aussagen wie die entsprechenden Schätzer auf Basis der Originaldaten: Sowohl IIA als auch Laisser-faire werden als relevant eingestuft, alle anderen Größen als irrelevant. Die Unterschiede zwischen den einzelnen Imputationsmethoden sind marginal. Betrachtet man beispielsweise den FMS-AIC-Schätzer, so wird die Phase von den verallgemeinerten Regressionsimputationen und der kNN-Methode als durchaus relevant eingestuft; unter Berücksichtigung der Modellselektionsunsicherheit und damit der Betrachtung des FMA-Akaike-Schätzers verschwinden diese Effekte jedoch vollständig.

- In diesem Beispiel liegen erneut die Complete Case Analyse, wie auch die Schätzer unter Verwendung des AIC_W in derselben Größenordnung. Insgesamt ist die Qualität der Schätzungen respektabel, sie liefern ebenfalls inhaltlich ähnliche Ergebnisse wie die Schätzungen auf Basis der Originaldaten. Es ist jedoch zu beobachten, dass die Schätzungen etwas konservativer sind: Für IIA ist bei den Originaldaten und den imputierten Daten die Parameterschätzung etwa vier- bis fünfmal so groß wie der zugehörige Standardfehler, bei der CC-Analyse und AIC_W-Methodik nur etwa dreimal so groß; für Laisser-faire sind die Parameterschätzungen auf Basis der Originaldaten bzw. der imputierten Daten etwa drei mal so groß wie der zugehörige Standardfehler, bei der Complete Case Analyse und den Schätzern unter Verwendung des AIC_W nur etwa zweimal so groß.

- Vergleicht man die Schätzer der Standardfehler zwischen den Imputationsmethoden mit nur einer Imputation (GAMRI, GLMRI, kNN, Amelia) und der multiplen Imputation (MI), so stellt man fest, dass diese ungefähr in einer ähnlichen Größenordnung liegen. Es liegt in diesem Beispiel also wenig Imputationsunsicherheit vor. Auch sind die Schätzungen der Standardfehler bei den Modellmittelungsschätzern oft nur marginal höher als bei dem betrachteten Selektionsschätzer. Dies deutet auf wenig Selektionsunsicherheit hin.

- Insgesamt lässt sich konstatieren, dass die Effektivität der Führung insbesondere von den den Führungsstilen IIA und Laisser-faire beeinflusst wird und unabhängig von der entsprechenden Unternehmenslebensphase ist. Es zeigt sich, dass sich eine besonders charismatische Führung (IIA) positiv, eine äußerst passive Führung (Laisser-faire) negativ auf den Erfolg auswirkt.

	I.		IIA		IIB		IM		IS		IC		CR		
FMA-Akaike-Schätzer															
1) Original	1.35	(0.38)	0.50	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.08	(0.11)	···
2) CC	1.40	(0.84)	0.37	(0.18)	-0.12	(0.19)	-0.04	(0.18)	-0.04	(0.24)	0.08	(0.22)	0.29	(0.25)	···
3) GAMRI	1.67	(0.46)	0.44	(0.11)	0.00	(0.00)	0.00	(0.01)	0.00	(0.00)	0.01	(0.01)	0.06	(0.10)	···
GLMRI	1.24	(0.39)	0.52	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.08	(0.11)	···
kNN	1.49	(0.45)	0.47	(0.10)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.07	(0.11)	···
Amelia	1.14	(0.36)	0.56	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.08	(0.11)	···
MI	1.45	(0.51)	0.47	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.09	(0.13)	···
4) AIC_W	0.94	(0.53)	0.39	(0.13)	-0.01	(0.02)	0.00	(0.01)	0.00	(0.01)	0.01	(0.03)	0.26	(0.18)	···
FMA-Hansen-Schätzer															
1) Original	1.44	(0.46)	0.43	(0.14)	0.00	(0.06)	0.02	(0.07)	0.01	(0.08)	0.04	(0.10)	0.03	(0.09)	···
2) CC	1.86	(0.60)	0.38	(0.18)	-0.03	(0.07)	-0.01	(0.05)	-0.01	(0.06)	0.02	(0.07)	0.07	(0.13)	···
3) GAMRI	1.59	(0.53)	0.31	(0.15)	-0.03	(0.08)	0.08	(0.11)	0.02	(0.10)	0.14	(0.16)	-0.04	(0.11)	···
GLMRI	1.36	(0.45)	0.46	(0.15)	-0.01	(0.06)	0.03	(0.08)	0.01	(0.08)	0.02	(0.09)	0.02	(0.08)	···
kNN	1.54	(0.51)	0.35	(0.16)	-0.02	(0.07)	0.06	(0.10)	0.03	(0.10)	0.10	(0.13)	-0.02	(0.10)	···
Amelia	1.43	(0.47)	0.50	(0.14)	0.00	(0.05)	0.03	(0.07)	0.01	(0.08)	0.01	(0.08)	0.04	(0.10)	···
MI	1.53	(0.50)	0.42	(0.15)	-0.01	(0.07)	0.04	(0.09)	0.02	(0.10)	0.00	(0.11)	0.06	(0.13)	···
FMS-AIC-Schätzer															
1) Original	1.08	(0.31)	0.49	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.16	(0.11)	···
2) CC	1.40	(0.84)	0.37	(0.18)	-0.12	(0.19)	-0.04	(0.18)	-0.04	(0.24)	0.08	(0.22)	0.29	(0.25)	···
3) GAMRI	1.77	(0.41)	0.42	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	···
GLMRI	1.35	(0.37)	0.50	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	···
kNN	1.62	(0.41)	0.45	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	···
Amelia	1.11	(0.28)	0.60	(0.08)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	···
MI	1.47	(0.50)	0.48	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.05	(0.13)	···
4) AIC_W	0.74	(0.41)	0.39	(0.12)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.29	(0.16)	···

Tab. 7.6: Schätzung der Regressionsparameter (Teil I), in Klammern die zugehörigen Standardfehler gemäß (5.24)

		MEA		MEP		L		P2		P3		P4	
FMA-Akaike-Schätzer													
1) Original	···	0.09	(0.09)	0.00	(0.00)	-0.34	(0.09)	-0.01	(0.06)	-0.11	(0.18)	-0.11	(0.18)
2) CC	···	0.12	(0.16)	0.06	(0.16)	-0.29	(0.15)	-0.27	(0.36)	-0.45	(0.41)	-0.44	(0.51)
3) GAMRI	···	0.07	(0.09)	0.00	(0.01)	-0.33	(0.11)	-0.02	(0.15)	-0.30	(0.31)	-0.31	(0.33)
GLMRI	···	0.10	(0.10)	0.00	(0.00)	-0.31	(0.09)	0.03	(0.10)	-0.12	(0.19)	-0.14	(0.21)
kNN	···	0.10	(0.10)	0.00	(0.00)	-0.34	(0.10)	-0.03	(0.13)	-0.24	(0.29)	-0.26	(0.31)
Amelia	···	0.09	(0.09)	0.00	(0.00)	-0.30	(0.09)	-0.02	(0.06)	-0.06	(0.11)	-0.06	(0.11)
MI	···	0.09	(0.10)	0.00	(0.00)	-0.33	(0.10)	-0.02	(0.15)	-0.22	(0.27)	-0.24	(0.32)
4) AIC_W	···	0.15	(0.12)	0.00	(0.01)	-0.29	(0.12)	-0.02	(0.06)	-0.04	(0.07)	-0.03	(0.06)
FMA-Hansen-Schätzer													
1) Original	···	0.08	(0.10)	-0.04	(0.07)	-0.15	(0.15)	-0.03	(0.19)	-0.39	(0.28)	-0.34	(0.28)
2) CC	···	0.03	(0.06)	0.01	(0.05)	-0.07	(0.12)	-0.39	(0.33)	-0.74	(0.45)	-0.81	(0.54)
3) GAMRI	···	0.09	(0.10)	-0.06	(0.08)	-0.14	(0.14)	-0.04	(0.23)	-0.47	(0.31)	-0.42	(0.35)
GLMRI	···	0.08	(0.11)	-0.04	(0.07)	-0.11	(0.14)	0.05	(0.21)	-0.37	(0.29)	-0.36	(0.29)
kNN	···	0.09	(0.11)	-0.06	(0.08)	-0.15	(0.14)	-0.06	(0.24)	-0.44	(0.33)	-0.44	(0.34)
Amelia	···	0.06	(0.09)	-0.05	(0.07)	-0.10	(0.12)	-0.18	(0.22)	-0.47	(0.31)	-0.43	(0.33)
MI	···	0.09	(0.10)	-0.05	(0.08)	-0.15	(0.16)	-0.07	(0.24)	-0.45	(0.29)	-0.44	(0.32)
FMS-AIC-Schätzer													
1) Original	···	0.12	(0.08)	0.00	(0.00)	-0.33	(0.09)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
2) CC	···	0.12	(0.16)	0.06	(0.16)	-0.29	(0.15)	-0.27	(0.36)	-0.45	(0.41)	-0.44	(0.51)
3) GAMRI	···	0.14	(0.09)	0.00	(0.00)	-0.30	(0.10)	0.01	(0.22)	-0.47	(0.25)	-0.47	(0.28)
GLMRI	···	0.17	(0.09)	0.00	(0.00)	-0.29	(0.09)	0.12	(0.18)	-0.28	(0.23)	-0.30	(0.24)
kNN	···	0.16	(0.09)	0.00	(0.00)	-0.31	(0.10)	-0.01	(0.22)	-0.44	(0.26)	-0.46	(0.28)
Amelia	···	0.15	(0.08)	0.00	(0.00)	-0.33	(0.08)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
MI	···	0.12	(0.11)	0.00	(0.00)	-0.33	(0.09)	0.01	(0.17)	-0.24	(0.31)	-0.28	(0.35)
4) AIC_W	···	0.18	(0.11)	0.00	(0.00)	-0.29	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)

Tab. 7.6: Schätzung der Regressionsparameter (Teil II), in Klammern die zugehörigen Standardfehler gemäß (5.24)

Analog zum vorhergehenden Abschnitt 7.1.1 sollen die Verlustfunktionen L_9 und L_{10} die Qualität der verschiedenen Methoden zur Berücksichtigung fehlender Werte quantifizieren. Die Tabellen 7.7 und 7.8 beinhalten die entsprechenden Resultate.

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	0.3841	0.0808	0.0040	0.0409	0.0103	0.0302	0.0622
FMA-Hansen	0.4911	0.0474	0.0104	0.0288	0.0460	0.0186	–
FMS-AIC	0.5295	0.4737	0.2126	0.4345	0.0386	0.1483	0.0321

Tab. 7.7: Übersicht über den Verlust L_9

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	0.3866	0.1832	0.0161	0.0605	0.0544	0.0402	0.2303
FMA-Hansen	0.6675	0.0699	0.0168	0.0388	0.0461	0.0267	–
FMS-AIC	0.6319	0.9498	0.2855	0.7261	0.0395	0.3004	0.1477

Tab. 7.8: Übersicht über den Verlust L_{10}

Auch in diesem Anwendungsbeispiel erwirken alle Strategien zur Berücksichtigung fehlender Werte eine Verbesserung gegenüber einer einfachen Complete Case Analyse. Insbesondere die verallgemeinerte Regressionsimputation auf Basis generalisierter Regressionsmodelle (GLMRI) sowie die Imputationen unter Verwendung des Amelia II-Pakets liefern besonders gute Ergebnisse.

Zuletzt sollen auch hier die Modellannahmen an das lineare Modell, insbesondere die Normalverteilungsannahme der Residuen und die Homoskedaszität der Varianzen, geprüft werden. Abbildung 7.2 zeigt den QQ-Plot und den Plot der gefitteten Werte gegen die Wurzel der Beträge der standardisierten Residuen für Modell 4 unter Verwendung der Originaldaten. Ähnlich dem vorangegangenen Beispiel ist die Normalverteilungsannahme an den Rändern etwas problematisch, insgesamt aber wohl auch hier noch im tolerierbaren Bereich; der Plot der gefitteten Werte $\hat{y}_i$ gegen die Wurzel der Beträge der standardisierten Residuen ist ein Chaos-Plot; es deutet nichts auf eine Heteroskedaszität der Varianzen hin.

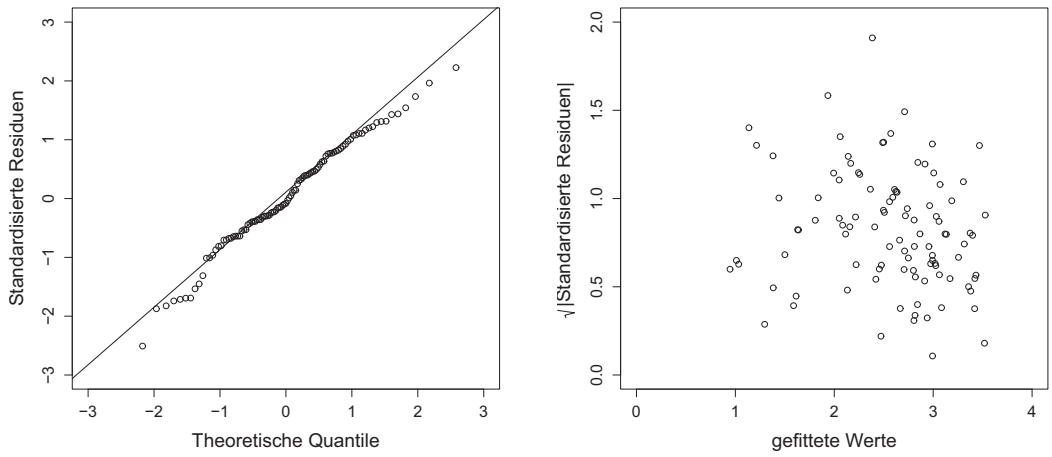


Abb. 7.2: QQ-Plot und Plot der gefitteten Werte gegen die Wurzel der Beträge der standardisierten Residuen für Modell 4 unter Verwendung der Originaldaten

7.2 Muskeldystrophie vom Typ Duchenne

Muskeldystrophie vom Typ Duchenne (DMD, Duchenne Muscular Dystrophy) ist eine genetisch hervorgerufene Muskelschwächekrankheit, die sich durch einen rapiden Muskelschwund in Armen, Beinen und andere Gliedmaßen bereits in sehr jungen Jahren (in der Regel vor dem sechsten Lebensjahr) bemerkbar macht; sie endet meist tödlich sobald die Herz- und Atemmuskulatur abgebaut wird. Obwohl auch Frauen DMD-Träger sein können, sind nur Männer durch die Muskeldystrophie beeinträchtigt und haben mit den meist schlimmen Folgen zu kämpfen. Unabhängig vom Geschlecht zeichnen sich Träger der Dystrophie vom Typ Duchenne durch erhöhte Werte bestimmter Proteine und Enzyme aus, beispielsweise *Creatine Kinase (CK), Hemopexin (H), Pyruvate Kinase (PK)* und *Lactate Dehydroginase (LD)*.

Der in diesem Abschnitt betrachtete Datensatz stammt aus Andrews und Herzberg (1985) und wurde in Grundzügen bereits in den Veröffentlichungen von Tibsharani und Hinton (1998) und Zhou, Wan und Wang (2008) analysiert. Er enthält 209 Beobachtungen, ausschließlich Frauen, von denen 75 DMD-Träger sind und 134 nicht. Der Response ist eine binäre Variable, bei der der Wert 1 einen Träger markiert und der Wert 0 einen Nicht-Träger. Potentielle Kovariablen sind neben dem Alter (AGE) auch das Niveau der

oben angesprochenen Proteine und Enzyme CK, H, PK und LD.[23] Die konkreten Werte für Creatine Kinase und Hemopexin lassen sich relativ kostengünstig aus gefrorenem Blutserum ermitteln, während für Pyruvate Kinase und Lactate Dehydroginase frisches Serum benötigt wird und die Bestimmung der Werte deswegen teurer ist. Dies ist auch der Grund weshalb sieben Werte von LD und acht Werte von PK fehlen; alle anderen Variablen sind vollständig beobachtet.

Eine Übersicht über die Verteilung der Werte der Kovariablen für DMD- und Nicht-DMD-Träger befindet sich in Tabelle 7.9 und 7.10. Offensichtlich sind die Level für die Proteine und Enzyme bei DMD-Trägern erhöht.

kein DMD	AGE	CK	H	PK	LD
Minimum	20.00	15.00	34.00	2.80	66.00
1. Quartil	25.00	26.25	76.30	9.73	136.00
Median	27.00	34.00	82.50	11.90	162.00
arith. Mittel	28.81	39.13	82.95	12.15	164.60
3. Quartil	32.75	45.00	91.00	15.30	185.00
Maximum	39.00	130.00	111.50	22.70	349.00

Tab. 7.9: Deskriptive Statistiken für die Gruppe der nicht an Muskeldystrophie erkrankten Personen

Es stellt sich die Frage, wie sich die Chancen ein DMD-Träger zu sein über ein logistisches Regressionsmodell am besten modellieren lassen, ob dafür bestimmte Proteine und Enzyme besonders geeignet sind und ob auf die teure Bestimmung der Level von PK und LD verzichtet werden kann oder nicht. Die oben angeführte Arbeit von Tibsharani und Hinton (1998) verwendet zur Illustration einiger ad-hoc Methoden für die Modellselektion auf Basis von Vorhersagefehlern die Werte von allen Kovariablen und schätzt für das entsprechende logistische Regressionsmodell die Effekte von CK, H und PK insgesamt am stärksten ein. Zhou, Wan und Wang (2008) demonstrieren ihre Inferenzprozeduren für Schätzgleichungen im Kontext fehlender Daten am selben Datensatz, verwenden aus nicht näher erläuterten Gründen als Kovariablen jedoch nur das Alter sowie LD. Insofern sind die in der Literatur bisher bekannten Resultate als durchaus

[23] das Alter ist in Jahren angegeben; die Einheit für CK und LD ist U/L (Unit je Liter) und misst die Enzymaktivität je Liter, vergleiche etwa Woods et al. (2004, Seite 276) für Details; die Enzymaktivität für PK wird dagegen in U/g (Units je Gramm) gemessen; H ist in mg/dl (Milligramm Hämoglobin je Deziliter) angegeben.

DMD	AGE	CK	H	PK	LD
Minimum	20.00	19.00	9.00	8.30	122.00
1. Quartil	30.50	54.00	83.50	14.30	198.00
Median	35.00	101.00	91.00	19.40	245.00
arith. Mittel	38.13	187.20	86.68	23.93	256.20
3. Quartil	42.50	235.00	100.30	25.25	297.00
Maximum	61.00	1288.00	118.00	110.00	593.00

Tab. 7.10: Deskriptive Statistiken für die Gruppe der an Muskeldystrophie erkrankten Personen

widersprüchlich einzuschätzen. Im Folgenden werden fünf logistische Regressionsmodelle betrachtet, die insbesondere auch die Interaktionseffekte zwischen CK und H sowie PK und LD berücksichtigen:

$$
\begin{aligned}
\mathrm{M}_1: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 \mathrm{AGE} + \gamma_2 \mathrm{CK} + \gamma_3 \mathrm{H} + \gamma_4 \mathrm{PK} + \gamma_5 \mathrm{LD} + \gamma_6 \mathrm{CK} \times \mathrm{H} + \gamma_7 \mathrm{PK} \times \mathrm{LD}\,, \\
\mathrm{M}_2: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 \mathrm{AGE} + \gamma_2 \mathrm{CK} + \gamma_3 \mathrm{H} + \gamma_4 \mathrm{PK} + \gamma_5 \mathrm{LD} + \gamma_6 \mathrm{CK} \times \mathrm{H}\,, \\
\mathrm{M}_3: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 \mathrm{AGE} + \gamma_2 \mathrm{CK} + \gamma_3 \mathrm{H} + \gamma_4 \mathrm{PK} + \gamma_5 \mathrm{LD} + \gamma_7 \mathrm{PK} \times \mathrm{LD}\,, \\
\mathrm{M}_4: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 \mathrm{AGE} + \gamma_2 \mathrm{CK} + \gamma_3 \mathrm{H} + \gamma_4 \mathrm{PK} + \gamma_5 \mathrm{LD}\,, \\
\mathrm{M}_5: \quad \ln \tfrac{p_i}{1-p_i} &= \alpha + \gamma_1 \mathrm{AGE} + \gamma_2 \mathrm{CK} + \gamma_3 \mathrm{H}\,.
\end{aligned}
$$

Um die in dieser Arbeit vorgestellten Methoden zur Modellselektion und Modellmittelung im Kontext fehlender Daten geeignet zu illustrieren, werden neben den bereits vorhandenen fehlenden Werten für PK und LD weitere Werte für diese beiden Kovariablen über einen MAR-Fehlendmechanismus[24] gemäß

$$
\pi_{\mathrm{PK,LD}}(y) \;=\; \{3 + \exp(\tilde{y} + 0.5 \cdot [\mathrm{sign}(\tilde{y}) + 1] + \epsilon)\}^{-1}\,,
$$

als fehlend deklariert, wobei $\tilde{y} = y - 0.5$ und $\epsilon \sim N(0, 0.5)$. Damit fehlen nun 22.01 % der Daten von PK und 22.49% der Daten von LD. Insgesamt fehlt bei 40.19% der

[24] Der in diesem Beispiel gewählte Fehlendmechanismus ist besonders interessant, da die Fehlwahrscheinlichkeit für PK und LD nur vom Response abhängt und damit – wie zu Beginn von Kapitel 5 erläutert – die Parameterschätzer der logistischen Regression auf Basis einer Complete Case Analyse nicht mehr konsistent sind.

Beobachtungen mindestens ein Wert. Eine Complete Case Analyse kann also nur auf 125 Fälle zurückgreifen.

Zur Veranschaulichung der in dieser Arbeit vorgestellten Verfahren werden die FMA-Akaike- und FMS-AIC-Schätzer wie im vorhergehenden Abschnitt 7.1 betrachtet, also die entsprechenden 14 Schätzer aus Tabelle 6.1, (a) und (c), und ferner der FMA-Akaike- bzw. FMS-AIC-Schätzer nach multipler Imputation gemäß Tabelle 6.7, (a) und (c). Der FMA-Hansen-Schätzer wird hier nicht weiter betrachtet, da seine Optimalitätseigenschaften nur im Kontext linearer Regression gelten, vergleiche auch Abschnitt 4.2.2. Der Referenzschätzer auf Basis der Originaldaten kann in diesem Beispiel nur auf die vollständigen Fälle vor Einführung des zusätzlichen Fehlendmechanismus, dies sind 194 Fälle, zurückgreifen. Die Resultate der Analyse befinden sich in Tabelle 7.11.

Resultate

Es lassen sich folgende Resultate bezüglich der inhaltlichen Interpretation, der betrachteten Schätzer und der verschiedenen Methoden zur Berücksichtigung der fehlenden Werte konstatieren:

- Betrachtet man den FMA-Akaike- und FMS-AIC-Schätzer für die Originaldaten, so fällt auf, dass nicht alle Effekte der betrachteten Proteine und Enzyme positiv sind. Die Schätzungen von CK und H sind negativ; der positive Interaktionseffekt $CK \times H$ wiegt dies jedoch wieder auf, so dass erhöhte Werte – wie zu erwarten – für ein erhöhtes DMD-Risiko sprechen. Berücksichtigt man zusätzlich die Schätzungen der Standardfehler, so zeigt sich, dass der FMS-AIC-Schätzer insbesondere PK, LD und CK (in Interaktion mit H) als relevant einstuft; die Parameterschätzungen sind hierbei jeweils etwa doppelt so groß wie der zugehörige Standardfehler. Die Betrachtung des FMA-Akaike-Schätzers offenbart, dass in diesem Beispiel eine große Modellselektionsunsicherheit vorherrscht und die Berücksichtigung derselben zu deutlich konservativeren Schätzungen führt; die Effekte sind bei weitem nicht so stark ausgeprägt wie durch die angegebene Literatur bzw. den Modellselektionsschätzer zu erwarten.

- Alle Imputationsmethoden modellieren das Risiko DMD-Träger zu sein etwas anders als die Originaldaten: Sowohl für die Modellselektion als auch für die Modellmittelung werden alle betrachteten Interaktionseffekte als nicht relevant erkannt. Der Effekt von CK wird als Ausgleich dazu positiv bewertet, der Effekt von H

dagegen als vernachlässigbar. Die Parameterschätzer von PK und LD liegen in etwa derselben Größenordnung wie bei den Originaldaten.

- Wie bereits in den vorangegangenen Simulationen und Beispielen häufig zu erkennen, führt die Verwendung des AIC_W zu qualitativ ähnlichen Ergebnissen wie die Verwendung der Complete Cases. Im Unterschied zu den Originaldaten wird in diesem Beispiel – unabhängig von der Betrachtung des FMS-AIC- bzw. des FMA-Akaike-Schätzers – der Interaktionseffekt $CK \times H$ noch stärker positiv modelliert; entsprechend liegen die Schätzungen für CK und H weit im negativen Bereich.

- Wie zu erwarten sind die Schätzungen der Standardfehler bei den Schätzern der Modellmittelung im Trend höher als bei den Schätzern der Modellselektion. Dies führt dazu, dass die Stärke der Effekte weitaus geringer eingeschätzt wird als man aufgrund der Resultate in der einschlägigen Literatur erwarten konnte. Insgesamt scheint der Effekt von Lactate Dehydroginase am stärksten ausgeprägt zu sein.

- Vergleicht man die Schätzungen der Standardfehler für einfache und multiple Imputationen, so ist zu erkennen, dass die Unsicherheit bezüglich der Imputation nicht besonders hoch ist.

- Unter Berücksichtigung aller oben angeführten Sachverhalte scheinen die Effekte von CK, PK und LD insgesamt am stärksten zu sein. Ein Prognosemodell sollte insbesondere unter Berücksichtigung der Ergebnisse der Originaldaten sowohl PK als auch LD beinhalten, obschon deren Effekte nicht so stark ausgeprägt sind wie eventuell zu erwarten.

Zur Beurteilung der Qualität der Schätzungen werden erneut die Verlustfunktionen L_9 und L_{10} betrachtet. Es ist an dieser Stelle jedoch zu beachten, dass die FMA- bzw. FMS-Schätzer auf Basis der Originaldaten, wie bereits angedeutet, nur auf 194 der 209 Beobachtungen zurückgreifen können und die entsprechenden Verluste daher mit Vorsicht zu bewerten sind. Die Resultate befinden sich in den Tabellen 7.12 und 7.13.

Es bestätigt sich das in dieser Arbeit weithin gewonnene Bild, dass Korrekturverfahren für den Umstand fehlender Daten in der Regel bessere Ergebnisse und damit geringere Verluste liefern als eine simple CC-Analyse. Dabei sind die Ergebnisse der Schätzer unter Verwendung von Imputationen meist besser und stabiler als die Schätzer unter Verwendung des AIC_W. Die Unterschiede zwischen den einzelnen Imputationsmethoden sind in diesem Beispiel marginal. Weitere Anmerkungen und Gedanken für die Analyse des vorliegenden Datensatzes finden sich auch bei Schomaker, Wan und Heumann (2010).

	Intercept		AGE		CK		H		
FMA-Akaike-Schätzer									
1) Original	−9.097	(8.342)	0.162	(0.047)	−0.243	(0.175)	−0.061	(0.089)	···
2) CC	1.854	(9.320)	0.171	(0.066)	−0.662	(0.272)	−0.203	(0.123)	···
3) GAMRI	−13.827	(3.312)	0.175	(0.046)	0.024	(0.027)	−0.006	(0.024)	···
GLMRI	−13.411	(3.332)	0.173	(0.046)	0.021	(0.030)	−0.008	(0.026)	···
kNN	−13.057	(3.269)	0.173	(0.043)	0.021	(0.026)	−0.007	(0.023)	···
Amelia	−13.712	(3.099)	0.181	(0.045)	0.032	(0.027)	0.000	(0.024)	···
MI	−12.611	(3.524)	0.174	(0.045)	0.022	(0.036)	−0.007	(0.029)	···
4) AIC_W	2.527	(9.306)	0.173	(0.066)	−0.679	(0.267)	−0.211	(0.120)	···
FMS-AIC-Schätzer									
1) Original	−5.132	(6.708)	0.166	(0.047)	−0.297	(0.160)	−0.088	(0.081)	···
2) CC	1.193	(7.572)	0.173	(0.066)	−0.657	(0.245)	−0.202	(0.112)	···
3) GAMRI	−13.943	(2.361)	0.173	(0.045)	0.035	(0.013)	0.002	(0.014)	···
GLMRI	−13.492	(2.276)	0.170	(0.045)	0.035	(0.012)	0.002	(0.015)	···
kNN	−13.082	(2.212)	0.170	(0.042)	0.031	(0.012)	0.001	(0.014)	···
Amelia	−14.396	(2.432)	0.178	(0.044)	0.043	(0.012)	0.008	(0.014)	···
MI	−13.381	(2.603)	0.171	(0.044)	0.039	(0.012)	0.005	(0.017)	···
4) AIC_W	1.193	(7.572)	0.173	(0.066)	−0.657	(0.245)	−0.202	(0.112)	···

Tab. 7.11: Schätzung der Regressionsparameter (Teil I), in Klammern die zugehörigen Standardfehler gemäß (5.24)

	PK		LD		PK×LD		CK×H	
FMA-Akaike-Schätzer								
1) Original	...	0.227	0.020	(0.014)	−0.001	(0.001)	0.003	(0.002)
2) CC	...	0.045	0.026	(0.016)	0.000	(0.001)	0.009	(0.003)
3) GAMRI	...	0.188	0.019	(0.010)	0.000	(0.000)	0.000	(0.000)
GLMRI	...	0.177	0.020	(0.010)	0.000	(0.000)	0.000	(0.000)
kNN	...	0.176	0.019	(0.010)	0.000	(0.001)	0.000	(0.000)
Amelia	...	0.115	0.015	(0.007)	0.000	(0.000)	0.000	(0.000)
MI	...	0.122	0.015	(0.008)	0.000	(0.000)	0.000	(0.000)
4) AIC_W	...	0.037	0.026	(0.016)	0.000	(0.001)	0.009	(0.003)
FMS-AIC-Schätzer								
1) Original	...	0.115	0.011	(0.006)	0.000	(0.000)	0.004	(0.002)
2) CC	...	0.085	0.028	(0.012)	0.000	(0.000)	0.009	(0.003)
3) GAMRI	...	0.155	0.017	(0.006)	0.000	(0.000)	0.000	(0.000)
GLMRI	...	0.133	0.017	(0.006)	0.000	(0.000)	0.000	(0.000)
kNN	...	0.138	0.016	(0.005)	0.000	(0.000)	0.000	(0.000)
Amelia	...	0.120	0.016	(0.006)	0.000	(0.000)	0.000	(0.000)
MI	...	0.110	0.015	(0.005)	0.000	(0.000)	0.000	(0.000)
4) AIC_W	...	0.085	0.028	(0.012)	0.000	(0.000)	0.009	(0.003)

Tab. 7.11: Schätzung der Regressionsparameter (Teil II), in Klammern die zugehörigen Standardfehler gemäß (5.24)

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	0.229	0.076	0.075	0.076	0.093	0.084	0.195
FMS-AIC	0.144	0.120	0.119	0.116	0.125	0.122	0.114

Tab. 7.12: Übersicht über den Verlust L_9

	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	120.149	22.444	18.684	15.752	21.390	12.434	113.884
FMS-AIC	40.142	77.763	70.003	63.317	85.943	68.168	38.863

Tab. 7.13: Übersicht über den Verlust L_{10}

7.3 Olympischer Zehnkampf

Die multivariate Analyse von Sportdaten besitzt eine lange Tradition in der Statistik. Viele Auswertungen beschäftigen sich dabei mit der Frage der Dimensionalität der Disziplinen des Sieben- bzw. Zehnkampfs unter Verwendung von Resultaten internationaler Leichtathletikveranstaltungen, vergleiche hierzu auch Dawkins, Andreae und O'Connor (1994) und Cox und Dunn (2002). Die Statistik hält für eine solche Fragestellung eine Vielzahl multivariater Verfahren bereit, beispielsweise Clusteranalysen, Faktorenanalysen, Hauptkomponentenanalysen, Korrespondenzanalysen, wie auch einfache grafische Darstellungen. In diesem Abschnitt werden Zehnkampfdaten der olympischen Spiele von Athen vom 23.8./24.8.2004 betrachtet und mit Hilfe einer Maximum-Likelihood-Faktorenanalyse unter Berücksichtigung von Modellselektionsunsicherheit und der Problematik fehlender Daten, mit den Methoden wie in den Abschnitten 4.2.5 und 5.2.2 beschrieben, ausgewertet und kritisch diskutiert; die betrachteten Daten, dargestellt in Tabelle C.9, enthalten dabei die Resultate aller 30 Athleten, die den Wettbewerb beendet haben, also die Ergebnisse der Teilnehmer in den zehn Disziplinen *100m, Weitsprung, Kugelstoßen, Hochsprung, 400m, 110m-Hürden, Diskuswurf, Stabhochsprung, Speerwurf* und *1500m*, gemessen jeweils in Metern bzw. Sekunden.

Typischerweise sind Zehnkampfdaten in dem Sinne hochdimensional, als dass das Verhältnis der Stichprobengröße zu der Anzahl der betrachteten Variablen relativ gering ist. Aufgrund dieser Datenstruktur muss damit gerechnet werden, dass jede durch statisti-

sche Verfahren gefundene Struktur instabil sein kann und insofern eine gewisse Modellselektionsunsicherheit vorherrscht. Diese Vermutung wird auch durch die Ergebnisse von Cox und Dunn (2002) belegt, die fünf verschiedene Datensätze mit Zehnkampfdaten aus den Jahren 1991-1999 über eine hierarchische Clusteranalyse (mit Ward-Linkage) analysieren und für die entsprechenden Datensätze jeweils äußerst verschiedene Cluster bezüglich der Variablen erhalten. Die Autoren kombinieren diese unterschiedlichen Resultate gewissermaßen ad-hoc und gelangen so zu drei verschiedenen Kernclustern, wie in Tabelle 7.14 beschrieben. Eine inhaltliche Einordnung und Interpretation dieser Resultate erfolgt nach Betrachtung der Ergebnisse der untenstehend durchgeführten Maximum-Likelihood-Faktorenanalyse zu Ende dieses Abschnitts.

Cluster	Disziplinen			
Cluster 1	100 m	400 m	Weitsprung	110 m Hürden
Cluster 2	Kugelstoßen	Diskuswurf	Speerwurf	Stabhochsprung
Cluster 3	Hochsprung	1500 m		

Tab. 7.14: Resultate der Clusteranalyse von Cox und Dunn (2002, S. 181)

Von den 30 vorliegenden Beobachtungen können 28 als vollständig angesehen werden und zwei als unvollständig, da sie aufgrund zu vieler Fehlversuche keine konkreten Ausprägungen bezüglich des Leistungsvermögens der Athleten besitzen. Dies betrifft sowohl das ungültige Kugelstoßergebnis von Eugene Martineau aus den Niederlanden als auch das ungültige Stabhochsprungresultat von Victor Covalenco aus Moldawien, vergleiche Tabelle C.9. Um diesen Umstand explizit in der Faktorenanalyse zu berücksichtigen, werden die FMS- und FMA-Schätzer (4.16) für die Ladungsmatrix Γ und die Einzelrestvarianz Ψ unter Verwendung des $\mathrm{AIC}_{\mathrm{FA}}$ gemäß (4.15)[25] sowohl für die vollständigen Fälle X_{CC} als auch für die imputierten Datensätze X_{GLMRI} bzw. X_{kNN} auf Basis einer GLMRI- bzw. kNN-Imputation berechnet; vergleiche auch Tabelle 7.15.

Die verallgemeinerte Regressionsimputation GLMRI imputiert einen Kugelstoßwert von 15.00 Metern für Eugene Martineau und einen Stabhochsprungswert von 4.65 Metern für Victor Covalenco; die k-Nächste-Nachbarn-Imputation ermittelt auf Basis der zwei nächsten Nachbarn Ergebnisse von 14.42 Metern bzw. ebenfalls 4.65 Metern für die feh-

[25] Meist stellt statistische Software Akaikes Informationskriterium in den faktoranalytischen Prozeduren nicht explizit zur Verfügung. Der Zusammenhang des AIC zur χ^2-Statistik kann im Kontext der Faktorenanalyse jedoch leicht zur Implementierung genutzt werden, vergleiche hierzu auch Akaike (1987, S. 321).

Auswahl an FMA- und FMS-Schätzern $\hat{\hat{\Gamma}}$ und $\hat{\hat{\Psi}}$ bei fehlenden Daten

(a) **FMA-Akaike-Schätzer**: die FMA-Schätzer (4.16), die auf Basis der exponentiellen AIC-Gewichte (4.6) den Umstand fehlender Daten wie folgt berücksichtigen:

1) *CC* – FMA-Schätzer unter Verwendung des Subdatensatzes der vollständig beobachteten Fälle $D^c_* = X_{\text{CC}}$.
2) *GLMRI* – entspricht dem FMA-Schätzer (5.20) unter Verwendung des aufgefüllten Datensatzes $D^{\text{imp}} = X_{\text{GLMRI}}$.
3) *kNN* – entspricht dem FMA-Schätzer (5.20) unter Verwendung des aufgefüllten Datensatzes $D^{\text{imp}} = X_{\text{kNN}}$.

(b) **FMS-AIC-Schätzer**: die FMS-Schätzer für die $\Gamma = \text{AIC}_{\text{FA}}$ und die den Umstand fehlender Daten wie folgt berücksichtigen:

1) *CC* – entspricht dem FMS-Schätzer (5.16).
2) *GLMRI* – entspricht dem FMS-Schätzer (5.8) unter Verwendung des aufgefüllten Datensatzes $D^{\text{imp}} = X_{\text{GLMRI}}$.
3) *kNN* – entspricht dem FMS-Schätzer (5.8) unter Verwendung des aufgefüllten Datensatzes $D^{\text{imp}} = X_{\text{kNN}}$.

Tab. 7.15: Die für die Zehnkampfdaten verwendeten Modellmittelungsschätzer und Modellselektionsschätzer für verschiedene Strategien zur Berücksichtigung der Problematik fehlender Daten

lenden Werte. Aufgrund des geringen Stichprobenumfangs im Verhältnis zu der Anzahl der Variablen und der damit verbundenen Problematik in der Schätzung generalisierter additiver Modelle liefert der GAMRI-Algorithmus keine Imputationen; das Amelia II-Paket stellt die Sinnhaftigkeit der gefundenen Ergebnisse aufgrund der schwachen Datengrundlage über einen Warnhinweis in Frage, weswegen diese beiden, in den bisherigen Auswertungen intensiv betrachteten Imputationsmethoden, im Weiteren nicht näher analysiert werden.

Konkret betrachtet werden an dieser Stelle nun die fünf faktoranalytischen Kandidatenmodelle $X' = \Gamma^{(k)} F^{(k)} + U$, $k = 1, \ldots, 5$, wobei X abhängig von der gewählten Methodik die $n \times 10$ Datenmatrix der vollständigen Fälle bzw. der imputierten Datensätze beschreibt, $\Gamma^{(k)}$ die entsprechende $10 \times k$ Ladungsmatrix und $F^{(k)}$ die $k \times n$ Matrix der k Faktoren darstellt. Untersucht wird also, ob sich die zehn Disziplinen des Zehnkampfs am besten über eine, zwei, drei, vier oder fünf Dimensionen erfassen lassen.

Resultate

In Tabelle 7.16 sind die Akaike-Gewichte (4.6) für die fünf Kandidatenmodelle abhängig von der gewählten Methodik dargestellt. M_1 beschreibt dabei das Modell mit $k = 1$ Faktor, M_2 das Modell mit $k = 2$ Faktoren, et cetera. Es lässt sich erkennen, dass die

	M_1	M_2	M_3	M_4	M_5
Complete Cases	0.00	0.25	0.73	0.01	0.00
GLMRI	0.00	0.35	0.62	0.02	0.00
kNN	0.00	0.39	0.60	0.01	0.00

Tab. 7.16: Akaike-Gewichte zur Modellmittelung

Modelle mit einem, vier bzw. fünf Faktoren, unabhängig von der gewählten Methodik zur Berücksichtigung der fehlenden Werte, wenig Erklärungskraft zur Beschreibung der Dimension eines Zehnkampfs liefern. Sowohl für die Analyse der vollständigen Fälle als auch für die Analyse der imputierten Datensätze besitzt das 3-Faktor-Modell die höchste Erklärungskraft, das 2-Faktor-Modell eine etwas geringere. Dabei lassen sich durchaus kleinere Unterschiede für die gewählten Methoden feststellen: Für die Complete Case Analyse erhält das Modell M_2 ein Gewicht von $w_2 = 0.25$, während für die Imputationsmethoden – insbesondere für das kNN-Verfahren – die Gewichte und damit die Evidenz des entsprechenden Modells weitaus höher eingeschätzt werden. Die Auswirkungen auf die entsprechenden FMA-Schätzer $\hat{\bar{\Gamma}}$ und $\hat{\bar{\Psi}}$ lassen sich in den Tabellen 7.17, 7.18 und 7.19 erkennen. Tabelle 7.17 beschreibt dabei die gewichtete Ladungsmatrix und die gewichteten Einzelrestvarianzen für den imputierten Datensatz unter Verwendung des GLMRI-Algorithmus, Tabelle 7.18 die entsprechenden Resultate für die kNN-Imputationen und Tabelle 7.19 die Modellmittelungsschätzer für die CC-Analyse. Die Schätzungen der Ladungen beruhen dabei stets auf einer Varimax-Rotation.[26] In Anhang C sind unter anderem alle konkreten Schätzungen für die Ladungsmatrizen und die Einzelrestvarianzen des 2- und 3-Faktor-Modells gemäß der betrachteten Methoden aufgelistet; Die Tabellen C.2, C.4 und C.6 zeigen dabei die Resultate der oben angeführten FMS-Schätzer, also die durch das $\mathrm{AIC}_{\mathrm{FA}}$ bestimmten Schätzungen $\hat{\Gamma}_3$ und $\hat{\Psi}_3$ des 3-Faktor-Modells gemäß den Daten X_{CC}, X_{GLMRI} und X_{kNN}. Alle $\gamma_{ii} \in \Gamma$ und $\Psi_i \in \Psi$

[26] Einige in dieser Arbeit nicht näher erläuterten Sensitivitätsanalysen erlauben den Schluss, dass für das vorliegende Anwendungsbeispiel die Wahl der Rotationsmethode nicht entscheidend ist und sich die Ergebnisse für andere Rotationsmethoden nur in Nuancen verändern.

mit Werten größer als 0.5 sind in den Tabellen unterstrichen, um die Variablen, die am höchsten auf die entsprechenden Faktoren laden, herauszustellen.

	Ladungsmatrix Γ					Ψ
100 m	0.81	-0.29	-0.04	0.00	0.00	0.24
Weitsprung	-0.80	0.06	-0.02	0.00	0.00	0.36
Kugelstoßen	-0.16	0.96	0.05	0.00	0.00	0.05
Hochsprung	-0.25	0.63	-0.01	0.00	0.00	0.54
400 m	0.76	-0.19	0.25	0.00	0.00	0.29
110 m Hürden	0.64	-0.26	0.00	0.00	0.00	0.52
Diskuswerfen	-0.20	0.65	0.15	0.01	0.00	0.49
Stabhochsprung	-0.29	-0.04	0.15	-0.01	0.00	0.88
Speerwurf	0.03	0.46	-0.18	0.00	0.00	0.73
1500 m	0.15	0.16	0.63	0.00	0.00	0.32

Tab. 7.17: FMA-Schätzungen $\hat{\hat{\Gamma}}$ und $\hat{\hat{\Psi}}$ unter Verwendung einer GLMRI-Imputation

	Ladungsmatrix Γ					Ψ
100 m	0.81	-0.29	-0.05	0.00	0.00	0.25
Weitsprung	-0.78	0.10	-0.02	0.00	0.00	0.38
Kugelstoßen	-0.18	0.92	0.05	0.00	0.00	0.12
Hochsprung	-0.23	0.65	-0.01	0.00	0.00	0.52
400 m	0.78	-0.16	0.24	0.00	0.00	0.28
110 m Hürden	0.63	-0.29	0.00	0.00	0.00	0.52
Diskuswerfen	-0.17	0.72	0.13	0.01	0.00	0.42
Stabhochsprung	-0.30	-0.07	0.15	-0.01	0.00	0.87
Speerwurf	0.02	0.44	-0.18	0.00	0.00	0.75
1500 m	0.17	0.19	0.60	0.00	0.00	0.34

Tab. 7.18: FMA-Schätzungen $\hat{\hat{\Gamma}}$ und $\hat{\hat{\Psi}}$ unter Verwendung einer kNN-Imputation

Betrachtet man die FMA-Schätzungen, so lässt sich feststellen, dass alle aufgeführten Resultate generell in einer ähnlichen Größenordnung liegen. Dabei erkennt man trotz Berücksichtigung des 2-Faktor-Modells insgesamt drei prägende Faktoren:

- Die Disziplinen 100m, Weitsprung, 400m und 110m Hürden laden hoch auf den ersten Faktor. Dabei gibt es nur marginale Unterschiede zwischen den verschiedenen Methoden zur Berücksichtigung der fehlenden Daten. Dieser Faktor könnte als Geschwindigkeits- und Athletikkomponente des Zehnkampfs interpretiert werden.

	Ladungsmatrix Γ					Ψ
100 m	0.80	-0.26	-0.06	0.00	0.00	0.29
Weitsprung	-0.83	0.15	0.01	0.00	0.00	0.29
Kugelstoßen	-0.14	0.89	0.02	0.00	0.00	0.19
Hochsprung	-0.21	0.65	-0.05	0.00	0.00	0.53
400 m	0.80	-0.09	0.30	0.00	0.00	0.23
110 m Hürden	0.61	-0.20	0.06	0.00	0.00	0.58
Diskuswerfen	-0.11	0.74	0.10	0.01	0.00	0.42
Stabhochsprung	-0.30	-0.09	0.20	0.00	0.00	0.84
Speerwurf	0.04	0.41	-0.26	0.00	0.00	0.74
1500 m	0.25	0.17	0.72	0.00	0.00	0.22

Tab. 7.19: FMA-Schätzungen $\hat{\hat{\Gamma}}$ und $\hat{\hat{\Psi}}$ unter Verwendung der vollständigen Fälle

- Die Disziplinen Kugelstoßen, Hochsprung und Diskuswurf laden hoch auf den zweiten Faktor. Auch hier gibt es nur marginale Unterschiede zwischen den verschiedenen Methoden zur Berücksichtigung der fehlenden Daten. Dieser Faktor könnte als Kraft- und Technikkomponente interpretiert werden.

- Auf den dritten Faktor lädt nur der 1500m-Lauf hoch. Dabei sind die Schätzungen der Imputationsmethoden aufgrund ihres höheren Gewichts für das 2-Faktor-Modell etwas konservativer als die Schätzung der Complete Case Analyse. Dieser Faktor könnte sowohl die Ausdauer als auch den speziellen Status der letzten Disziplin repräsentieren.[27]

Bei der Interpretation dieser Ergebnisse und der Einschätzung ihrer Sensitivität sind einige wichtige Sachverhalte zu beachten:

- Sowohl der Stabhochsprung als auch der Speerwurf laden auf keinen der Faktoren besonders hoch. Ihre Einzelrestvarianzen sind entsprechend hoch; dies bedeutet, dass ein Großteil ihrer Streuung nicht durch das entsprechende faktoranalytische

[27] Da nach den ersten neun Disziplinen die Rangfolge der Athleten, insbesondere in der Nähe der Medaillenränge, in vielen Fällen sehr deutlich ausgeprägt ist und nur noch geringfügige Veränderungen erwartet werden können, wird das eigentliche Leistungsvermögen in der letzten Disziplin – dem 1500m-Lauf – nicht mehr von jedem Sportler voll ausgeschöpft. Dies verdeutlicht den speziellen Status der letzten Disziplin. Als einprägendes Beispiel hierfür lässt sich der Zehnkampf der olympischen Spiele 2008 in Peking anführen: Der US-Athlet Bryan Clay dominierte die ersten neun Disziplinen des Wettbewerbs. Aufgrund seines deutlichen Vorsprungs zu Platz 2 landete er siegesgewiss im abschließenden 1500m-Lauf auf dem letzten Platz.

Modell erklärt werden kann. Dies mag mit dem hohen technischen Anspruch der beiden Disziplinen zu erklären sein. Es ist jedoch auch zu erkennen, dass der Speerwurf zu einem geringen Anteil auf den zweiten Faktor lädt und dabei dasselbe Vorzeichen wie die Disziplinen, die bei diesem Faktor hoch laden, besitzt. Dies ergibt Sinn und gibt dem Speerwurf ein gewisses Technik-Attribut.

- Die Interpretation der FMS-Schätzer, dargestellt in den Tabellen C.2, C.4 und C.6, führt – unter inhaltlichen Gesichtspunkten – prinzipiell zu denselben Schlüssen wie die Interpretation der FMA-Schätzer. Dies liegt vor allem daran, dass der dritte Faktor sehr klar ausgeprägt ist; bei allen betrachteten FMS-Schätzern wird die entsprechende Ladung stets größer als 0.96 geschätzt, so dass die durchaus vorhandene Erklärungskraft des 2-Faktor-Modells nicht dazu ausreicht, die Schätzung der Ladungen für den entsprechenden FMA-Schätzer unter die in diesem Beispiel ad-hoc vorgegebene Relevanz von 0.5 zu drücken.

- Wie in Abschnitt 4.2.5 beschrieben, unterliegt die ML-Faktorenanalyse gewissen Beschränkungen, etwa bei der Schätzung der Einzelrestvarianzen, wo durch den iterativen Schätzprozess oder besondere Datenbegebenheiten, wie beispielsweise kleine Stichproben, ein oder mehrere Ψ_i kleiner Null geschätzt werden können. Die in diesem Abschnitt verwendete Prozedur der statistischen Software *R* (`factanal()`) setzt für den Optimierungsprozess als untere Schranke $\Psi_{i,\min} = 0.005$ fest. Die Schätzungen bei denen diese untere Grenze erreicht wird, sind in den entsprechenden Tabellen exakt mit 0.00 versehen. Die für den Modellmittelungsschätzer relevanten Modelle M_2 und M_3 weisen solche geringen Einzelrestvarianzen nur vereinzelt und nur für den 1500m-Lauf auf, vergleiche Anhang C. Dies führt an dieser Stelle für die finalen Resultate zu keiner besonders großen Problematik, da der dritte Faktor sehr deutlich ausgeprägt ist, inhaltlich nachvollziehbar ist und die entsprechende Einzelrestvarianz somit keine große Bedeutung besitzt. Die weiteren Fälle, bei denen die untere Schranke erreicht wird (vgl. Tabelle C.7 und C.8), sind irrelevant, da das entsprechende 5-Faktor-Modell aufgrund des geringen Akaike-Gewichts nicht in den Modellmittelungsschätzer eingeht.

- Es stellt sich die Frage, wie sensitiv die vorgestellten Ergebnisse bezüglich der Stichprobengröße sind. Wie in Abschnitt 3.3.4 erläutert, sind die in der Herleitung des AIC verwendeten Approximationen für kleine Stichproben gegebenenfalls sehr ungenau. Die entsprechenden Korrekturen nach Sugiura (1978) bzw. Hurvich und Tsai (1989) sind im faktoranalytischen Kontext nicht gültig. Insofern muss davon

ausgegangen werden, dass für die vorliegenden Daten das AIC keine unverzerrte Schätzung der Kullback-Leibler-Distanz mehr darstellt. Dies mag für die relativ eindeutigen Ergebnisse an dieser Stelle vernachlässigbar sein, kann aber generell Probleme, auch für die entsprechenden FMA-Schätzer, nach sich ziehen.

- Die Problematik der Datenstruktur, die sich in den Schwierigkeiten der GAMRI- bzw. Amelia II-Imputationsmethoden widerspiegelt, verdeutlicht die Herausforderung der Modellselektion und Modellmittelung in der Faktorenanalyse unter Berücksichtigung fehlender Daten. Auch wenn nur zwei der 30 Beobachtungen (ca. 7% der Daten) fehlende Werte aufweisen, so lassen sich bereits hierfür durchaus relevante Unterschiede zwischen dem FMA-Schätzer bei einer Complete Case Analyse und den FMA-Schätzern für die Imputationsmethoden konstatieren. Eine höhere Anzahl an fehlenden Werten bei einer ähnlichen Datenstruktur, kann zu weiteren Schwierigkeiten für geeignete Imputationsverfahren führen und die Sensitivität der vorgestellten Methoden erhöhen.

- Ein Vergleich mit den Ergebnissen von Cox und Dunn (2002), wie in Tabelle 7.14 vorgestellt, liefert einige interessante Aspekte: Die Grundstruktur von drei relevanten Faktoren bestätigt sich trotz unterschiedlicher Datengrundlage in beiden Auswertungen. Die Gestalt dieser Dimensionen wird von der Cluster- und der FMA-Faktorenanalyse unterschiedlich bewertet. Inhaltlich erscheinen die oben angeführten Resultate der Maximum-Likelihood-Faktorenanalyse plausibler; es ergibt keinen Sinn anzunehmen, dass beispielsweise der Hochsprung und der 1500m-Lauf bzw. der Stabhochsprung und das Kugelstoßen dieselben Dimensionen abbilden wie von Cox und Dunn (2002) vermutet. Diese Einteilung hängt natürlich auch stark mit dem Kernkonzept der Clusteranalyse an sich zusammen: Alle zehn Disziplinen müssen einem Cluster zugeordnet werden, so dass häufig eine leichte Ähnlichkeit zwischen verschiedenen Variablen genügt um sie zu einem Cluster zusammenzufassen. Das Ausmaß und die Stärke dieser Ähnlichkeit ist in den finalen Resultaten jedoch nicht mehr zu erkennen, weswegen schwache Strukturen häufig nicht mehr als solche erkannt werden können. Die Betrachtung einer Faktorenanalyse, im Speziellen unter Berücksichtigung der vorgestellten Modellmittelungsschätzer, erlaubt dagegen eine fundierte Einordnung der Stärke des Beitrags der einzelnen Variablen zu den gefundenen Faktoren: Je höher die Ladungen, desto mehr trägt eine Variable zu dem entsprechenden Faktor bei. Ist die Einzelrestvarianz sehr groß, so kann die Variabilität dieser Variable nicht adäquat durch das faktoranalytische Modell erfasst werden, wodurch die Erklärungskraft der Variablen zusätzlich eingeschätzt

werden kann. Wird statt eines Selektionsschätzers ein Mittelungsschätzer verwendet so ist zu erwarten, dass die Volatilität der Ergebnisse in Grenzen gehalten werden kann.

Die in diesem Abschnitt vorgestellten Resultate zeigen, dass sich die Grenzen von Modellmittelungsverfahren jenseits der Regressionsanalyse befinden und die Berücksichtigung von Selektionsunsicherheit prinzipiell in weiten Teilen statistischer Modellierung möglich ist. So lassen sich die Kernkonzepte aus Kapitel 4, insbesondere der Transfer dieser Methoden zur Faktorenanalyse – wie in Abschnitt 4.2.5 aufgezeigt und hier verwendet–, in weiten Teilen einfach und zielgerecht umsetzen. Die Interpretation der FMA-Schätzer ist, wie hier am Beispiel der Ladungsmatrix und der Einzelrestvarianzen gesehen, in der Regel unproblematisch.

Dennoch ergeben sich einige kritische Punkte, die verdeutlichen, dass die Sinnhaftigkeit von Modellmittelungsverfahren außerhalb der Regressionsanalyse teilweise deutlich eingeschränkter ist und die Sensitivität der Ergebnisse von einigen technischen Begebenheiten geprägt ist, deren Problematik in zukünftigen Arbeiten noch intensiver diskutiert werden muss. Dies betrifft zum einen die durch den Selektionsprozess verursachte zusätzliche Variabilität, die durch die Verwendung von Modellmittelungsverfahren erfasst werden soll: Im Fall multivariater Verfahren, wie beispielsweise der Faktorenanalyse, sind häufig vor allem die Punktschätzungen interessant. Die zugehörigen Varianzschätzungen, etwa für die Elemente der geschätzten Ladungsmatrix $\hat{\Gamma}$, können zwar konstruiert werden, vergleiche etwa Fahrmeir, Hamerle und Tutz (1996, Abschnitt 11.2.3), stehen aufgrund des explorativen Charakters der Verfahren jedoch nicht im Vordergrund und damit im Fokus des Interesses. Insofern stellt sich die Frage, welchen zusätzlichen Informationsgewinn die vorgestellten Verfahren in diesem Kontext überhaupt erwirken können. Zu einem gewissen Maß gehört dazu sicherlich eine stabile und wenn möglich unverzerrte Punktschätzung, die die angesprochene Unsicherheit sinnhaft reflektiert. Auch hier darf jedoch bemerkt werden, dass die Anzahl der Kandidatenmodelle in der Maximum-Likelihood-Faktorenanalyse typischerweise gering ausfällt, wohl sehr selten im zweistelligen Bereich liegt. Im Gegensatz zur Regressionsanalyse, wo bedingt durch Transformationen, Interaktionen und auch unterschiedlichen Verteilungsannahmen für den Response eine große Anzahl an plausiblen Modellen vorliegen kann, wird die Selektionsunsicherheit in vielen konkreten Anwendungen verhältnismäßig gering sein. Ferner wirken untypische, schwer zu modellierende Unsicherheitskomponenten auf die finalen Resultate, so etwa die Entscheidung zugunsten eines bestimmten Rotationsprinzips, was – wenn auch nicht im vorgestellten Beispiel – die Ergebnisse zu

einem gewissen Teil beeinflussen kann. Ebenfalls problematisch erscheinen die technischen Grenzen, die sich bei einer kleinen Stichprobengröße, wie etwa in diesem Beispiel, ergeben können. Dies führt zu teils irregulären oder fragwürdigen Ergebnissen für die Einzelrestvarianzen und kann verteilungsbasierte Imputationen wie beispielsweise den GAMRI-Algorithmus bzw. korrekte multiple Imputationen über das Amelia II-Paket erschweren oder auch verhindern.

Die Verwendung des AIC_W zur Modellselektion oder zur Bestimmung des FMA-Schätzers (5.17) ist prinzipiell auch in der Faktorenanalyse möglich. Darauf wurde in diesem Abschnitt bei der Analyse der olympischen Zehnkampfdaten jedoch verzichtet, da die für die gewichtete Likelihood notwendigen Gewichte (5.5) in der Regel über additive Modelle geschätzt werden, was an dieser Stelle aufgrund der geringen Stichprobengröße nicht möglich war. Dies unterstreicht erneut die Herausforderung in der Konstruktion passender Verfahren für die Berücksichtigung fehlender Daten und der Modellselektionsunsicherheit auch im Kontext der Faktorenanalyse. Welchen Erfolg ein gewichtetes Akaike-Kriterium an dieser Stelle versprechen kann, müssen zukünftige Arbeiten zeigen.

8. Résumé

Die in dieser Arbeit gewonnenen Erkenntnisse sind äußerst vielfältig und in vielerlei Hinsicht aufschlussreich für die Beurteilung des Erfolgs unterschiedlicher Strategien zur Berücksichtigung fehlender Werte im Kontext von Modellwahl und Modellmittelung. Auch wenn Resultate auf Basis von Monte-Carlo-Simulationen und Datenbeispielen immer einer gewissen Beschränkung unterliegen und nur mit Vorsicht auf Situationen außerhalb der betrachteten übertragen werden können, so ergeben sich doch Verhaltensweisen der verschiedenen Schätzer, die relativ durchgängig beobachtet und erklärt werden können und insofern auch einige typische Charakteristika aufweisen:

- Ungeachtet der Problematik fehlender Daten lässt sich festhalten, dass eine einfache, kriteriumsbasierte Modellmittelungsstrategie, wie die Verwendung exponentieller AIC-Gewichte und damit des FMA-Akaike-Schätzers, innerhalb einer linearen oder logistischen Regressionsanalyse zu sehr guten Ergebnissen führt. In der Regel erhält man bei einem solchen Vorgehen bessere Punktschätzungen als bei den entsprechenden Selektionsschätzern. Die zugehörigen Varianzschätzungen sind in der Regel ebenfalls von besserer Qualität. Wie vielfach diskutiert, ist die Konstruktion von FMA-Schätzern unter Optimalitätseigenschaften durchaus erstrebenswert, auch unter Berücksichtigung des bisher relativ wenig einheitlichen Vorgehens innerhalb der frequentistischen Modellmittelung. Der in dieser Arbeit betrachtete MMA-Schätzer von Hansen (2007) führt jedoch in vielen Fällen zu keiner Verbesserung im Vergleich zu einer simplen Modellselektion, etwa unter Verwendung des AIC. Dies mag häufig, jedoch nicht immer, an seiner restriktiven Auswahl der Kandidatenmodelle liegen.

- Werden die fehlenden Werte einer Datenmatrix schlicht verworfen, also nur die vollständigen Fälle (CC) zur Analyse verwendet, so führt dies in den betrachteten Regressionsbeispielen, unabhängig von dem gewählten Selektions- oder Mittelungsverfahren, bei einem MAR-Fehlendmechanismus zu minderwertigen, insgesamt relativ unbefriedigenden Schätzungen. Dies war zu erwarten. Die vorliegende Arbeit

zeigt jedoch an vielen Stellen, dass bei einer CC-Strategie in der Regel immerhin die entsprechenden Varianzschätzungen von passabler Qualität sind.

- Die Verwendung des für fehlende Daten adjustierten AIC_W, entweder direkt zur Selektion des Modells oder alternativ zur Konstruktion von Modellmittelungsgewichten, führt insbesondere im Kontext des linearen Modells meist zu besseren Punktschätzungen als eine Complete Case Analyse, aber auch zu etwas schlechteren Ergebnissen als die Imputationsverfahren. In den betrachteten Situationen der logistischen Regressionsanalyse kann die AIC_W-Methodologie jedoch häufig keinen zusätzlichen Gewinn erbringen. Berechnet man die Varianz des FMA-AIC_W-Schätzers unter Berücksichtigung der Modellselektionsunsicherheit wie in (5.19) angeführt, so ist eine deutliche Unterschätzung der Varianz zu erkennen. Wie in dieser Arbeit bereits vielfach erläutert, müsste korrekterweise eigentlich noch die Unsicherheit bei der Wahl des Glättungsparameters und den Kovariablen für das zur Schätzung der Gewichte (5.5) notwendige GAM berücksichtigt werden. Dies zeigt, dass die Konzepte des *inverse probability weighting* innerhalb des Themenkomplexes von Modellwahl und -mittelung zu einer weiteren Unsicherheitskomponente führen, deren Modellierung gegebenenfalls sehr aufwändig sein kann.

- Die fehlenden Werte einer Datenmatrix zu ersetzen und anschließend die entsprechenden FMS- und FMA-Schätzer zu berechnen, kann als relativ weitläufige Strategie angesehen werden, da sowohl die Imputations- als auch die Selektions- bzw. Mittelungsverfahren frei gewählt werden können, was insbesondere in einem so schwer einzugrenzenden Feld wie der Modellwahl einen großen Vorteil darstellt. Insgesamt lassen sich in den angeführten Beispielen sehr stabile und gute Punktschätzungen erkennen, wobei deren Varianz bei Verwendung nicht-multipler Imputationen aufgrund der nicht modellierten Imputationsunsicherheit prinzipiell etwas unterschätzt wird. Die in dieser Arbeit verwendeten Imputationsverfahren stehen stellvertretend für eine Vielzahl an Methoden unterschiedlichster Konzeption. Häufig liefert dabei die Amelia-Methode, also die Verwendung des *R*-Pakets „Amelia II“, sehr gute und stabile Ergebnisse. In einigen Situationen, etwa bei Betrachtung komplexerer Abhängigkeitsstrukturen oder zufälligen Effekten, kann jedoch auch die Verwendung der GAMRI-,GLMRI- bzw. kNN-Methodik zu den besten Resultaten führen. Prinzipiell kann nicht konstatiert werden, dass bei Gebrauch des in Abschnitt 5.1.2 vorgestellten verallgemeinerten, rekursiven Regressionsimputationsalgorithmus die Verwendung von generalisierten additiven Model-

len (GAMRI) der Verwendung simpler generalisierter linearer Regressionsmodelle (GLMRI) vorzuziehen ist. Der Erfolg dieser beiden Methoden variiert mit dem gewählten Setting.

- Die Verwendung von korrekten multiplen Imputationen unter Verwendung des Amelia II-Pakets der statistischen Software *R* kann prinzipiell zu recht guten und realistischen Schätzungen führen, insbesondere in Bezug auf den FMA-Akaike-Schätzer. Selbst dort kann jedoch unter gewissen Umständen ein interessantes Artefakt beobachtet werden: Wirkt der Fehlendmechanismus auf eine Variable, die einen komplexeren Effekt, etwa einen Interaktionseffekt, mitgestaltet und ist die Imputationsunsicherheit groß, so wird diese Unsicherheit wie gewünscht explizit mitmodelliert und kann zu einer Überschätzung der Varianz führen. Auch wenn nach einer Reflexion über diese Methodik ein solches Verhalten des Schätzers als plausibel erscheint, so ist dies bei der Interpretation konkreter Anwendungsbeispiele zu beachten.

Es hat sich gezeigt, dass die vorgestellten Konzepte zur Berücksichtigung fehlender Daten im Kontext von Modellselektion und Modellmittelung für lineare und logistische Regressionsanalysen einfach umzusetzen sind und in den gewählten Anwendungsbeispielen in Anbetracht ihrer oben angeführten Charakteristika zu entsprechend guten bzw. weniger guten Resultaten führen, auf jeden Fall aber gut interpretierbar sind. Sehr einfache Ansätze (kriteriumsbasierte Modellmittelungsgewichte, Imputationen aus vorhandener Software) scheinen dabei oft zu bereits sehr guten Resultaten zu führen.

Etwas differenzierter gestaltet sich der Blick bei Verwendung der Verfahren in der Faktorenanalyse. Die vorliegende Arbeit macht deutlich, weshalb sich die aktuelle Literatur vorwiegend einfachster Regressionsanalysen bedient, um ihre Methoden zu illustrieren: Eine Übertragung der grundlegenden Konzepte ist zwar prinzipiell auf viele Typen statistischer Modellierung möglich und den Umständen entsprechend technisch unkompliziert; liegen der Interpretationsschwerpunkt und die generelle Konzipierung etwas weiter von den Prinzipien der Regressionsanalyse entfernt, so ergeben sich jedoch neue Schwierigkeiten, Unsicherheitskomponenten und gegebenenfalls auch ein geringerer Informationsgewinn. In der Faktorenanalyse wird dies besonders deutlich, da hier in der Regel die Punkt- und nicht die zugehörigen Varianzschätzungen im Vordergrund stehen und insofern die Modellierung zusätzlicher, durch die Modellselektionsunsicherheit hervorgerufener, Unsicherheit keinen wirklichen Gewinn erbringt. Auch ist die Anzahl der konkurrierenden Kandidatenmodelle im Vergleich zur Regressionsanalyse, wo In-

teraktionen, logarithmische und quadratische Effekte oft zu einer großen Auswahl an Modellierungsmöglichkeiten führen, sehr beschränkt, was den Erfolg ebenfalls limitiert.

Ausblick

Wie die Problematik fehlender Daten im Kontext von Modellmittelungsverfahren berücksichtigt werden kann und welche Chancen, aber auch Schwierigkeiten, die vorgestellten Schätzer bieten, hat die vorliegende Arbeit aufgezeigt. Viele Ideen sind sehr allgemein formuliert worden und eröffnen damit die Möglichkeit, jenseits einfacher linearer und logistischer Regressionsanalysen verwendet zu werden. Tatsächlich stellt sich die Frage, wie einfach die Konzepte auf komplexere Modelle übertragen werden können.

In den letzten Jahren haben einige erste Arbeiten aufgezeigt, welchen Erfolg neuere (frequentistische) Modellmittelungskonzepte in anderen Bereichen versprechen: Hjort und Claeskens (2006) verwenden kriteriumsbasierte Gewichte unter Verwendung des Focused Information Criterion (FIC) von Claeskens und Hjort (2003), um einen FMA-Schätzer im Cox-Modell zu konstruieren, Hansen (2008b) überträgt einige Kerngedanken seines MMA-Schätzers (Hansen (2007)) auf den Kontext spezieller autoregressiver Prozesse und Zhang, Wan und Zhou (2010) verwenden die Ideen von Claeskens und Hjort (2003) und Hjort und Claeskens (2006) im Kontext des Tobit-Modells. Sowohl die „Mittelung nach Imputation“-Strategie als auch die Idee des *inverse probability weighting* könnten auch hier beim Vorhandensein fehlender Werte verwendet werden. Welchen Erfolg die in dieser Arbeit behandelten Imputationsmethoden jedoch versprechen und wie das Gewichtungskonzept konkret etwa auf das FIC angewendet werden kann, bedarf noch einer genaueren, möglicherweise simulationsgestützten Untersuchung.

Auch stellt sich die Frage, wie Modellmittelung und fehlende Daten, etwa im Bereich nonparametrischer Regression, kombiniert werden können. Die einfachste Möglichkeit besteht sicher darin, multiple Imputationen und exponentielle Gewichte auf Basis von Vorhersagefehlern zu verwenden, etwa mit Hilfe des generalisierten Kreuzvalidierungskriteriums, das für viele nonparametrische Ansätze – etwa bei additiven Modellen – von statistischer Software standardmäßig ausgegeben wird. Diskussionswürdig erscheint dagegen der Gegenstand der Mittelung: Soll über den gesamten Prädiktor gemittelt werden, ähnlich dem Vorgehen beim Bagging (Breimann (1996b))? Dies würde sicherlich einen Teil der Selektionsunsicherheit abdecken, für den Großteil der Verfahren im Bereich nonparametrischer Regression umsetzbar sein und möglicherweise auch zu besseren Vorhersagen führen. Damit kann aber zum einen keine Interpretation der einzelnen Effekte

und Variablen erfolgen und zum anderen wird die Unsicherheit bezüglich der Wahl des Glättungsparameters und der Variablen nicht ausreichend abgedeckt. Der erstgenannte Punkt kann sicher dahingehend berücksichtigt werden, dass nicht über den gesamten Prädiktor, sondern über die einzelnen Effekte oder bei Verwendung stückweiser Polynome sogar über die Parameterschätzungen gemittelt wird. Konkret kann die Entscheidung wohl von der bevorzugten Schätzmethodik und dem Ziel der Analyse abhängig gemacht werden. Der zweite Punkt lässt sich möglicherweise erfassen, indem doppelt kombiniert wird, was bedeutet, dass die Mittelung über einen Grid erfolgt, der sowohl eine Variablen- und Effektauswahl als auch mögliche Smoothing-Parameter enthält. Ob sich ein solch großer Aufwand überhaupt lohnt, ist aber sicherlich fraglich.

Insgesamt erscheinen auch unabhängig vom Auftreten fehlender Daten die Konzepte frequentistischer Modellmittelung im Gegensatz zur bayesianischen Herangehensweise noch nicht wirklich einheitlich und in sich konsistent. Unter welchem Paradigma die Modelle kombiniert werden sollen und welche Eigenschaften und Verteilung diese Schätzer besitzen, ist bisher nicht ausreichend geklärt. Speziell für die Konstruktion von Konfidenzintervallen ist dies jedoch besonders interessant. Hjort und Claeskens (2003) zeigen zwar, dass innerhalb eines lokalen Misspezifikationskonstrukts und unter gewissen Regularitätsvoraussetzungen die asymptotische Verteilung eines Modellmittelungsschätzers einer konvexen Kombination von Normalverteilungen entspricht, Leeb und Pötscher (2006b) zweifeln die Nützlichkeit dieses Resultats jedoch an, da die Konvergenz gegen die wahre Verteilungsfunktion nicht gleichmäßig, sondern nur punktweise in θ gilt und damit für jede fixe Stichprobengröße die wahren und geschätzten Verteilungen potentiell weit voneinander entfernt liegen können. Ob diese Aussagen tatsächlich von praktischer Relevanz sind und ob im Kontext fehlender Werte für die Berechnung von Konfidenzintervallen nicht die Verwendung von simplen Normalverteilungsquantilen, wie auch in Burnham und Anderson (2002) vorgeschlagen, zusammen mit den Standardfehlern, wie in Kapitel 5 beschrieben, eine ausreichende Approximation liefern, müssen zukünftige Arbeiten zeigen. Die Simulationsergebnisse aus Abschnitt 6.3 lassen zumindest erahnen, dass die Berechnung der Standardfehler gemäß (5.24) unter Verwendung korrekter multipler Imputationen für einen FMA-Akaike-Schätzer sehr gute Ergebnisse erzielen kann und die entsprechenden Konfidenzintervalle damit zumindest in einem einigermaßen sinnvollen und adäquaten Bereich liegen können.

Prinzipiell scheinen noch viele Erweiterungen der vorgeschlagenen Methodik möglich: So kann durchaus angedacht werden, die angeführten Korrekturverfahren für fehlende Daten im Bereich bayesianischer Modellmittelung zu untersuchen. Dort stellen sich

natürlich weiterhin die in den Abschnitten 3.4 und 4.1 aufgeworfenen, offenen Fragen. Um es kurz zu sagen: Wieviel Bayes soll sein? Weniger bis gar kein Bayes, indem die SBC-Approximation zur schnellen Berechnung des BMA-Schätzers verwendet wird und damit keine priori-Wahrscheinlichkeiten $p(M_\kappa)$ spezifiziert und interpretiert werden müssen, oder mehr Bayes, indem mit Laplace-Approximationen genauere, aber auch aufwändigere Lösungen gesucht werden. Eine Verwendung von (ggf. multiplen) Imputationsmethoden ist hier immer möglich, auch ein Gewichtungsansatz könnte insbesondere bei Verwendung von exponentiellen SBC-Gewichten konstruiert werden.

Auch ist weiterhin offen, ob die multiplen Imputationen über den im Amelia II-Paket implementierten Bootstrap-Ansatz zu qualitativ ähnlichen Ergebnissen führen wie multiple Imputationen auf Basis eines streng bayesianischen Vorgehens, wie etwa unter Verwendung des IP-Algorithmus. Möglicherweise gibt es an dieser Stelle noch etwas Spielraum um die Varianzschätzungen nach (5.24) noch etwas stabiler zu machen und die Überschätzung der Standardfehler in komplexen Situation eher zu unterbinden. Die Analyse und die Konstruktion von Varianzschätzern scheint im Rahmen frequentistischer Modellmittelung allgemein noch ausbaufähig zu sein. Ein Großteil aktueller Artikel, so etwa von Yang (2003), Yuan und Yang (2005), Hansen (2007, 2008a,b, 2009), Magnus, Powell und Prüfer (2008), Liang et al. (2010), Hansen und Racine (2009) und Zhang, Wan und Zhou (2010) betrachtet vor allem die Qualität der Punktschätzungen, obwohl ein Hauptziel von Analysen post model selection sicherlich in einer geeigneten Erfassung der Variabilität dieser Schätzungen liegt.

Es bleibt festzuhalten, dass das Bewusstsein, geeignete Inferenzverfahren zur Berücksichtigung der Modellselektionsunsicherheit zu konstruieren, in den letzten Jahren deutlich geschärft wurde und viele Denkanstöße der Fachliteratur auch in den angewandten Wissenschaften immer häufiger verwendet werden. Ob dies auch im Kontext fehlender Daten gilt, werden zukünftige Arbeiten zeigen.

Anhang

A. Symbolverzeichnis

Im Folgenden sind sowohl die Verwendung von Symbolen und Abkürzungen als auch die wesentlichen Elemente der Notation dieser Arbeit erläutert. Die meisten Symbole besitzen nur eine Bedeutung. Einige wenige Symbole werden jedoch an verschiedenen Stellen mit verschiedener Bedeutung verwendet; dies ist gegebenenfalls aus dem Kontext zu erkennen. Die Auflistung ist nicht vollständig, umfasst jedoch alle elementaren und häufig verwendeten Konzepte der Arbeit.

A.1 Lateinische Symbole

Symbol	Bedeutung
b	Bootstrap-Stichprobe
B	Anzahl an Bootstrap-Stichproben
c	Konstante
d	Anzahl der Parameter, die auf jeden Fall in ein Endmodell M_{κ^*} aufgenommen werden sollten
$\mathcal{D}$	Datensatz in beliebiger Form
D	Datensatz in Form einer $n \times p + 1$ Matrix
$\mathcal{F}$	parametrisierte Familie von Wahrscheinlichkeitsverteilungen
f, g	beliebige Funktionen, häufig Dichten
F	Indikatormatrix, die angibt welche Elemente eines Datensatzes vollständig beobachtet worden sind und welche nicht
$\mathcal{H}$	Menge von Gewichten, die sich zu Eins aufsummieren
H	Hessematrix
i, j, l, m	Indexvariablen
I	Einheitsmatrix
J	Fisher-Informationsmatrix
k	Anzahl an Kandidatenmodellen
K	Anzahl zu schätzender Parameter

Symbol	Bedeutung
$\mathcal{L}$	Likelihoodfunktion
L	Verlustfunktion
M	Kandidatenmodell
$\mathcal{M}$	Menge aller Kandidatenmodelle
n	Stichprobengröße
p	Anzahl der Kovariablen in einer Regressionanalyse
P	Projektionsmatrix
q	Anzahl der Parameter, die potentiell in ein Endmodell M_{κ^*} integriert werden können
R	Risikofunktion
$\mathcal{R}$	Anzahl der betrachteten Simulationsläufe
w	Gewichtsvektor
X	Designmatrix bzw. Kovariable
x	Beobachtung eines Datensatzes D
$\mathbf{x}$	Zeilenvektor der Kovariablen in einer Regressionanalyse
y	Responsevektor
$\mathcal{Z}$	Zielfunktion

A.2 Griechische Symbole

Symbol	Bedeutung
α	Parametervektor, der zu den Variablen gehört, die auf alle Fälle in eine Endmodell M_{κ^*} aufgenommen werden, $\alpha \subset \beta$
β	Parametervektor in Regressionsmodellen, $\beta = (\alpha', \gamma')'$
γ	Parametervektor, der zu den Variablen gehört, die potentiell in ein Endmodell M_{κ^*} aufgenommen werden, $\gamma \subset \beta$
Γ	Kriterium oder Verfahren zur Bestimmung eines „besten“ Modells $M_{\kappa^*} \in \mathcal{M}$
δ	Indikatorvariable
ϵ	stochastischer Fehlerterm
η	Linkfunktion in generalisierten Regressionsmodellen
θ	Parametervektor der parametrisierten Dichte $f(y;\theta)$
ϑ	kanonischer Parametervektor im generalisierten linearen Regressionsmodell

Symbol	Bedeutung
κ	Indexvariable
λ	Indexvariable
μ	Erwartungswert einer Zufallsvariablen
ν	Pönalisierungsgewicht
ξ	Parametervektor zur Charakterisierung eines Fehlendmechanismus
π	Funktion zur Beschreibung einer Wahrscheinlichkeit
σ	Varianz einer Zufallsvariablen
ϕ	Dispersionsparameter in generalisierten linearen Regressionsmodellen
φ	Funktion zur Bestimmung eines M-Schätzers; enthält als Spezialfall die Likelihoodfunktion
χ	Menge aller Kovariablen in einer Regressionsanalyse
ψ	Die erste Ableitung von φ, enthält als Spezialfall die Score-Funktion
Θ	Parameterraum

A.3 Notation

Notation	Bedeutung
D_*	Datensatz, der auch fehlende Werte enthält
D_*^c	Datensatz, der nur die vollständigen Fälle enthält
D^{obs}	vollständig beobachtete Werte eines Datensatzes D
D^{mis}	fehlende Werte eines Datensatzes D
D^{imp}	Aufgefüllter Datensatz nach einer Imputation
$f(y;\theta)$	parametrisierte Wahrscheinlichkeitsverteilung
$f(y\|\theta)$	Likelihoodfunktion; entspricht $\mathcal{L}(\theta\|y)$
M_κ	Ein Modell, das Element der Menge $\mathcal{M}$ ist
M_{κ^*}	Ein Modell, das Element der Menge $\mathcal{M}$ ist und anhand eines Verfahrens oder Kriteriums gewählt wird
M_κ^*	Das wahre, datengenerierende Modell
M_κ^L	Ein Modell, das Element der Menge $\mathcal{M}$ ist und eine gegebene Verlustfunktion $L(\cdot)$ über alle $M_\kappa \in \mathcal{M}$ minimiert
$p(D^{\text{mis}}\|D^{\text{obs}})$	prädiktive a-posteriori-Verteilung der fehlenden Daten gegeben die beobachteten Daten

Notation	Bedeutung
$p(M_\kappa)$	a-priori-Wahrscheinlichkeit für Modell M_κ
$p(M_\kappa \vert y)$	posteriori-Wahrscheinlichkeit für M_κ
$p(\theta \vert D^{\text{obs}})$	posteriori-Verteilung von θ gegeben die beobachteten Daten
$p(\theta_\kappa \vert M_\kappa)$	a-priori-Wahrscheinlichkeit für θ_κ im Modell M_κ
$p(\theta_\kappa \vert M_\kappa, y)$	posteriori-Verteilung von θ_κ
$\hat{\theta}$	ML- bzw. KQ-Schätzer für θ
$\hat{\theta}^M$	beliebiger Schätzer nach multipler Imputation
$\hat{\theta}^{(m)}$	beliebiger Schätzer des m-ten imputierten Datensatzes
θ^*	Maximum A Posteriori-Schätzung für θ
$\tilde{\theta}$	M-Schätzer für θ
$\hat{\theta}_W$	gewichteter ML- bzw. KQ-Schätzer für θ
$\hat{\bar{\theta}}$	gemittelter Schätzer für θ

A.4 Abkürzungen

Abkürzung	Bedeutung
Abb.	Abbildung
AIC	Akaikes Informationskriterium
AIC_c	Akaikes Informationskriterium; korrigiert für kleine Stichproben
AIC_W	Akaikes Informationskriterium; adjustiert für fehlende Beobachtungen
BMA	Bayesian Model Averaging; Bayesianische Modellmittelung
BMS	Bayesian Model Selection; Bayesianische Modellselektion
bzw.	beziehungsweise
C_p	Kriterium von Colin Mallows
CC	Complete Cases; vollständige Beobachtungen
CV	Cross Validation Criterion; Kreuzvaliderungskriterium
det	Determinante
df	Freiheitsgrade

Abkürzung	Bedeutung
dim	die Dimension eines Modells; entspricht meist der Anzahl der Elemente von θ
DMD	Duchenne Muscular Dystrophy
EMP	Entropy Maximization Principle
EPE	Expected Prediction Error
FMA	Frequentist Model Averaging; Frequentistische Modellmittelung
FMS	Frequentist Model Selection; Frequentistische Modellselektion
ff.	und folgende Seiten
f.s.	fast sicher
GAMRI	Generalized Additive Model based Recursive Imputation; verallgemeinerter Imputationsalgorithmus
GLMRI	Generalized Linear Model based Recursive Imputation; verallgemeinerter Imputationsalgorithmus
GCV	Generalized Cross Validation Criterion; Generalisiertes Kreuzvalidierungskriterium
Hrsg.	Herausgeber
i.d.R.	in der Regel
kNN	k-Nächste-Nachbarn-Imputationsmethode
KL	Kullback-Leibler-Distanz
LOCF	Last Observation Carried Forward
MAP	Maximum-A-Posteriori
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
MI	Multiple Imputation
ML	Maximum Likelihood
MLQ	Multifactor Leadership Questionnaire
MMA	Mallows Model Averaging
MNAR	Missing Not At Random

Abkürzung	Bedeutung
MSE	Mean Squared Error
MSPE	Mean Squared Prediction Error
o.B.d.A.	ohne Beschränkung der Allgemeinheit
PMSE	Post Model Selection Estimator
rg	Rang einer Matrix
S.	Seite
SBC	Schwarzsches Bayes-Kriterium
se	standard error
sp	Spur einer Matrix
SSE	Sum of Squares Error; Residuenquadratsummen im linearen Regressionsmodell
Tab.	Tabelle
TIC	Takeuchis Informationskriterium
Var	Varianz

B. Detaillierte Simulationsergebnisse

B.1 Lineare Regression

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.2304	–	0.1337	7.4189
2) CC	1.5224	1.3751	1.4253	8.9009
3) GAMRI	0.8403	0.5893	0.5619	7.9840
GLMRI	1.0047	0.7439	0.7006	8.1222
kNN	0.6021	0.3981	0.4013	7.7832
Amelia	0.4644	0.2614	0.2758	7.6916
4) AIC_W	1.0714	0.8290	0.7135	8.0802
FMA-Hansen-Schätzer				
1) Original	0.2555	–	0.1418	7.4234
2) CC	1.9288	1.6008	1.6159	9.0921
3) GAMRI	0.9599	0.6477	0.6357	8.0616
GLMRI	1.1453	0.8173	0.7937	8.2210
kNN	0.6755	0.4153	0.4120	7.8004
Amelia	0.5694	0.2851	0.3116	7.7385
FMS-AIC-Schätzer				
1) Original	0.2610	–	0.1628	7.4412
2) CC	1.6116	1.4881	1.5277	9.0011
3) GAMRI	0.8949	0.6668	0.6077	8.0357
GLMRI	1.0605	0.8254	0.7444	8.1724
kNN	0.6434	0.4350	0.4354	7.8144
Amelia	0.5020	0.3057	0.3080	7.7185
4) AIC_W	1.1123	0.8766	0.7468	8.1042

Tab. B.1: Resultate im Grundszenario

	se(β_0)		se(β_1)		se(β_2)		se(β_3)		se(β_4)		se(β_5)	
FMA-Akaike-Schätzer												
1) Original	0.29	(0.29)	0.14	(0.14)	0.14	(0.16)	0.07	(0.08)	0.29	(0.29)	0.07	(0.08)
2) CC	0.41	(0.41)	0.21	(0.22)	0.20	(0.22)	0.09	(0.09)	0.37	(0.38)	0.08	(0.09)
3) GAMRI	0.28	(0.42)	0.15	(0.21)	0.13	(0.21)	0.10	(0.16)	0.29	(0.42)	0.09	(0.18)
GLMRI	0.28	(0.39)	0.15	(0.20)	0.13	(0.18)	0.10	(0.16)	0.28	(0.41)	0.09	(0.18)
kNN	0.30	(0.41)	0.17	(0.20)	0.14	(0.18)	0.09	(0.12)	0.30	(0.41)	0.09	(0.12)
Amelia	0.30	(0.41)	0.15	(0.19)	0.14	(0.20)	0.09	(0.12)	0.30	(0.38)	0.09	(0.15)
4) AIC_W	0.45	(0.66)	0.22	(0.28)	0.21	(0.37)	0.14	(0.23)	0.43	(0.56)	0.13	(0.20)
FMA-Hansen-Schätzer												
1) Original	0.34	(0.29)	0.16	(0.14)	0.15	(0.14)	0.14	(0.13)	0.38	(0.31)	0.04	(0.08)
2) CC	0.51	(0.44)	0.24	(0.23)	0.23	(0.22)	0.18	(0.16)	0.54	(0.43)	0.04	(0.10)
3) GAMRI	0.31	(0.43)	0.17	(0.21)	0.14	(0.21)	0.14	(0.17)	0.39	(0.44)	0.07	(0.18)
GLMRI	0.30	(0.41)	0.17	(0.20)	0.14	(0.19)	0.14	(0.17)	0.39	(0.44)	0.08	(0.17)
kNN	0.35	(0.42)	0.19	(0.21)	0.16	(0.17)	0.15	(0.16)	0.41	(0.43)	0.06	(0.12)
Amelia	0.36	(0.42)	0.17	(0.19)	0.16	(0.19)	0.15	(0.17)	0.41	(0.40)	0.07	(0.14)
FMS-AIC-Schätzer												
1) Original	0.25	(0.32)	0.14	(0.15)	0.11	(0.17)	0.02	(0.10)	0.28	(0.29)	0.02	(0.09)
2) CC	0.31	(0.46)	0.20	(0.22)	0.10	(0.25)	0.02	(0.12)	0.37	(0.38)	0.02	(0.11)
3) GAMRI	0.25	(0.44)	0.15	(0.21)	0.10	(0.21)	0.06	(0.18)	0.28	(0.42)	0.05	(0.20)
GLMRI	0.24	(0.41)	0.15	(0.20)	0.10	(0.19)	0.05	(0.18)	0.28	(0.41)	0.05	(0.20)
kNN	0.26	(0.44)	0.16	(0.21)	0.12	(0.20)	0.03	(0.14)	0.30	(0.41)	0.03	(0.14)
Amelia	0.26	(0.43)	0.15	(0.19)	0.11	(0.21)	0.03	(0.15)	0.30	(0.39)	0.04	(0.17)
4) AIC_W	0.41	(0.67)	0.21	(0.29)	0.19	(0.37)	0.08	(0.25)	0.43	(0.56)	0.07	(0.23)

Tab. B.2: Mittlere geschätzte Standardfehler unter Verwendung von (6.8) im Grundszenario; in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.2165	–	0.1226	7.4317
2) CC	0.3250	0.1175	0.1854	7.6072
3) GAMRI	0.3339	0.1011	0.1824	7.6056
GLMRI	0.3386	0.1074	0.1865	7.6101
kNN	0.2938	0.0780	0.1616	7.5873
Amelia	0.2856	0.0914	0.1600	7.5891
4) AIC_W	0.3297	0.1186	0.1904	7.6108
FMA-Hansen-Schätzer				
1) Original	0.2372	–	0.1309	7.4374
2) CC	0.3819	0.1350	0.2092	7.6367
3) GAMRI	0.3721	0.1138	0.2101	7.6383
GLMRI	0.3825	0.1229	0.2170	7.6453
kNN	0.3427	0.0886	0.1812	7.6135
Amelia	0.3410	0.1010	0.1749	7.6100
FMS-AIC-Schätzer				
1) Original	0.2522	–	0.1518	7.4594
2) CC	0.3859	0.1741	0.2365	7.6536
3) GAMRI	0.3825	0.1464	0.2212	7.6444
GLMRI	0.3876	0.1551	0.2234	7.6484
kNN	0.3403	0.1134	0.2000	7.6243
Amelia	0.3273	0.1236	0.1974	7.6257
4) AIC_W	0.3722	0.1675	0.2244	7.6500

Tab. B.3: Resultate aus Experiment 2

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.2365	–	0.1426	7.4286
2) CC	0.9586	0.8048	0.8716	8.2773
3) GAMRI	0.3721	0.1350	0.1864	7.6084
GLMRI	0.3803	0.1450	0.1861	7.6034
kNN	0.3363	0.1008	0.1788	7.5945
Amelia	0.3552	0.1390	0.1864	7.5910
4) AIC_W	0.4644	0.2350	0.2892	7.6953
FMA-Hansen-Schätzer				
1) Original	0.2593	–	0.1498	7.4259
2) CC	1.1303	0.8570	0.9527	8.3552
3) GAMRI	0.3545	0.1405	0.1927	7.6159
GLMRI	0.3552	0.1506	0.1909	7.6090
kNN	0.3816	0.1122	0.1913	7.6083
Amelia	0.4458	0.1553	0.2054	7.6115
FMS-AIC-Schätzer				
1) Original	0.2726	–	0.1747	7.4573
2) CC	1.0079	0.8550	0.9108	8.3271
3) GAMRI	0.4198	0.1695	0.2259	7.6440
GLMRI	0.4277	0.1771	0.2265	7.6429
kNN	0.3724	0.1311	0.2103	7.6272
Amelia	0.3945	0.1655	0.2194	7.6184
4) AIC_W	0.5038	0.2796	0.3214	7.7335

Tab. B.4: Resultate aus Experiment 3

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.3410	–	0.1906	11.4441
2) CC	2.4211	2.3074	2.2490	13.6521
3) GAMRI	1.3835	1.0135	0.9093	12.3127
GLMRI	1.7384	1.3243	1.1941	12.5775
kNN	0.9221	0.6256	0.6117	12.0563
Amelia	0.6656	0.3555	0.3767	11.7790
4) AIC_W	1.9596	1.6332	1.3324	12.6928
FMA-Hansen-Schätzer				
1) Original	0.3907	–	0.2083	11.4474
2) CC	3.2278	2.7977	2.6401	14.0419
3) GAMRI	1.5759	1.1206	1.0152	12.4152
GLMRI	1.9657	1.4498	1.3188	12.6981
kNN	1.0483	0.6622	0.6274	12.0790
Amelia	0.8652	0.4195	0.4460	11.8527
FMS-AIC-Schätzer				
1) Original	0.3962	–	0.2373	11.4828
2) CC	2.5391	2.4817	2.3800	13.7806
3) GAMRI	1.4729	1.1453	0.9789	12.3768
GLMRI	1.8373	1.4745	1.2723	12.6558
kNN	0.9843	0.6989	0.6581	12.1069
Amelia	0.7308	0.4352	0.4362	11.8338
4) AIC_W	2.0333	1.7620	1.3889	12.7640

Tab. B.5: Resultate aus Experiment 4

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.2130	–	0.0992	7.4151
2) CC	1.4696	1.3411	1.4782	8.9157
3) GAMRI	1.1593	0.8950	0.6899	8.1260
GLMRI	2.3707	2.0398	1.5713	9.0299
kNN	0.6271	0.4373	0.3252	7.7598
Amelia	0.5398	0.3440	0.2551	7.6625
4) AIC_W	1.0194	0.8101	0.6118	8.0334
FMA-Hansen-Schätzer				
1) Original	0.2405	–	0.1049	7.4194
2) CC	1.7662	1.5031	1.5949	9.0362
3) GAMRI	1.3253	1.0382	0.7819	8.2224
GLMRI	2.5298	2.1851	1.6492	9.1118
kNN	0.7387	0.4869	0.3639	7.7995
Amelia	0.6747	0.3846	0.2914	7.6983
FMS-AIC-Schätzer				
1) Original	0.2470	–	0.1215	7.4301
2) CC	1.5545	1.4493	1.5596	9.0024
3) GAMRI	1.2445	1.0080	0.7431	8.1801
GLMRI	2.4698	2.1714	1.6414	9.0963
kNN	0.6709	0.5084	0.3520	7.7884
Amelia	0.5904	0.4138	0.2860	7.6992
4) AIC_W	1.0675	0.8753	0.6412	8.0684

Tab. B.6: Resultate aus Experiment 5

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.4179	–	0.1459	7.3305
2) CC	1.6352	1.3177	0.8607	8.2028
3) GAMRI	0.9184	0.4838	0.3568	7.6711
GLMRI	0.9016	0.4936	0.3830	7.6998
kNN	0.7689	0.3595	0.2743	7.6126
Amelia	0.7776	0.4488	0.2431	7.6009
4) AIC_W	1.3046	0.9840	0.5906	7.9238
FMA-Hansen-Schätzer				
1) Original	0.5654	–	0.1679	7.3367
2) CC	2.5182	1.5750	1.0244	8.3719
3) GAMRI	1.2465	0.5211	0.4487	7.7709
GLMRI	1.2451	0.5316	0.4842	7.8095
kNN	1.0006	0.3713	0.2986	7.6416
Amelia	1.2457	0.4912	0.2983	7.6552
FMS-AIC-Schätzer				
1) Original	0.4153	–	0.1594	7.3407
2) CC	1.7204	1.4458	0.9123	8.2564
3) GAMRI	0.9435	0.5599	0.3907	7.6995
GLMRI	0.9389	0.5836	0.4215	7.7370
kNN	0.7895	0.4075	0.3020	7.6379
Amelia	0.7827	0.4822	0.2699	7.6244
4) AIC_W	1.3424	1.0340	0.6232	7.9543

Tab. B.7: Resultate aus Experiment 6

	L_1	L_2	L_3	L_4
FMA-Akaike-Schätzer				
1) Original	0.5003	–	0.1050	7.3305
2) CC	3.4687	3.2761	14.1101	21.3577
3) GAMRI	4.9016	4.4069	2.3102	9.7724
GLMRI	6.3662	5.9210	6.3293	13.7643
kNN	1.3049	0.8027	0.8719	8.3156
Amelia	7.9882	8.0278	9.1899	16.6371
4) AIC_W	5.2356	4.7386	25.4760	32.7809
FMA-Hansen-Schätzer				
1) Original	0.6475	–	0.1732	7.3628
2) CC	4.3269	3.7676	10.7959	18.0960
3) GAMRI	5.5303	4.5376	3.0721	10.5303
GLMRI	6.9261	5.8800	7.2165	14.6520
kNN	1.7729	1.0003	1.3188	8.7743
Amelia	8.3941	7.6953	9.0832	16.5423
FMS-AIC-Schätzer				
1) Original	0.5546	–	0.1318	7.3540
2) CC	3.8104	3.6662	13.9921	21.2542
3) GAMRI	5.0376	4.5441	2.6047	10.0683
GLMRI	6.4684	6.0407	6.1844	13.6125
kNN	1.4271	0.9222	0.9227	8.3721
Amelia	8.0571	8.1495	9.5958	17.0411
4) AIC_W	5.4094	4.9167	25.5961	32.8970

Tab. B.8: Resultate aus Experiment 7

B.2 Logistische Regression

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	0.3606	–	0.2060	0.2456	0.0052	0.0125	0.8609
2) CC	0.6755	0.3055	0.4520	0.2568	0.0091	0.0236	0.8549
3) GAMRI	0.6158	0.1926	0.3141	0.2570	0.0075	0.0190	0.8559
GLMRI	0.6423	0.2266	0.3125	0.2602	0.0077	0.0193	0.8560
kNN	0.5960	0.1799	0.3092	0.2607	0.0077	0.0192	0.8558
Amelia	0.5095	0.3763	0.2249	0.2459	0.0058	0.0141	0.8553
4) AIC_W	0.8840	0.4813	0.6183	0.2600	0.0120	0.0324	0.8527
FMS-AIC-Schätzer							
1) Original	0.3790	–	0.2177	0.2468	0.0055	0.0135	0.8606
2) CC	0.7248	0.3529	0.4855	0.2584	0.0100	0.0260	0.8533
3) GAMRI	0.6395	0.2124	0.3293	0.2582	0.0079	0.0200	0.8553
GLMRI	0.6619	0.2450	0.3228	0.2612	0.0080	0.0200	0.8557
kNN	0.6203	0.1999	0.3251	0.2605	0.0080	0.0203	0.8555
Amelia	0.5333	0.4087	0.2412	0.2468	0.0062	0.0153	0.8541
4) AIC_W	0.9072	0.5128	0.6348	0.2610	0.0124	0.0335	0.8523

Tab. B.9: Resultate im Grundszenario

	se(β_0)		se(β_1)		se(β_2)		se(β_3)		se(β_4)		se(β_5)	
FMA-Akaike-Schätzer												
1) Original	0.35	(0.37)	0.25	(0.26)	0.11	(0.12)	0.06	(0.07)	0.06	(0.07)	0.34	(0.35)
2) CC	0.45	(0.49)	0.37	(0.37)	0.16	(0.16)	0.10	(0.12)	0.08	(0.09)	0.44	(0.46)
3) GAMRI	0.39	(0.44)	0.26	(0.29)	0.14	(0.18)	0.07	(0.08)	0.08	(0.10)	0.37	(0.41)
GLMRI	0.39	(0.43)	0.26	(0.28)	0.14	(0.18)	0.07	(0.08)	0.08	(0.10)	0.38	(0.40)
kNN	0.39	(0.44)	0.26	(0.28)	0.13	(0.17)	0.06	(0.08)	0.07	(0.09)	0.37	(0.41)
Amelia	0.32	(0.35)	0.25	(0.26)	0.11	(0.14)	0.07	(0.07)	0.07	(0.09)	0.31	(0.36)
4) AIC_W	0.47	(0.55)	0.34	(0.42)	0.15	(0.18)	0.12	(0.24)	0.10	(0.15)	0.46	(0.50)
FMS-AIC-Schätzer												
1) Original	0.34	(0.38)	0.25	(0.26)	0.10	(0.13)	0.02	(0.09)	0.02	(0.09)	0.33	(0.35)
2) CC	0.43	(0.51)	0.37	(0.38)	0.11	(0.19)	0.03	(0.15)	0.02	(0.12)	0.43	(0.47)
3) GAMRI	0.37	(0.45)	0.26	(0.29)	0.12	(0.19)	0.02	(0.10)	0.03	(0.12)	0.37	(0.42)
GLMRI	0.38	(0.43)	0.26	(0.28)	0.12	(0.18)	0.02	(0.10)	0.03	(0.12)	0.37	(0.40)
kNN	0.38	(0.45)	0.26	(0.28)	0.11	(0.18)	0.02	(0.10)	0.03	(0.12)	0.37	(0.41)
Amelia	0.31	(0.36)	0.25	(0.26)	0.08	(0.16)	0.02	(0.10)	0.02	(0.11)	0.30	(0.36)
4) AIC_W	0.47	(0.55)	0.34	(0.42)	0.15	(0.18)	0.08	(0.26)	0.05	(0.16)	0.45	(0.50)

Tab. B.10: Mittlere geschätzte Standardfehler unter Verwendung von (6.8) im Grundszenario; in Klammern die zugehörigen empirischen Standardfehler nach (6.7)

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	0.6344	–	0.3507	0.2398	0.0078	0.0197	0.8652
2) CC	2.1565	1.2095	1.0098	0.2522	0.0122	0.0420	0.8579
3) GAMRI	1.1596	0.3683	0.6895	0.2492	0.0105	0.0287	0.8587
GLMRI	1.2197	0.4150	0.7197	0.2485	0.0106	0.0292	0.8585
kNN	0.9778	0.2590	0.5578	0.2498	0.0104	0.0274	0.8585
Amelia	0.7014	0.4839	0.3856	0.2462	0.0081	0.0203	0.8569
4) AIC_W	2.2087	1.2498	1.0240	0.2504	0.0122	0.0424	0.8574
FMS-AIC-Schätzer							
1) Original	0.6736	–	0.3744	0.2390	0.0084	0.0213	0.8641
2) CC	2.2209	1.2770	1.0664	0.2528	0.0134	0.0455	0.8549
3) GAMRI	1.2126	0.4118	0.7232	0.2505	0.0112	0.0309	0.8576
GLMRI	1.2797	0.4708	0.7515	0.2504	0.0111	0.0308	0.8576
kNN	1.0268	0.2980	0.5886	0.2509	0.0110	0.0292	0.8575
Amelia	0.7519	0.5377	0.4218	0.2484	0.0091	0.0228	0.8549
4) AIC_W	2.2457	1.2951	1.0460	0.2506	0.0127	0.0438	0.8571

Tab. B.11: Resultate aus Experiment 2

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	0.5224	–	0.3053	0.2406	0.0067	0.0169	0.8631
2) CC	2.2050	1.5043	0.9980	0.2528	0.0109	0.0397	0.8558
3) GAMRI	1.1177	0.4747	0.5438	0.2520	0.0103	0.0296	0.8564
GLMRI	1.1282	0.5108	0.5394	0.2531	0.0104	0.0297	0.8562
kNN	1.0595	0.4568	0.5242	0.2535	0.0105	0.0297	0.8557
Amelia	0.7183	0.4756	0.3511	0.2472	0.0083	0.0210	0.8550
4) AIC_W	2.3387	1.6099	1.0788	0.2528	0.0119	0.0431	0.8542
FMS-AIC-Schätzer							
1) Original	0.5648	–	0.3305	0.2406	0.0074	0.0187	0.8616
2) CC	2.2508	1.5581	1.0484	0.2553	0.0122	0.0434	0.8531
3) GAMRI	1.1662	0.5149	0.5725	0.2537	0.0110	0.0316	0.8551
GLMRI	1.1732	0.5470	0.5657	0.2542	0.0110	0.0316	0.8551
kNN	1.0934	0.4863	0.5536	0.2562	0.0113	0.0320	0.8540
Amelia	0.7547	0.5244	0.3748	0.2485	0.0089	0.0225	0.8536
4) AIC_W	2.3924	1.6776	1.1096	0.2532	0.0123	0.0448	0.8538

Tab. B.12: Resultate aus Experiment 3

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	0.6753	–	0.3771	0.2549	0.0082	0.0202	0.8465
2) CC	1.0680	0.4664	0.6229	0.2652	0.0125	0.0320	0.8382
3) GAMRI	0.8045	0.2704	0.4636	0.2678	0.0106	0.0269	0.8385
GLMRI	0.7678	0.3125	0.4408	0.2652	0.0099	0.0251	0.8383
kNN	0.8509	0.2649	0.4740	0.2674	0.0108	0.0271	0.8382
Amelia	1.1618	0.3285	0.5025	0.2636	0.0118	0.0294	0.8371
4) AIC_W	1.3311	0.7410	0.8146	0.2707	0.0165	0.0442	0.8347
FMS-AIC-Schätzer							
1) Original	0.7112	–	0.4000	0.2558	0.0088	0.0218	0.8444
2) CC	1.1550	0.5585	0.6767	0.2655	0.0134	0.0353	0.8364
3) GAMRI	0.8578	0.3183	0.4997	0.2683	0.0115	0.0294	0.8373
GLMRI	0.8151	0.3542	0.4743	0.2680	0.0108	0.0275	0.8368
kNN	0.8938	0.3040	0.5020	0.2681	0.0114	0.0291	0.8374
Amelia	1.1960	0.3736	0.5254	0.2654	0.0124	0.0310	0.8352
4) AIC_W	1.3714	0.7952	0.8397	0.2700	0.0170	0.0461	0.8341

Tab. B.13: Resultate aus Experiment 4

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	0.3631	–	0.3216	0.2377	0.0060	0.0166	0.8298
2) CC	0.6691	0.2909	0.6240	0.2490	0.0100	0.0290	0.8232
3) GAMRI	0.5602	0.1816	0.4343	0.2496	0.0085	0.0238	0.8238
GLMRI	0.5470	0.2192	0.4023	0.2517	0.0081	0.0222	0.8238
kNN	0.5143	0.1637	0.4110	0.2513	0.0084	0.0229	0.8240
Amelia	0.8507	0.7705	0.5581	0.2815	0.0166	0.0392	0.8236
4) AIC_W	0.8823	0.4949	0.8355	0.2564	0.0135	0.0398	0.8207
FMS-AIC-Schätzer							
1) Original	0.3643	–	0.3206	0.2389	0.0061	0.0166	0.8295
2) CC	0.7171	0.3356	0.6682	0.2517	0.0111	0.0324	0.8217
3) GAMRI	0.5734	0.1946	0.4426	0.2519	0.0088	0.0245	0.8232
GLMRI	0.5553	0.2236	0.4076	0.2516	0.0083	0.0226	0.8235
kNN	0.5318	0.1746	0.4271	0.2515	0.0088	0.0241	0.8231
Amelia	0.8884	0.7894	0.5915	0.2858	0.0179	0.0424	0.8224
4) AIC_W	0.8979	0.5235	0.8481	0.2565	0.0137	0.0406	0.8201

Tab. B.14: Resultate aus Experiment 5

	L_1	L_2	L_3	L_5	L_6	L_7	L_8
FMA-Akaike-Schätzer							
1) Original	3.5851	–	3.8569	0.2499	0.0231	0.0669	0.8195
2) CC	5.2233	1.9233	4.5507	0.2683	0.0317	0.0934	0.8154
3) GAMRI	4.1877	1.9175	3.9425	0.2519	0.0224	0.0677	0.8121
GLMRI	3.6152	1.1149	3.1731	0.2475	0.0214	0.0641	0.8128
kNN	4.7901	1.8613	3.8357	0.2549	0.0251	0.0780	0.8123
Amelia	3.4105	2.3410	2.8265	0.2530	0.0223	0.0623	0.8142
4) AIC_W	6.3385	3.1122	6.6343	0.2813	0.0374	0.1094	0.8120
FMS-AIC-Schätzer							
1) Original	4.0024	–	4.1934	0.2539	0.0247	0.0727	0.8191
2) CC	6.7579	3.0584	5.4951	0.2809	0.0388	0.1173	0.8146
3) GAMRI	5.4528	3.1749	5.1074	0.2643	0.0289	0.0880	0.8114
GLMRI	4.6244	1.8854	3.4729	0.2583	0.0266	0.0802	0.8124
kNN	6.0965	3.0447	4.5345	0.2662	0.0313	0.0977	0.8123
Amelia	3.8880	2.9796	3.3857	0.2567	0.0249	0.0704	0.8131
4) AIC_W	6.6409	3.4768	6.6817	0.2832	0.0388	0.1144	0.8112

Tab. B.15: Resultate aus Experiment 6

Ergänzungen zu Abschnitt 6.3

	Org	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	0.43	1.54	0.93	0.92	0.74	0.73	0.64	1.28
FMA-Hansen	0.61	2.48	1.29	1.28	0.99	1.21	1.12	–
FMS-AIC	0.46	1.68	0.99	0.96	0.78	0.76	0.65	1.34

Tab. B.16: Übersicht über den Verlust L_1 für das Grundszenario

	Org	CC	GAMRI	GLMRI	kNN	Amelia	MI	AIC_W
FMA-Akaike	0.39	2.68	0.57	0.57	0.55	0.51	0.49	1.25
FMA-Hansen	0.57	3.68	0.77	0.76	0.74	0.71	0.69	–
FMS-AIC	0.41	2.74	0.60	0.60	0.58	0.54	0.50	1.31

Tab. B.17: Übersicht über den Verlust L_1 für Experiment 2

C. Weitere Analysen

	Ladungsmatrix Γ					Ψ
100 m	0.80	-0.30	0.00	0.00	0.00	0.27
Weitsprung	-0.82	0.05	0.00	0.00	0.00	0.33
Kugelstoßen	-0.15	0.99	0.00	0.00	0.00	0.00
Hochsprung	-0.27	0.61	0.00	0.00	0.00	0.55
400 m	0.77	-0.17	0.00	0.00	0.00	0.38
110 m Hürden	0.65	-0.24	0.00	0.00	0.00	0.52
Diskuswerfen	-0.20	0.64	0.00	0.00	0.00	0.55
Stabhochsprung	-0.26	0.00	0.00	0.00	0.00	0.93
Speerwurf	-0.01	0.43	0.00	0.00	0.00	0.81
1500 m	0.21	0.22	0.00	0.00	0.00	0.91

Tab. C.1: Ladungsmatrix (GLMRI, 2 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.83	-0.28	-0.07	0.00	0.00	0.23
Weitsprung	-0.79	0.06	-0.03	0.00	0.00	0.37
Kugelstoßen	-0.17	0.94	0.08	0.00	0.00	0.08
Hochsprung	-0.24	0.64	-0.01	0.00	0.00	0.53
400 m	0.76	-0.19	0.39	0.00	0.00	0.23
110 m Hürden	0.64	-0.26	0.00	0.00	0.00	0.52
Diskuswerfen	-0.20	0.66	0.23	0.00	0.00	0.47
Stabhochsprung	-0.30	-0.06	0.23	0.00	0.00	0.85
Speerwurf	0.05	0.48	-0.28	0.00	0.00	0.69
1500 m	0.11	0.13	0.98	0.00	0.00	0.00

Tab. C.2: Ladungsmatrix (GLMRI, 3 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.79	-0.31	0.00	0.00	0.00	0.28
Weitsprung	-0.79	0.10	0.00	0.00	0.00	0.37
Kugelstoßen	-0.16	0.94	0.00	0.00	0.00	0.09
Hochsprung	-0.24	0.64	0.00	0.00	0.00	0.53
400 m	0.79	-0.15	0.00	0.00	0.00	0.36
110 m Hürden	0.63	-0.29	0.00	0.00	0.00	0.51
Diskuswerfen	-0.16	0.72	0.00	0.00	0.00	0.46
Stabhochsprung	-0.27	-0.03	0.00	0.00	0.00	0.93
Speerwurf	-0.01	0.41	0.00	0.00	0.00	0.83
1500 m	0.25	0.26	0.00	0.00	0.00	0.87

Tab. C.3: Ladungsmatrix (kNN, 2 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.83	-0.27	-0.07	0.00	0.00	0.23
Weitsprung	-0.78	0.10	-0.03	0.00	0.00	0.39
Kugelstoßen	-0.19	0.91	0.08	0.00	0.00	0.14
Hochsprung	-0.23	0.66	-0.03	0.00	0.00	0.52
400 m	0.77	-0.17	0.38	0.00	0.00	0.23
110 m Hürden	0.63	-0.29	0.00	0.00	0.00	0.52
Diskuswerfen	-0.18	0.72	0.20	0.00	0.00	0.41
Stabhochsprung	-0.32	-0.10	0.24	0.00	0.00	0.83
Speerwurf	0.04	0.46	-0.29	0.00	0.00	0.70
1500 m	0.12	0.16	0.98	0.00	0.00	0.00

Tab. C.4: Ladungsmatrix (kNN, 3 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.75	-0.32	0.00	0.00	0.00	0.33
Weitsprung	-0.82	0.20	0.00	0.00	0.00	0.29
Kugelstoßen	-0.08	0.90	0.00	0.00	0.00	0.18
Hochsprung	-0.19	0.66	0.00	0.00	0.00	0.53
400 m	0.81	-0.12	0.00	0.00	0.00	0.32
110 m Hürden	0.61	-0.24	0.00	0.00	0.00	0.57
Diskuswerfen	-0.06	0.74	0.00	0.00	0.00	0.44
Stabhochsprung	-0.27	-0.05	0.00	0.00	0.00	0.93
Speerwurf	0.01	0.37	0.00	0.00	0.00	0.86
1500 m	0.36	0.19	0.00	0.00	0.00	0.84

Tab. C.5: Ladungsmatrix (Complete Cases, 2 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.81	-0.24	-0.08	0.00	0.00	0.28
Weitsprung	-0.83	0.13	0.01	0.00	0.00	0.29
Kugelstoßen	-0.16	0.89	0.02	0.00	0.00	0.19
Hochsprung	-0.21	0.65	-0.06	0.00	0.00	0.53
400 m	0.80	-0.08	0.40	0.00	0.00	0.19
110 m Hürden	0.61	-0.19	0.08	0.00	0.00	0.58
Diskuswerfen	-0.13	0.74	0.13	0.00	0.00	0.42
Stabhochsprung	-0.32	-0.10	0.28	0.00	0.00	0.81
Speerwurf	0.06	0.42	-0.34	0.00	0.00	0.70
1500 m	0.22	0.16	0.96	0.00	0.00	0.00

Tab. C.6: Ladungsmatrix (Complete Cases, 3 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.77	-0.23	-0.08	-0.19	0.24	0.26
Weitsprung	-0.61	0.11	-0.05	0.76	-0.18	0.00
Kugelstoßen	-0.28	0.93	0.20	-0.09	-0.02	0.00
Hochsprung	-0.24	0.62	0.07	0.16	0.05	0.53
400 m	0.83	-0.13	0.35	-0.09	0.02	0.16
110 m Hürden	0.65	-0.19	-0.04	-0.12	0.01	0.52
Diskuswerfen	-0.32	0.59	0.40	0.20	0.59	0.00
Stabhochsprung	-0.19	-0.05	0.15	0.10	-0.47	0.71
Speerwurf	0.00	0.49	-0.21	0.01	0.08	0.71
1500 m	0.13	0.03	0.98	-0.02	-0.13	0.00

Tab. C.7: Ladungsmatrix (kNN, 5 Faktoren)

	Ladungsmatrix Γ					Ψ
100 m	0.76	-0.21	-0.07	0.13	0.11	0.35
Weitsprung	-0.92	0.14	0.01	-0.06	0.29	0.05
Kugelstoßen	-0.19	0.90	0.17	0.08	-0.34	0.00
Hochsprung	-0.26	0.64	0.05	0.11	0.09	0.50
400 m	0.84	0.01	0.41	-0.08	0.33	0.00
110 m Hürden	0.57	-0.16	0.08	-0.01	0.01	0.64
Diskuswerfen	-0.24	0.57	0.29	0.71	-0.03	0.04
Stabhochsprung	-0.28	-0.04	0.22	-0.41	0.00	0.70
Speerwurf	0.01	0.52	-0.25	0.06	0.12	0.65
1500 m	0.17	0.00	0.98	-0.03	0.00	0.00

Tab. C.8: Ladungsmatrix (Complete Cases, 5 Faktoren)

		100m	Weit-sprung	Kugel-stoßen	Hoch-sprung	400m	110m Hürden	Diskus-wurf	Stabhoch-sprung	Speer-wurf	1500m
Roman Sebrle	(CZE)	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01
Bryan Clay	(USA)	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00
Dmitriy Karpov	(KAZ)	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11
Dean Macey	(GBR)	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42
Chiel Warners	(NED)	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05
Attila Zsivoczky	(HUN)	10.91	7.14	15.31	2.12	49.40	14.95	45.62	4.70	63.45	269.54
Laurent Hernu	(FRA)	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.80	57.76	264.35
Erki Nool	(EST)	10.80	7.53	14.26	1.88	48.81	14.80	42.05	5.40	61.33	276.33
Claston Bernard	(JAM)	10.69	7.48	14.80	2.12	49.13	14.17	44.75	4.40	55.27	276.31
Roland Schwarzl	(AUT)	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.10	56.32	273.56
Aleksandr Pogorelov	(RUS)	10.95	7.31	15.10	2.06	50.79	14.21	44.60	5.00	53.45	287.63
Florian Schönbeck	(GER)	10.90	7.30	14.77	1.88	50.30	14.34	44.41	5.00	60.89	278.82
Romain Barras	(FRA)	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.60	64.55	267.09
Marice Smith	(JAM)	10.85	6.81	15.24	1.91	49.27	14.01	49.02	4.20	61.52	272.74
Nikolay Averyanov	(RUS)	10.55	7.34	14.44	1.94	49.72	14.39	39.88	4.80	54.51	271.02
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tab. C.9: Resultate des Olympischen Zehnkampfs am 23.8./24.8.2004 in Athen (Teil I)

		100m	Weit-sprung	Kugel-stoßen	Hoch-sprung	400m	110m Hürden	Diskus-wurf	Stabhoch-sprung	Speer-wurf	1500m
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Jaako Ojaniemi	(FIN)	10.68	7.50	14.97	1.94	49.12	15.01	40.35	4.60	59.26	275.71
Vitaliy Smirnov	(UZB)	10.89	7.07	13.88	1.94	49.11	14.77	42.47	4.70	60.88	263.31
Haifeng Qi	(CHN)	11.06	7.34	13.55	1.97	49.65	14.78	45.13	4.50	60.79	272.63
Stefan Drews	(GER)	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5.00	51.53	274.21
Aleksandr Parkhomenko	(BLR)	11.14	6.61	15.69	2.03	51.04	14.88	41.90	4.80	65.82	277.94
Paul Terek	(USA)	10.92	6.94	15.15	1.94	49.56	15.12	45.62	5.30	50.62	290.36
David Gomez	(ESP)	11.08	7.26	14.57	1.85	48.61	14.41	40.95	4.40	60.71	269.70
Indrek Turi	(EST)	11.08	6.91	13.62	2.03	51.67	14.26	39.83	4.80	59.34	290.01
Santiago Lorenzo	(ARG)	11.10	7.03	13.22	1.85	49.34	15.38	40.22	4.50	58.36	263.08
Janis Karlivans	(LAT)	11.33	7.26	13.30	1.97	50.54	14.98	43.34	4.50	52.92	278.67
Prodromos Korkizoglou	(GRE)	10.86	7.07	14.81	1.94	51.16	14.96	46.07	4.70	53.05	317.00
Hans Olav Uldal	(NOR)	11.23	6.99	13.53	1.85	50.95	15.09	43.01	4.50	60.00	281.70
Paolo Casarsa	(ITA)	11.36	6.68	14.92	1.94	53.20	15.39	48.66	4.40	58.62	296.12
Eugene Martineau	(NED)	10.99	6.84	NA	2.00	49.10	15.02	40.00	4.80	63.62	271.79
Victor Covalenco	(MDA)	11.28	7.20	13.04	1.85	51.82	15.80	38.19	NA	53.46	263.81

Tab. C.9: Resultate des Olympischen Zehnkampfs am 23.8/24.8.2004 in Athen (Teil II)

Literaturverzeichnis

Agostinelli, C. (2002) *Robust model selection in regression via weighted likelihood methdology*. Statistics & Probability Letters 64:583–639

Akaike, H. (1969) *Fitting autoregressive models for prediction*. Annals of the Institute of Statistical Mathematics 21:243–247

Akaike, H. (1970) *Statistical predictor identification*. Annals of the Institute of Statistical Mathematics 22:203–217

Akaike, H. (1971) *Determination of the number of factors by an extended maximum likelihood principle*. Institute of Statistical Mathematics, Memo. 44

Akaike, H. (1973) *Information theory and an extension of the maximum likelihood principle*. Proceeding of the Second International Symposiumon Information Theory Budapest 267–281

Akaike, H. (1974) *A new look at the statistical model identification*. IEEE Transactions on Automatic Control 19:716–723

Akaike, H. (1987) *Factor analysis and AIC*. Psychometrika 52:317–332

Allen, D.M. (1974) *The relationship between variable selection and data augmentation and a method for prediction*. Technometrics 16:125–127

Amemiya, T. (1980) *Selection of regressors*. International Econometric Review 21:331–354

Andrews, D.F., Herzberg, A.M. (1985) *Data: A collection of problems from many fields for the student and research worker*. Springer, New York

Bass, B.M., Aviolo, B., Active, P. (2003) *Multifactor Leadership Questionnaire: Feedback Report*. Mind Garden Inc., Menlo Park

Bass, B.M., Steyrer, J. (1995) *Transaktionale und transformationale Führung*. In: Kieser, A., Reber, G., Wunderer, R. (Hrsg.) Handwörterbuch der Führung S. 2053–2062. Schäffer-Poeschel, Stuttgart

Bhansali, R.H., Downham, D.Y. (1977) *Some properties of the order of an autoregressive model selected by a generalization of Akaikes FPE criterion*. Biometrika 67:413–418

Blackwell, D. (1953) *Equivalent comparisons of experiments*. Annals of Mathematical Statistics 24:265–272

Bozdogan, H. (1987) *Model selection and Akaikes Information Criterion (AIC): the general theory and its analytical extensions*. Psychometrika 52:345–370

Breimann, L. (1992) *The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error*. Journal of the American Statistical Association 87:738–754

Breimann, L. (1996a) *Heuristics of instability and stabilization in model selection*. Annals of Statistics 24:2350–2383

Breimann, L. (1996b) *Bagging predictors*. Machine Learning 24:95–122

Breimann, L., Freedmann, D. (1983) *How many variables should be entered in a regression equation*? Journal of the American Statistical Association 78:131–136

Buckland, S.T., Burnham, K.P., Augustin, N.H. (1997) *Model selection: an integral part of inference*. Biometrics 53:603–618

Burnham, K., Anderson, D. (2002) *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York

Burton, A., Altmann, D.G. (2004) *Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines*. British Journal of Cancer 91:4–8

Candolo, C., Davison, A.C., Demétrio, C.G.B. (2003) *A note on model uncertainty in linear regression*. The Statistician 52:165–177

Cattell, R. B. (1966) *The scree test for the number of factors*. Multivariate Behaviorial Research 1:245–276

Cavanaugh, J., Shumway, R. (1998) *An Akaike information criterion for model selection in the presence of incomplete data*. Journal of Statistical Planning and Inference 67:45–65

Chatfield, C. (1995) *Model uncertainty, data mining and statistical inference*. Journal of the Royal Statistical Society A 158:419–466

Chen, J.H., Shao, J. (2000) *Nearest neighbor imputation for survey data*. Journal of Official Statistics 16:113–131

Chen, J.H., Shao, J. (2001) *Jackknife variance estimation for nearest-neighbor imputation*. Journal of the American Statistical Association 96:260–269

Claeskens, G., Consentino, F. (2008) *Variable selection with incomplete covariate data*. Biometrics 64:1062–1069

Claeskens, G., Hjort, N.L. (2003) *The focused information criterion*. Journal of the American Statistical Association 98:900–916

Clayton, D., Spiegelhalter, D., Dunn, D., Pickles, A. (1998) *Analysis of longitudinal binary data from multiphase sampling*. Journal of the Royal Statistical Society B 60:71–87

Cook, R.J., Zeng, L., Yi, G.Y. (2004) *Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation*. Biometrics 60:820–828

Costa, M. (1996) *Factor analysis and information criteria*. QÜESTIIÓ 20:409–425

Cox, T.F., Dunn, R.T. (2002) *An analysis of decathlon data*. The Statistician 51:179–187

Danilov, D., Magnus, J.R. (2004) *On the harm that ignoring pretesting can cause*. Journal of Econometrics 122:27–46

Dawkins, B.P., Andreae, P.M., O'Connor, P.M. (1994) *Analysis of olympic hepathlon data*. Journal of the American Statistical Association 89:1100–1106

de Leeuw, J. (1988) *Model selection in multinomial experiments*. In: Dijkstra, T.K. (Hrsg.) On model uncertainty and its statistical implications. Lecture Notes in Economics and Mathematical Systems S. 118-138. Springer, New York

de Leeuw, J. (1992) *Introduction to Akaike (1973) Information theory and an extension of the maximum likelihood principle*. In: Kotz, S., Johnson, N.L. (Hrsg.) Breakthroughs in Statistics S. 599–609. Springer, London

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977) *Maximum likelihood estimation from incomplete data via the EM algorithm*. Journal of the Royal Statistical Association B 39:1–38

Dempster, A.P., Rubin, D.B. (1983) *Overview*. In: Madow, W.G., Olkin, I., Rubin, D.B. (Hrsg.) Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography S. 3–10. Academic Press, New York

Detel, W. (2007) *Erkenntnis- und Wissenschaftstheorie, Grundkurs Philosophie Band 4*. Reclam, Stuttgart

Draper, D. (1995) *Assesment and propagation of model uncertainty*. Journal of the Royal Statistical Society B 57:45–97

Draper, N.R., Smith, H. (1998) *Applied regression analysis*. Wiley, New York

Drechsler, J., Rässler, S. (2008) *Does convergence really matter*? In: Shalabh, Heumann, C. (Hrsg.) Recent Advances in Linear Models and Related Areas S. 341–355. Physica-Verlag, Heidelberg

Efron, B. (1979) *Bootstrap methods: another look at the jackknife*. Annals of Statistics 7:1–26

Efron, B. (1983) *Estimating the error rate of a prediction rule: improvement on cross-validation* Journal of the American Statistical Association 78:316–331

Efron, B. (1986) *How biased is the apparent error rate of a prediction rule*? Journal of the American Statistical Association 81:461–470

Efron, B., Tibsharani, R. (1991) *Improvements on cross-validation: the 632+ bootstrap method*. Journal of the American Statistical Association 92:548–560

Efroymson, M.A. (1960) *Multiple regression analysis*. In: Ralston, A., Wilf, H.S. (Hrsg.) Mathematical Methods for Digital Computers S. 191–203. Wiley, New York

Eubank, R.L. (1999) *Nonparametric regression and spline smoothing*. CRC Press, New York

Fahrmeier, L., Hamerle, A., Tutz, G. (1996) *Multivariate statistische Verfahren*. de Gruyter, Berlin

Fahrmeier, L., Kneib, T., Lang, S. (2007) *Regression - Modelle, Methoden und Anwendungen*. Springer, Heidelberg

Fieger, A. (2001) *Fehlende Kovariablenwerte bei linearen Regressionsmodellen*. Peter Lang Verlag, Frankfurt am Main

Fisher, R.A. (1924) *The influence of rainfall on the yield of wheat at Rothamsted*. Philosophical Transactions of the Royal Society of London B 213:89–142

Forster, M.R. (1998) *Parsimony and simplicity*. Universität Wisconsin-Madison, `http://philosophy.wisc.edu/forster/220/simplicity.html`

Forster, M.R., Sober, E. (1994) *How to tell wenn simpler, more unified, or less ad hoc theories will provide more accurate predictions*. British Journal for the Philosophy of Science 45:1–35

Frank, I.E., Friedmann, J.H. (1993) *A statistical view of some chemometrics regression tools*. Technometrics 35:109–135

Freedman, D.A. (1983) *A note on screening regression equations*. The American Statistician 37:152–155

Frigg, R., Hartmann, S. (2006) *Models in science*. Stanford Encyclopedia of Philosophy, `http://plato.stanford.edu/entries/models-science/`

Geisser, S. (1975) *The predictive sample reuse method with applications*. Journal of the American Statistical Association 70:320–328

George, E.I., McCulloch, R.E. (1993) *Variable selection via Gibbs sampling*. Journal of the American Statistical Association 88:881-889

George, E.I., McCulloch, R.E. (1997) *Approaches for bayesian variable selection*. Statistica Sinica 7:339–373

Gettier, E.L. (1963) *Is justified true belief knowledge*? Analysis 23:121–123

Golub, G.H., Heath, M., Wahba, G. (1979) *Generalized cross-validation as a method for choosing a good ridge parameter*. Technometrics 21:215–223

Gorman, J.W., Toman, R.J. (1966) *Selection of variables for fitting equations to data*. Technometrics 8:27–51

Gottardo, R. (2008) *EMV: Estimation of missing values for a data matrix*. R package version 1.3.1

Grünwald, P.D. (2005) *A tutorial introduction to the minimum description length principle*. In: Grünwald, P.D., Myung, I.J., Pitt, M.A. (Hrsg.) Advances in Minimum Description Length: Theory and Applications. MIT Press, Cambridge

Grünwald, P.D. (2007) *The minimum description length principle*. MIT Press, Cambridge

Guttman, L. (1954) *Some necessary conditions for common factor analysis*. Psychometrika 19:149–161

Hannan, E.J. (1980) *The estimation of the order of an ARMA process*. Annals of Statistics 8:1071–1081

Hannan, E.J. (1981) *Estimating the dimension of a linear system*. Journal of Multivariate Analysis 11:459–473

Hannan, E.J., Quinn, B.G. (1979) *The determination of the order of an autoregression*. Journal of the Royal Statistical Society B 41:190–195

Hansen, B.E. (2007) *Least squares model averaging*. Econometrica 75:1175–1189

Hansen, B.E. (2008a) *Least squares forecast averaging*. Journal of Econometrics 146:342:342–350

Hansen, B.E. (2008b) *Averaging estimators for autoregressions with a near unit root*. Journal of Econometrics, akzeptiert

Hansen, B.E. (2009) *Averaging estimators for regressions with a possible structural break*. Econometric Theory 35:1498–1514

Hansen, B.E., Racine, J. (2009) *Jackknife model averaging*. Arbeitspapier
`http://www.ssc.wisc.edu/~bhansen/papers/jma.pdf`

Hastie, T., Tibsharani, R. (1990) *Generalized additive models*. Chapman and Hall, London

Hastie, T., Tibsharani, R., Friedman, J. (2001) *The elements of statistical learning*. Springer, New York

Hayes, J.P., Weikel, J.M., Huso, M.M.P. (2003) *Response of birds to thinning young Douglas-Fir forests*. Ecological Applications 13:1222–1232

Hens, N., Aerts, M., Molenberghs G. (2006) *Model selection for incomplete and design based samples*. Statistics in Medicine 25:2502–2520

Hjort, U. (1982) *Model selection and forward validation*. Scandinavian Journal of Statistics 9:95–105

Hjort, L., Claeskens, G. (2003) *Frequentist model average estimators*. Journal of the American Statistical Association 98:879–945

Hjort, N.L., Claeskens, G. (2006) *Focused information criteria and model averaging for Cox's hazard regression model*. Journal of the American Statistical Association 101:1449–1464

Hodges, J.S. (1987) *Uncertainty, policy analysis and statistics*. Statistical Science 2:259-291

Hoerl, A.E., Kennard, R.W. (1970) *Ridge regression: biased estimation for nonorthogonal problems*. Technometrics 12:55–67

Hoeting, J., Madigan, D., Raftery, E., Volinsky, C. (1999) *Bayesian model averaging: a tutorial*. Statistical Science 14:382–417

Honaker, J., King, G. (2010) *What to do about missing data in time series cross-section data*. American Journal of Political Science, akzeptiert

Honaker, J., King, G., Blackwell, M. (2008) *Amelia II: A program for missing data*. R Package version 1.1–33, `http://gking.harvard.edu/amelia`.

Horton, N.J., Kleinman, K.P. (2007) *Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models*. The American Statistician 61:79–90

Horton, N.J., Switzer, S.S. (2005) *Statistical methods in the journal*. New England Journal of Medicine 353:1977–1979

Horvitz, D., Thompson, D. (1952) *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association 47:663–685

Huber, P.J. (1964) *Robust estimation of a location parameter*. Annals of Mathematical Statistics 35:73–101

Huber, P.J. (1981) *Robust statistics*. Wiley, New York

Hurvich, C.M., Tsai, C.L. (1989) *Regression and time series model selection in small samples*. Biometrika 76:297–307

Hurvich, C.M., Tsai, C.L. (1995) *Model selection for extended quasi-likelihood models in small samples*. Biometrics 51:1077–1084

Kabaila, P. (2002) *On variable selection in linear regression*. Econometric Theory 18:913–925

Karagrigoriou, A. (1997) *Asymptotic efficiency of the order selection of a nongaussian AR process*. Statistica Sinica 7:407–423

Kass, R.E., Raftery, A.E. (1995) *Bayes factors*. Journal of the American Statistical Association 90:773–795

Kastner, C. (2001) *Fehlende Werte bei korrelierten Beobachtungen*. Peter Lang Verlag, Frankfurt am Main

Khattree, R., Naik, D.N. (2000) *Multivariate data reduction and discrimination with SAS software*. SAS Publishing, Wiley, New York

King, G., Honaker, J., Joseph, A., Scheve, K. (2001) *Analyzing incomplete political science data: an alternative algorithm for multiple imputation*. American Political Science Review 95:49–69

Klaußner, A. (2007) *Phasenangepasste Führung von Wachstumsunternehmen*. Dissertation, International University Schloss Reichartshausen

Kockelkorn, U. (2000) *Lineare statistische Methoden*. Oldenbourg Verlag

Konishi, S., Kitagawa, G. (1996) *Generalised information criteria in model selection*. Biometrika 83:875–890

Kraft, L.G. (1949) *A device for quantizing, grouping, and coding amplitude modulated pulses*. MS Thesis, Electrical Engineering Department, Massachusetts Institute of Technology

Kullback, S. (1959) *Information theory and statistics*. Wiley, New York

Kullback, S., Leibler, R. (1951) *On information and sufficiency*. Annals of Mathematical Statistics 22:79-86

Lantermann, A. (2001) *Schwarz, Wallace, and Rissanen: intertwining themes in theories of model selection*. International Statistical Review 69:185–212.

Laud, P.W., Ibrahim, J.G. (1995) *Predictive model selection*. Journal of the Royal Statistical Society B 57:247–262

Leamer, E.E. (1978) *Specification searches*. Wiley, New York

Lebreton, J-D., Burnham, K.P., Clobert, J., Anderson, D.R. (1992) *Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies*. Ecological Monograph 62:67–118

Leeb, H., Pötscher, B.M. (2003) *The finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations*. Econometric Theory 19:100–142

Leeb, H., Pötscher, B.M. (2005) *Model selection and inference: facts and fiction*. Econometric Theory 21:21–59

Leeb, H., Pötscher, B.M. (2006a) *Can one estimate the conditional distribution of post-model-selection estimators*? Annals of Statistics 34:2554–2591

Leeb, H., Pötscher, B.M. (2006b) *The distribution of model averaging estimators and an impossibilty result regarding its estimation*. IMS Lecture Notes – Time Series and Related Topics 52:113–129

Leeb, H., Pötscher, B.M. (2008a) *Model selection*. In: Andersen, T.G., Davis, R.A., Kreiß, J.-P., Mikosch, T. (Hrsg.) The Handbook of Financial Time Series S. 785–821. Springer, New York

Leeb, H., Pötscher, B.M. (2008b) *Can one estimate the unconditional distribution of post-model-selection estimators*? Econometric Theory 24:338–376

Leung, G., Barron, A.R. (2006) *Information theory and mixing least squares regressions*. IEEE Transaction on Information Theory 52:3396–3410

Li, K. (1987) *Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set*. Annals of Statistics 15:958–975

Li, M., Vitànyi, P. (1997) *An introduction to Kolmogorov complexity and its applications*. Springer, New York

Liang, H., Zou, G., Wan, A.T.K., Zhang, X. (2010) *On optimal weight choice in a frequentist model average estimator*. `http://fbstaff.cityu.edu.hk/msawan/research1.htm`

Liang, K.-Y., Zeger, S.L. (1986) *Longitudinal data analysis using generalized linear models*. Biometrika 73:13-22

Linhart, H., Zucchini, W. (1986) *Model selection*. Wiley, New York

Little, R. (1992) *Resgression with missing X's: a review*. Journal of the American Statistical Association 87:1227–1237

Little R., Rubin D. (2002) *Statistical analysis with missing data*. Wiley, New York.

Machado, J.A.F. (1993) *Robust model selection and M-estimation*. Econometric Theory 9:478–493

Mackenzie, D.I., Nichols, J.D., Sutton, N., Kawanishi, K., Bailey, L.L. (2005) *Improving inferences in population studies of rare species that are detected imperfectly*. Ecology 86:1101–1113

Madigan, D., Raftery, A.E. (1994) *Model selection and accounting for model uncertainty in graphical models using Occam's window*. Journal of the American Statistical Association 89:1535–1546

Magnus, J.R., Durbin, J. (1999) *Estimation of regression coefficients of interests when other regression coefficients are of no interest*. Econometrica 67:639–643

Magnus, J.R., Powell, O., Prüfer, P. (2008) *A comparsion of two averaging techniques with an application to growth empirics*. Tilburg University, CentER Discussion Paper 2008–39

Mallows, C. L. (1964) *Choosing variables in a linear regression: a graphical aid*. Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas

Mallows, C.L. (1973) *Some comments on Cp*. Technometrics 15:661–675

Maronna, R., Martin, D., Yohai, V. (2006) *Robust statistics - theory and methods*. Wiley, New York

McCullagh, P., Nelder J.A. (1989) *Generalized linear models*. Chapmann and Hall, London

Miller, A.J. (1990) *Subset selection in regression*. Chapman and Hall, London

Molenberghs, G., Kenward, M.G. (2007) *Missing data in clinical studies*. Wiley, New York

Nishii, R. (1984) *Asymptotic properties of criteria for selection of variables in multiple regression*. Annals of Statistics 12:758–765

Nittner, T. (2003) *Fehlende Daten in Additiven Modellen*. Peter Lang Verlag, Frankfurt am Main

O'Hagan, A. (1995) *Fractional bayes factors for model comparisons*. Journal of the Royal Statistical Society B 57:99–138

Paulsen, J. (1984) *Order determination of multivariate autoregressive time series with unit roots*. Journal of Time Series Anaysis 5:115–127

Pötscher, B.M. (1989) *Model selection under nonstationary: autoregressive models and stochastic linear regression models*. Annals of Statistics 17:1257–1274

Pötscher, B.M. (1991) *Effect of model selection on inference*. Econometric Theory 7:163–185

Pümpin, C., Prange, J. (1991) *Management der Unternehmensentwicklung. Phasengerechte Führung und der Umgang mit Krisen*. Campus Verlag, Frankfurt am Main

Quinn, B.G. (1980) *Order determination for a multivariate autoregression*. Journal of the Royal Statistical Society B 42:182–185

Raftery, A.E., Hoeting, J., Volinsky, C., Painter I., Yeung, K.Y. (2006) *BMA: Bayesian Model Averaging. R package version 3.03.* `http://www.r-project.org`, `http://www.research.att.com/~volinsky/bma.html`

Raftery, A.E., Madigan, D., Hoeting, J. (1997) *Bayesian model averaging for linear regression models*. Journal of the American Statistical Association 92:179–191

Raftery, A.E., Madigan, D., Volinsky, C.T. (1996) *Accounting for model uncertainty in survival analysis improves predictive performance*. In: Bernado, J., Berger, J., Dawid, A., Smith, A. (Hrsg.) Bayesian Statistics 5 S. 323–349. Oxford University Press

Rao, C.R., Toutenburg, H., Shalabh, Heumann, C. (2008) *Linear models and generalizations – least squares and alternatives*. Springer, New York

Rao, C.R., Wu, Y. (2001) *On model selection*. IMS Lecture Notes 38:1–64

Reid, J.M., Bignal, E.M., Bignal, S., McCracken, D.I., Monaghan, P. (2003) *Age-specific reproductive performance in red-billed choughs pyrrhocorax pyrrhocorax: patterns and processes in a natural population*. The Journal of Animal Ecology 72:765–776

Rissanen, J. (1978) *Modeling by the shortest data description*. Automatica 14:465–471

Rissanen, J. (1986) *Stochastic complexity and modeling*. Annals of Statistics 14:1080–1100

Ronchetti, E. (1985) *Robust model selection in regression*. Statistics & Probability Letters 3:21–23

Ronchetti, E. (1997) *Robustness aspects of model choice*. Statistica Sinica 7:327–338

Ronchetti, E., Field, C., Blanchard, W. (1997) *Robust linear model selection by cross-validation*. Journal of the American Statistical Association 92:1017–1023

Ronchetti, E., Staudte, R.G. (1994) *A robust version of Mallows C_p*. Journal of the American Statistical Association 89:550–559

Rubin, D.B. (1976) *Inference and missing data*. Biometrika 63:581–592

Rubin, D.B. (1978) *Multiple imputation in sample surveys – a phenomenological bayesian approach to nonresponse*. American Statistical Association Proceedings of the section on Survey Research Methods 20–40

Rubin, D.B. (1996) *Multiple Imputation After 18+ Years*. Journal of the American Statistican Association 91:473–489.

Schafer, J. (1997) *Analysis of incomplete multivariate data*. Chapman and Hall

Schomaker, M. (2006) *Neuere Ansätze für Kriterien zur Modellselektion bei Regressionsmodellen unter Berücksichtigung der Problematik fehlender Daten*. Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München

Schomaker, M., Wan, A.T.K., Heumann, C. (2010) *Frequentist model averaging with missing observations*. Computational Statistics & Data Analysis, akzeptiert

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

Searle, S.R. (1971) *Linear Models*. Wiley, New York

Sen, A., Srivastava, M. (1990) *Regression analysis - theory, methods and applications*. Springer, New York

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005) *ROCR: visualizing classifier performance in R*. Bioinformatics 21:3940–3941

Shannon, E. (1948) *A mathematical theory of communication*. The Bell System Technical Journal 27:379–423, 623–653

Shao, J. (1996) *Bootstrap model selection*. Journal of the American Statistical Association 91:655–665

Shao, J. (1997) *An asymptotic theory for linear model selection*. Statistica Sinica 7:221–264

Shibata, R. (1980) *Asymptotically efficient selection of the order of the model for estimating parameters of a linear process*. Annals of Statistics 8:147–164

Shibata, R. (1981) *An optimal selection of regression variables*. Biometrika 68:45–54

Shibata, R. (1984) *Approximate efficiency of a selection procedure for the number of regression variables*. Biometrika 71:43–49

Shibata, R. (1989) *Statistical aspects of model selection*. In: Willems, J.C. (Hrsg.) From Data to Model S. 215–40. Springer, Berlin

Shimodaira, H. (1994) *A new criterion for selecting models from partially observed data* In: Cheesman, P., Oldford, R.W. (Hrsg.) Selecting Models from Data: Artificial Intelligence and Statistics IV S. 21–29. Springer, New York

Sober, E. (1981) *The principle of parsimony.* British Journal for the Philosophy of Science 32:145–156

Sober, E. (2000) *Instrumentalism, parsimony and the Akaike framework.* Philosophy of Science 69:112–123

Sober, E. (2002) *What is the problem of simplicity?* In: Keuzenkamp, H., McAleer, M., Zellner, A. (Hrsg.) Simplicity, Inference, and Econometric Modelling S. 13–32. Cambridge University Press

Sommer, S., Huggins, R.M. (1996) *Variables selection using the Wald test and a robust C_p.* Applied Statistics 45:15–29

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002) *Bayesian measures of model complexity and fit.* Journal of the Royal Statistical Society B 65:583–639

Stone, M. (1974) *Cross-validatory choice and assessment of statistical predictions.* Journal of the Royal Statistical Society B 36:111–147

Sugiura, N. (1978) *Further analysis of the data by Akaike's Information Criterion and the finite corrections.* Communications in Statistics – Theory and Methods A7:13–26

Takeuchi, K. (1976) *Distribution of informational statistics and a criterion of model fitting.* Suri-Kagaku (auf Japanisch) 153:12–18

Thall, P.F., Russell, K.E., Simon, R.M. (1997) *Variable selection in regression via repeated data splitting.* Journal of Computational & Graphical Statistics 6:416–434

Thall, P.F., Simon, R.M., Grier, D.A. (1992) *Test-based variable selection via cross-validation.* Journal of Computational & Graphical Statistics 1:41–61

Tibsharani, R. (1996) *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society B 58:267–288

Tibsharani, R., Hinton, G. (1998) *Coaching variables for regression and classification.* Statistics and Computing 8:25–33

Tierney, L., Kadane, J.B. (1986) *Accurate approximation for posterior moments and marginal densities.* Journal of the American Statistical Association 81:82–86

Tsay, R.S. (1984) *Order selection in nonstationary autoregressive models.* Annals of statistics 12:1425–1433

Ulbricht, J. (2010) *Variable selection in generalized linear models.* Dissertation, Institut für Statistik, Ludwig-Maximilian-Universität München

Vach, W. (1994) *Logistic regression with missing values in the covariates*. In: Fienberg, S., Gani, J., Krickeberg, K., Olkin, I., Wermuth, N. (Hrsg.) Lecture Notes in Statistics Vol. 86. Springer, New York

van Buuren, S., Boshuizen, H.C., Knook, D.L. (1999) *Multiple imputation of blood pressure covariates in survival analysis*. Statistics in Medicine 18:681–694

van Fraasen, B. (1980) *The scientific image*. Oxford University Press, New York

Volinsky, C.T., Madigan, D., Raftery, A.E., Kronmal, A.E. (1997) *Bayesian model averaging in proportional hazards models: assessing the risk of a stroke*. Journal of the Royal Statistical Society C 46:433–448

Wallace, T.D. (1972) *Weaker criteria and tests for linear restrictions in regression*. Econometrica 40:689-698

Wallace, C., Boulton, D. (1968) *An information measure for classification*. The Computer Journal 11:195–209

Wan, A.T.K., Zhang, X., Zou, G. (2010) *Least squares model combining by Mallows criterion*. Journal of Econometrics, akzeptiert

Wei, C.Z. (1992) *On predictive least squares principle*. Annals of Statistics 20:1–42

Wiener, N. (1948) *Cybernetics: or control and communication in the animal and the machine*. MIT Press, Cambridge

Woods, S.L., Froelicher, S.L., Adams-Motzer, S., Bridges, E. (2004) *Cardiac nursing*. Lippincott Williams & Wilkins, Philadelphia

Yan, J. (2007) *Enjoy the joy of copulas: with package copula*. Journal of Statistical Software 21:1–21

Yang, Y. (2001) *Adaptive regression by mixing*. Journal of the American Statistical Association 96:574–586

Yang, Y. (2003) *Regression with multiple candidate models: selecting or mixing*? Statistica Sinica 13:783–809.

Yuan, Z., Yang, Y. (2005) *Combining linear regression models: when and how*? Journal of the American Statistical Association 100:1202–1214

Zhang, X., Wan, A.T.K., Zhou, S.Z. (2010) *Focussed information criteria, model selection and model averaging in a Tobit model with non-zero threshold*.
`http://fbstaff.cityu.edu.hk/msawan/research1.htm`

Zhou, Y., Wan, A.T.K., Wang, X. (2008) *Estimating equation inference with missing data*. Journal of the American Statistical Association 103:1187–1199

Zou, H., Hastie, T. (2005) *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society B 67:301–320