Neuere Ansätze für Kriterien zur Modellselektion bei Regressionsmodellen unter Berücksichtigung der Problematik fehlender Daten

Diplomarbeit
von
Michael Schomaker

Betreuung:
PD Dr. Christian Heumann
Prof. Dr. Dr. Helge Toutenburg

Oktober 2006

Institut für Statistik
Ludwig-Maximilians-Universität München

„Classical statistics as developed in the first half of the 20th century has two obvious deficiencies from practical applications: an overreliance on the normal distribution and failure to account for model selection. The first of these was dealt with in the century's second half [...] Model selection, the data-based choice [...] remains mostly terra incognita as far as statistical inference is concerned. “

Bradley Efron

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Sowohl die Modellselektion, als auch der Umgang mit fehlenden Daten sind wichtige Teilbereiche der Statistik, die bei einer Datenanalyse ihre Anwendung finden [15]. Für sich allein gesehen wird beiden Methoden große Aufmerksamkeit zuteil, in ihrer Kombination fällt die Betrachtung meist jedoch sehr spärlich aus.

Fehlende Daten zu verwerfen kann zu erheblichem Informations- und Aussageverlust führen, weswegen die Statistik eine Vielzahl an Methoden bereitstellt, die ermöglichen fehlende Werte zu ersetzen [9]. Selbstverständlich kann im Zuge einer Regressionsanalyse Modellselektion anhand eines aufgefüllten Datensatzes durchgeführt werden, ob und welche Methoden dabei jedoch den größten Erfolg versprechen und ob eine Verallgemeinerung in diesem Sinne überhaupt möglich ist, ist bisher noch ungeklärt.

Neuere Forschungsergebnisse [6] lassen vermuten, dass auch alternative Lösungsansätze Erfolg versprechen können. Dabei werden fehlende Werte nicht durch andere ersetzt, sondern nur die vorhandenen Fälle – gewichtet – zur Entscheidungsfindung miteinbezogen.

In dieser Arbeit soll nun erörtert werden, wie sich der Umgang mit fehlenden Daten auf die Modellwahl auswirkt, und ob für eine solche Situation ein gewichtetes Gütemaß oder das Ersetzen fehlender Werte zu bevorzugen ist.

In Kapitel 2 richtet sich der Fokus in erster Linie auf den Umgang mit fehlenden Daten im Allgemeinen, Kapitel 3 erläutert Konzeption und Idee verschiedener Gütemaße der Modellselektion.

Der Kern der Arbeit ist Kapitel 4, in dem anhand zahlreicher Simulationsstudien die oben beschriebene Fragestellung genauestens untersucht werden soll. Verschiedene Methoden der Imputation werden dabei alternativen Lösungsansätzen gegenübergestellt. Kapitel 5 liefert einen Vergleich zu bisherigen Forschungsarbeiten, speziell zu den Untersuchungen von Hens, Aerts und Molenberghs aus dem Jahre 2005 [6].

In Kapitel 6 wird die Problematik der Modellselektion diskutiert, alternative Methoden werden vorgestellt und untersucht. Kapitel 7 konkretisiert die

Methoden und Ergebnisse dieser Arbeit noch einmal an einem Datenbeispiel.

Zum Abschluss werden in Kapitel 8 noch einmal die Kernthesen dieser Arbeit dargelegt, Hinweise auf Modifikationen und weitere Forschungsarbeiten werden ebenfalls gegeben.

2. Umgang mit fehlenden Daten

Bei verschiedensten Fragestellungen und Sachverhalten kann es vorkommen, dass die einem vorliegenden Daten nur teilweise vollständig sind, also Beobachtungen fehlen.
Dies kann mehrere Gründe haben. Bei Befragungen können Daten verloren gegangen sein oder Teilnehmer haben möglicherweise keine Antwort gegeben. Bei Langzeitstudien sind Versuchspersonen eventuell unbekannt verzogen und bei naturwissenschaftlichen Experimenten war es vielleicht nicht möglich eine Messung durchzuführen. So kann beispielsweise ein Objekt zerstört worden sein oder eine Röntgenaufnahme unscharf vorliegen.
Egal um welche Fragestellung es sich auch handeln mag, so stellt sich immer die Frage, ob den fehlenden Werten eine Systematik unterliegt. Kann man davon ausgehen, dass ein Wert nur aus 'Zufall' fehlt, oder dass andere Komponenten bzw. Variablen das Fehlen systematisch beeinflussen? Um dies zu klären bedarf es zunächst einiger grundlegender Definitionen.

2.1 Zufallsmechanismen und grundlegende Begriffe

Betrachtet man eine Variable Y bei der n Einheiten vollständig beobachtet wurden, und eine Variable X für die nur $n - m$ Werte vorliegen, so können sich unter anderem die drei folgenden Situationen ergeben:

- Fall 1: Die fehlenden Werte hängen weder von X, noch von Y ab.
- Fall 2: Die fehlenden Werte hängen von Y, nicht aber von X ab.
- Fall 3: Die fehlenden Werte hängen von X und Y ab.

Für den ersten Fall können die fehlenden Werte als 'Missing at Random' (MAR), die beobachteten Werte als 'Observed at random' bezeichnet werden. Passenderweise werden nun die fehlenden Daten als 'Missing completely at random' (MCAR) definiert [9].

Im zweiten Fall können die fehlenden Werte als MAR angesehen werden, die beobachteten x-Werte dagegen bilden hier jedoch nicht notwendigerweise eine zufällige Substichprobe von x, und können daher auch nicht mehr als MCAR angesehen werden. Im dritten Fall sind die Daten weder MAR noch OAR.

Weicht man nun von der Vorstellung ab, dass nur zwei Variablen in Einfluss und Beziehung stehen, so kann die Standardsituation innerhalb einer Datenanalyse wie folgt verallgemeinert werden [15]:

$$X_* = \begin{pmatrix} x_{11} & \dots & \dots & x_{1m} \\ \vdots & * & & \vdots \\ \vdots & & & * \\ \vdots & & * & \vdots \\ x_{n1} & \dots & \dots & x_{nm} \end{pmatrix}$$

In der Datenmatrix X_* ist dabei durch ein '$*$'-Symbol angedeutet, dass einzelne Werte innerhalb der Daten fehlen können. Im Zuge einer Regressionsanalyse würde ein Vektor $x_{.j}$ also die Zielgröße darstellen, andere Vektoren würden mögliche Einflussgrößen repräsentieren.

2.2 Betrachtung der beobachteten Werte

Im Folgenden sollen nun Methoden im Umgang mit fehlenden Werten kurz vorgestellt und analysiert werden. Sie gehen im wesentlichen auf Little und Rubin [9] zurück. In den Arbeiten der AG Toutenburg (Projekt C3 im SFB 386) wurden zahlreiche Modifikationen für Lineare Modelle erarbeitet.

Complete Case Analysis

Dem Namen entsprechend werden bei dieser Methode nur diejenigen Fälle betrachtet, für die alle Beobachtungen vorliegen [9]. Beim Fehlen mindestens eines Wertes in der Zielvariable oder bei einer der Einflussgrößen, wird der entsprechende Fall nicht in der weitergehenden Betrachtung miteinbezogen.

Nachteilig wirkt sich dabei vor allem der potentielle Informationsverlust beim Verwerfen aller unvollständigen Fälle aus. Besonders bei einer großen Anzahl an Variablen besteht die Gefahr viel an Information zu verlieren. Schätzungen können verzerrt werden, wenn das Fehlen der Werte systematisch bedingt

ist, und die MCAR-Annahme unter diesen Umständen nicht mehr vertreten werden kann. Die Art dieser Verzerrungen hängt dabei sehr stark vom Mechanismus des Fehlens ab [9].

Ein weiterer, nicht zu vernachlässigender, Problempunkt besteht darin, dass bei der Streichung einzelner Fälle Schichtungseffekte entstehen können. So könnte beispielsweise innerhalb einer bestimmten Personengruppe der Anteil an nicht beantworteten Fragen höher sein als in einer anderen. Um diesem Sachverhalt Rechnung zu tragen, sollten Homogenitätstests durchgeführt werden [15], um festzustellen ob ein Schichtungseffekt vorliegt oder nicht.

Available Case Analysis

Eine Alternative zur Complete Case Analysis bietet die Available Case Analysis. Es werden hier all diejenigen Fälle betrachtet, die bei den interessierenden Variablen vollständig beobachtet wurden.
Auch wenn hier alle erhältlichen Werte verarbeitet werden, so ist doch das Problem unterschiedlicher Samplegröße bei unterschiedlichen Variablen unübersehbar. Dies kann zu nicht unerheblichen Schwierigkeiten führen. So ist anzumerken, dass jede einzelne Schätzung der Erwartungswerte (bzw. der Kovarianzen) auf unterschiedlichen Stichprobenumfängen beruht. Daher ist eine für alle Spalten aus X_* geschätzte Kovarianzmatrix unter Umständen nicht mehr positiv definit [15].

2.3 Ersetzen der fehlenden Werte

2.3.1 Gängige Methoden

Im Folgenden sollen Methoden aufgezeigt werden, die die fehlenden Werte durch neue Werte ersetzen (Imputation). Die Wahl der Methode sollte vom jeweiligen Sachverhalt abhängig gemacht werden.

Mean Imputation

Das arithmetische Mittel der beobachteten Fälle ersetzt hierbei alle anderen nicht beobachteten – und damit fehlenden – Werte innerhalb einer Variable. Für einen fehlenden Wert innerhalb einer Datenmatrix folgt damit:

$$x_{ij,mis} = \bar{x}_{\cdot j,C}$$

Auch gewichtete Mittelwertsschätzungen oder die Betrachtung des Medians stellen in entsprechendem Zusammenhang eine sinnvolle Möglichkeit der Imputation dar.

Durch eine Mittelwertsimputation erhöht sich – im Verhältnis zu den 'Complete Cases' – zwar der Stichprobenumfang, nicht jedoch die Varianz. Dies führt zu einer Unterschätzung derselben, und sollte stets beachtet werden [9].

Ein großer Nachteil ist des Weiteren, dass es schwer zu sagen ist, wann und unter welchen Umständen diese Methode zu vertreten ist. Wird für eine Variable 'Einkommen' beispielsweise eine Mittelwertsimputation durchgeführt, neigt die Methode dazu Armut und Reichtum zu unterschätzen.

Für eine genauere Diskussion dieser Imputationsmethode sei auf Kapitel 2.3.2 verwiesen.

Hot deck Imputation

Hot deck Imputation kann im Allgemeinen als eine Methode angesehen werden, bei der die imputierten Werte aus einer geschätzten Verteilung ausgewählt werden. Größtenteils wird hierfür die empirische Verteilung genommen, die aus den beobachteten Werten gewonnen wird [9].

Positiv zu bewerten ist die Tatsache, dass das Aussehen der Verteilung dabei keine Rolle spielt und daher unter wenig komplexen Bedingungen sehr schnell effiziente Ergebnisse erzielt werden können. Ist hinter dem Fehlen der Werte jedoch eine Systematik zu erkennen (z.B. eine höhere Fehlwahrscheinlichkeit bei großen Werten), so treten schnell Verzerrungen auf [9].

Auch wenn diese Methode in der Praxis sehr populär ist, so sei darauf hingewiesen, dass bisher noch kein konsistentes, theoretisches Konstrukt hierfür geschaffen wurde und daher auch größere tiefgründige Untersuchungen Mangelware sind.

Cold deck imputation

Fehlende Werte werden durch einen konstanten Wert aus einer externen Quelle oder einen Erfahrungswert aus früheren Untersuchungen ersetzt.

Dieses Vorgehen mag in Spezialfällen sinnvoll und vielversprechend klingen, im Zuge allgemeiner Untersuchungen bietet sich jedoch kein gutes Fundament um weitergehende Erörterungen anzustreben.

Regression Imputation

Es werden die fehlenden Werte durch die vorhergesagten Werte einer Hilfsregression ersetzt [15]. Die entsprechenden Parameter werden dabei durch eine Regression innerhalb der vollständigen Fälle bestimmt und dann auf die unvollständigen Fälle angewandt. Betrachtet man die unvollständige Datenmatrix X_*, so ergibt sich für die Regression innerhalb der vollständigen Fälle bei einer Zielgröße $X_{\cdot j}$ und den möglichen Einflussgrößen $X_{\cdot k}$:

$$X_{\cdot j,C} = \gamma_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \gamma_k X_{\cdot k,C} + \epsilon\,, \qquad \epsilon \sim N(0, \sigma^2) \tag{2.1}$$

Vier Möglichkeiten der Imputation sollen nun kurz vorgestellt und erläutert werden:

1. Die fehlenden (missing) Werte werden durch die geschätzten Koeffizienten der Regression (2.1) vorhergesagt:

$$X^{(1)}_{\cdot j,mis} = \hat{\gamma}_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \hat{\gamma}_k X_{\cdot k,mis} \tag{2.2}$$

2. Alternativ kann dem so ermittelten Wert noch ein Residuum hinzugefügt werden um die Unsicherheit des vorhergesagten Wertes widerzuspiegeln. Man erhält also für die fehlenden Werte:

$$X^{(2)}_{\cdot j,mis} = \hat{\gamma}_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \hat{\gamma}_k X_{\cdot k,mis} + \epsilon\,, \qquad \epsilon \sim N(0, \hat{\sigma}^2) \tag{2.3}$$

3. Als weitere Möglichkeit bietet es sich an, auch die Unsicherheit bei den Schätzungen der Regressionskoeffizienten zu berücksichtigen. Man zieht daher $\tilde{\gamma}$ aus einer $N((\hat{\gamma}_0, \ldots, \hat{\gamma}_k); \hat{\sigma}^2(Z'Z)^{-1})$-Verteilung. Z bezeichnet dabei die Designmatrix. Die fehlenden Werte können für diesen Ansatz daher wie folgt bestimmt werden:

$$X^{(3)}_{\cdot j,mis} = \tilde{\gamma}_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \tilde{\gamma}_k X_{\cdot k,mis} + \epsilon, \qquad \epsilon \sim N(0, \hat{\sigma}^2) \tag{2.4}$$

4. Die Eigenschaft

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \tag{2.5}$$

berücksichtigt die Unsicherheit bezüglich der Schätzung von $\hat{\sigma}^2$. Um diese Unsicherheit in einen Imputationsansatz miteinzubeziehen, müsste die wahre Varianz σ^2 bekannt sein. Da dies in der Regel jedoch nicht der Fall ist, kann für eine Schätzung $\tilde{\sigma}^2$, die diesem Sachverhalt Rechnung tragen soll, Formel 2.5 wie folgt modifiziert werden:

$$(n-p)\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \sim \chi^2_{n-p} \tag{2.6}$$

Dadurch ist es möglich einen alternativen Wert für $\hat{\sigma}^2$, nämlich $\tilde{\sigma}^2$, aus einer $\hat{\sigma}^2 \cdot \chi^2_{n-p} \backslash (n-p)$–Verteilung zu erhalten. Die Variable p steht hierbei für die Anzahl der zu schätzenden Parameter. Man erhält als weitere Möglichkeit der Imputation:

$$X^{(4)}_{\cdot j,mis} = \tilde{\gamma}_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \tilde{\gamma}_k X_{\cdot k,mis} + \epsilon, \qquad \epsilon \sim N(0, \tilde{\sigma}^2) \tag{2.7}$$

Fehlen bei mehr als bei einer Variable Werte, und soll für jede davon eine Regressionsimputation durchgeführt werden, so stellt sich die Frage, ob für alle Variablen nur die 'Complete Cases' betrachtet werden, oder ob bei einer sukzessiven Auffüllung auch bereits imputierte Werte in die Berechnung miteinbezogen werden sollen. Je nach Problemstellung kann für die eine oder die andere Methode argumentiert werden.

Im Rahmen der Simulationsstudien in Kapitel 4 wurden stets nur die komplett vollständigen Fälle untersucht.

Multiple Imputation

Es werden $k \geq 2$ Werte für einen fehlenden Wert eingesetzt [9] [15]. Die k vervollständigten Datensätze können dann mit der gewünschten Methode analysiert werden, so dass man dann k Schätzungen für den interessierenden Parameter erhält. Aus diesen k Schätzungen wird anschließend eine endgültige Schätzung kombiniert, zum Beispiel über deren Mittelwert. Es ergibt sich also die Schätzung

$$\hat{\bar{\theta}} = \sum_{l=1}^{k} \frac{\hat{\theta}_l}{k}$$

für die multiple Imputation. Seien nun $\hat{\theta}_l$, W_l, $l = 1, \ldots, k$ die Schätzungen und deren Varianz einer k-fachen Imputation. Da k vervollständigte Datensätze vorliegen, existiert für den Schätzer der multiplen Imputation folglicherweise sowohl eine Variabilität zwischen den Imputationen

$$B_k = \frac{1}{k-1}\sum_{l=1}^{k}(\hat{\theta}_l - \bar{\hat{\theta}})^2 \,,$$

als auch innerhalb der verschiedenen Imputationen

$$\bar{W}_k = \sum_{l=1}^{k}\frac{\hat{W}_l}{k} \,.$$

Dass die Gesamtvariabilität

$$T_k = \bar{W}_k + \frac{k+1}{k}B_k$$

die Variabilität zwischen den einzelnen Datensätzen berücksichtigt, erweist sich insofern als großer Vorteil, als dass damit einer Varianzunterschätzung entgegengewirkt wird [9].

Einzig nennenswerter Nachteil einer multiplen Imputation im Vergleich zu einer einfachen, ist ein Mehraufwand in der Programmierung.

Alternative Multiple Imputationen

Auch wenn eine klassische multiple Imputation eine Varianzunterschätzung vermeidet, und somit große Vorteile erwirken kann, so stellt sich die Frage, wie genau unter der Problemstellung der Modellselektion eine konkrete Anwendung der Methode aussehen könnte. Hierbei interessieren ja nicht die Schätzungen der Parameter, sondern ein konkreter Wert eines Gütekriteriums (siehe auch Kapitel 3 und 4). Auch im Hinblick auf die Simulationsstudien in Kapitel 4 bieten sich im Wesentlichen zwei Möglichkeiten an:

1. Gemäß der Grundidee der multiplen Imputation werden für k verschiedene Datensätze die Parameter des Modells geschätzt und für jedes Modell ein Gütekriterium berechnet. Der Mittelwert aller k Werte des Gütekriteriums dient zur Entscheidungsfindung für ein Modell

2. Es wird eine Kombination der k verschiedenen Imputationen in einen Datensatz eingesetzt, und daraus ein Gütemaß errechnet, welches dem Vergleich mit anderen Modellen dient

In Kapitel 4 wurde die zweite Möglichkeit der hier vorgestellten alternativen multiplen Imputationen angewendet. Die grundlegende Eigenschaft einer Vermeidung der Varianzunterschätzung bei einer multiplen Imputation konnte dadurch jedoch nicht ausgenützt werden.

Ergänzende Bemerkungen

Die Liste der hier aufgeführten Methoden im Umgang mit fehlenden Daten ist keinesfalls vollständig. Zahlreiche weitere Methoden, sowie Modifikationen und Kombinationen der aufgelisteten Ideen kann erarbeitet, beziehungsweise in entsprechender Literatur (z.B. Little und Rubin [9], Toutenburg [15]) nachgelesen werden.

2.3.2 Eigenschaften und Verhalten der Methoden

Mixed Schätzer

Um das Verhalten der verschiedenen Imputationsmethoden näher zu analysieren bedarf es zunächst einiger grundlegender Betrachtungen. Für den Fall, dass im klassischen linearen Modell Zusatzinformationen vorhanden sind, also beispielsweise der Parametervektor β teilweise bekannt ist oder Einschränkungen vorgegeben sind, so kann dieses Wissen mit Hilfe einer 'Restriktionsbetrachtung' in das Modell mitaufgenommen werden [15]. Eine lineare Restriktion kann wie folgt notiert werden:

$$r \;=\; R\beta\,. \tag{2.8}$$

Wird die Zusatzinformation (bzw. Restriktion) mit Hilfe des Lagrange-Ansatzes in die Minimierung der Fehlerquadratsummen miteinbezogen, so ergibt sich für den restriktiven KQ-Schätzer:

$$\beta(R) \;=\; \beta + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\beta)\,. \tag{2.9}$$

Die restriktive KQ-Schätzung ist also nichts anderes als die Summe aus der normalen KQ-Schätzung β und einem Korrekturglied, das die Erfüllung der exakten Restriktion $r = R\beta$ sichert.

In vielen Modellen in der Praxis kann eine *stochastische* lineare Restriktion postuliert werden:

$$r = R\beta + \phi, \qquad \phi \sim (0, \sigma^2 V) \tag{2.10}$$

Dabei seien $r : J \times 1$, $R : J \times K$ und R und V bekannt.

Theil und Goldberger (1961) begründeten die sogenannte 'mixed estimation technique', deren Grundidee die Mischung der beiden Informationen zu einem gemeinsamen Modell war. Dies bedeutet, dass die Zusammenführung der beiden Modelle $y = X\beta + \epsilon$ und (2.10) als 'Mixed Modell' bezeichnet werden kann:

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \beta + \begin{pmatrix} \epsilon \\ \phi \end{pmatrix} \tag{2.11}$$

Wichtigste Voraussetzung ist dabei, dass die beiden Zufallsgrößen ϕ und ϵ unkorreliert sind:

$$E(\epsilon\phi') = 0$$

Die gemeinsame Kovarianzmatrix kann daher wie folgt notiert werden:

$$E\begin{pmatrix} \epsilon \\ \phi \end{pmatrix}(\epsilon, \phi)' = \sigma^2 \begin{pmatrix} W & 0 \\ 0 & V \end{pmatrix} \tag{2.12}$$

Mit der Bezeichnung

$$\tilde{y} = \begin{pmatrix} y \\ r \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ R \end{pmatrix}, \quad \tilde{\epsilon} = \begin{pmatrix} \epsilon \\ \phi \end{pmatrix}, \quad \tilde{W} = \begin{pmatrix} W & 0 \\ 0 & V \end{pmatrix} \tag{2.13}$$

lässt sich das mixed Modell also wie folgt bezeichnen:

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}, \quad \epsilon \sim (0, \sigma^2\tilde{W}) \,. \tag{2.14}$$

Dieses Modell kann als ein verallgemeinertes lineares Regressionsmodell aufgefasst werden. Es ergibt sich für die beste lineare erwartungstreue Schätzung von β der Schätzer

$$\begin{aligned} \hat{\beta}(R) &= (X'X + R'V^{-1}R)^{-1}(X'W^{-1}y + R'V^{-1}r) \\ &= b + (X'X)^{-1}R'(V + R(X'X)^{-1}R')^{-1}(r - Rb) \end{aligned} \tag{2.15}$$

mit

$$V(\hat{\beta}(R)) = \sigma^2(X'X + R'V^{-1}R)^{-1}. \tag{2.16}$$

Dieser Schätzer wird auch *Mixed-Schätzer* genannt.

KQ-Schätzer für allgemeine Imputation

Im Zuge einer Analyse über fehlende Daten stellt sich zu allererst die Frage an welcher Stelle die fehlenden Werte zu finden sind. Unter dem Gesichtspunkt der Durchführung einer Regressionsanalyse können fehlende Daten sowohl in der Zielgröße, als auch in den Einflussgrößen auftreten. Da auch im Verlauf einiger Simulationsstudien (siehe Kapitel 4) in erster Linie das Augenmerk auf dem Fehlen innerhalb der Einflussgrößen liegen soll, sei hier vorrangig diese Situation beschrieben und erläutert.

Allgemein kann ein Regressionsmodell mit fehlenden Werten wie folgt notiert werden:

$$\begin{pmatrix} y_{obs} \\ y_{mis} \\ y_{obs} \end{pmatrix} = \begin{pmatrix} X_{obs} \\ X_{obs} \\ X_{mis} \end{pmatrix} \beta + \epsilon \,. \tag{2.17}$$

Der Index 'obs' steht dabei für die beobachteten (observed) Werte, der Index 'mis' dagegen für die fehlenden (missing) Werte. Da für den Spezialfall von fehlenden Werten nur in der y-Variable der klassische KQ-Schätzer β die Lösung des Minimierungsproblems der Residuenquadratsummen darstellt (siehe auch Toutenburg [15] für Details), kann man sich auf folgende Problemstellung beschränken:

$$y_{obs} = \begin{pmatrix} X_{obs} \\ X_{mis} \end{pmatrix} \beta + \epsilon \,. \tag{2.18}$$

Formel 2.18 kann alternativ als

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \qquad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim (0, \sigma^2 I) \tag{2.19}$$

formuliert werden. Der Index 'c' steht dabei für die vollständig beobachteten Fälle, der Index '$*$' steht für die fehlenden Fälle. Die Daten seien entsprechend umsortiert worden.

Betrachtet man die Darstellung (2.19) als eine Kombination zweier Submodelle, so kann man diese als Mixed Modell interpretieren (siehe auch Formel 2.11).

Als allgemeiner KQ-Schätzer bei fehlenden Daten in der X-Matrix ergibt sich mit Anwendung von 2.15 der mixed Schätzer

$$\hat{\beta}(X_*) = (X_c' X_c + X_*' X_*)^{-1} (X_c' y_c + X_*' y_*) \tag{2.20}$$

$$= \beta_c + S_c^{-1} X_*' (I + X_* S_c^{-1} X_*')^{-1} (y_* - X_* \beta_c) \,. \tag{2.21}$$

Dabei bezeichnet β_c den normalen KQ-Schätzer für die vollständigen Fälle und S_c steht für den Term $X_c' X_c$.

Verhalten eines Schätzers bei Mittelwert-Imputation

Exemplarisch soll am Beispiel der Mittelwert-Imputation erläutert werden, dass theoretische Grundüberlegungen bezüglich einer Imputationsmethode eigentlich unerlässlich sind.

Wie in Kapitel 2.3.1 angedeutet, wird bei der Mittelwertsimputation jeder fehlende Wert innerhalb einer Variable, also eines Vektors $x_{\cdot j}$ in der Datenmatrix X_*, durch das arithmetische Mittel der beobachteten Fälle ersetzt. Es folgt also:

$$x_{ij,mis} = \bar{x}_{\cdot j,C}$$

Ist das Stichprobenmittel eine gute Schätzung für den Mittelwert der j–ten Spalte, so können bezüglich des Bias mit Mittelwertsimputationen gute Ergebnisse erzielt werden [15]. Falls die Werte der j–ten Spalte Trends oder Nichtlinearitäten wie Wachstumskurven unterliegen, ist $\bar{x}_{\cdot j}$ jedoch keine gute Schätzung, mit Verzerrungen ist zu rechnen. Das Ersetzen aller fehlenden x_{ij} durch die entsprechenden Spaltenmittel führt die Matrix X_* in eine vollständig bekannte Matrix $X_{(1)}$ über. Eine Darstellung im Sinne eines Mixed Modells (siehe Formel 2.11) ist nun möglich:

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_{(1)} \end{pmatrix} \beta + \begin{pmatrix} \epsilon \\ \epsilon_{(1)} \end{pmatrix}. \tag{2.22}$$

Für den Fehlervektor $\epsilon_{(1)}$ gilt

$$\epsilon_{(1)} = (X_* - X_{(1)})\beta + \epsilon_* \tag{2.23}$$

mit

$$\epsilon_{(1)} \sim \{(X_* - X_{(1)})\beta, \sigma^2 I\}. \tag{2.24}$$

Das Ersetzen der fehlenden Werte führt also im Allgemeinen zu einem verzerrten mixed Modell, da im Allgemeinen $X_* - X_{(1)} \neq 0$ gelten wird. Falls X stochastisch ist, kann man günstigenfalls $E(X_* - X_{(1)}) = 0$ erwarten.

Es sei angemerkt, dass die Mittelwertsimputation oft auch als Zero-order Regression (ZOR) bezeichnet wird [15].

3. Maße zur Modellgüte

Bei der Durchführung einer Regression, und der damit verbundenen Entscheidung für ein Modell, gibt es viele Maße, die eine Aussage über Modellgüte und Modellwahl liefern. Im Folgenden sollen die gebräuchlichsten davon in ihrer Konzeption und Idee vorgestellt werden.

3.1 Maße bei vollständigen Daten

3.1.1 Akaikes Informationskriterium

Kullback-Leibler Information

Das Akaike Informationskriterium (AIC) findet seinen Ursprung in der Informationstheorie. Diese wurde im wesentlichen durch Arbeiten von Wiener [17] und Shannon [12] in der Mitte des 20. Jahrhunderts begründet.
Unter verschiedensten Gesichtspunkten und Fragestellungen kann dabei der ursprüngliche Informationsbegriff modifiziert, interpretiert und auch quantifiziert werden. Im Jahr 1951 publizierten S. Kullback und R.A. Leibler ihre Idee von der Quantifizierung des Informationsbegriffes [8].

Als grundlegende Annahme soll eine Funktion 'f' die volle Realität oder Wahrheit widerspiegeln, eine Funktion 'g' dagegen soll für ein Modell stehen, das versucht die Wahrheit so gut wie möglich, jedoch nur approximativ, auszudrücken. Die Kullback Leibler-Information $I(f, g)$ steht dann für die verlorengegangene Information bei der Verwendung des Modells 'g' anstelle von 'f'. Für stetige Funktionen ist diese wie folgt definiert [4]:

$$I(f, g) = \int f(x) \cdot \log\Big(\frac{f(x)}{g(x|\theta)}\Big) dx\,. \tag{3.1}$$

Trivialerweise ist unter einer Menge von Modellen dasjenige das Beste, das in Relation zu den anderen am wenigsten Information bezüglich des wahren Modells verliert. Alternativ kann die Kullback-Leibler Information auch als 'Distanz' zwischen der Realität und einem Modell interpretiert werden.

Im Rahmen der Modellselektion kann das Kriterium $I(f, g)$ natürlich nicht in der ursprünglichen Definition verwendet werden. Hierfür wäre es notwendig sowohl eine umfassende Kenntnis der ganzen Realität zu besitzen, als auch die Parameter θ der Modelle g_i zu kennen. In der Datenanalyse müssen die Modellparameter jedoch geschätzt werden, worin ein großer Unsicherheitsfaktor liegt. Als Grundidee wird deshalb die *erwartete* anstelle der ursprünglichen Kullback-Leibler Information minimiert [4]. Um diesem Sachverhalt Rechnung zu tragen, wird folgende Umformung benötigt:

$$\begin{aligned} I(f,g) &= \int f(x) \cdot \log\Big(\frac{f(x)}{g(x|\theta)}\Big) dx \\ &= \int f(x) \cdot \log(f(x)) dx - \int f(x) \cdot \log(g(x|\theta)) dx \\ &= E_f[\log(f(x))] - E_f[\log(g(x|\theta))]\,. \end{aligned} \tag{3.2}$$

Da der erste Term $E_f[\log(f(x))]$ als Konstante angesehen werden kann, genügt es als Gütekriterium für ein Modell die relativ erwartete Kullback-Leibler-Information $E_f[\log(g(x|\theta))]$ zu schätzen.

Akaike Informationskriterium

Im Jahr 1974 gelang es Akaike [2] zu zeigen, dass es im Rahmen eines Kullback-Leibler basierten Gütekriteriums der wesentliche Punkt war, den Erwartungswert

$$E_y E_x[\log(g(x|\hat{\theta}(y)))]$$

zu schätzen. Der Parameter $\hat{\theta}$ steht dabei für die Maximum-Likelihood Schätzung von θ – basierend auf einem Modell g und Daten y.

Akaike gelang der große Durchbruch, indem er einen formalen Zusammenhang zwischen der Kullback-Leibler Information (also einem informationstheoretischen Konstrukt) und der Likelihood-Theorie (einem großen statistischen Konstrukt) herstellen konnte. Er fand heraus, dass der maximierte Log-likelihood Wert ein verzerrter Schätzer von $E_y E_x[\log(g(x|\hat{\theta}(y)))]$ war. Die Verzerrung entsprach näherungsweise jedoch gerade der Anzahl der zu schätzenden Parameter. Damit ergab sich als Schätzung für die relative erwartete Kullback-Leibler Information:

$$\text{rel } \hat{E}(KL) = \log(L(\hat{\theta}|\text{Daten})) - p\,. \tag{3.3}$$

Der Parameter 'p' steht dabei für die Anzahl der zu schätzenden Parameter. Akaike multiplizierte dieses Ergebnis aus historischen Gründen mit -2 (Information kann auch als negative Entropie aufgefasst werden [11]) und formulierte damit sein Informationskriterium:

$$AIC \quad = \quad -2L(\hat{\theta}) + 2p\,. \tag{3.4}$$

Im Spezialfall einer kleinsten Quadrate-Schätzung mit normalverteiltem Fehlerterm kann der AIC auch wie folgt notiert werden:

$$AIC \quad = \quad n \cdot \ln(SSE) + 2p - n \cdot \ln(n)\,. \tag{3.5}$$

p bezeichnet die Anzahl der zu schätzenden Parameter und kann als Strafterm für die Komplexität eines Modells interpretiert werden. 'SSE' steht hierbei für die Residuenquadratsumme.

3.1.2 Schwarzsches Bayes-Kriterium

Das Schwarzsche Bayes-Kriterium, auch BIC oder SBC genannt, ist ein im Resultat dem AIC ähnliches Gütemaß, entstand jedoch aus einer völlig eigenständigen Idee. Im Gegensatz zu Akaikes Informationskriterium beruht hier der Ansatz nicht auf einer informationstheoretischen Grundlage, sondern auf einer bayesianischen [4].

Sei eine Menge an Modellen $m \,\epsilon\, \mathcal{M}$ und ein Datensatz x gegeben. Jedes Modell habe eine *priori* Wahrscheinlichkeit von $p_m(m)$, die priori Dichten für jedes Modell seien durch $p_l(x|\theta, m)$ und $p_p(\theta|m)$ gegeben. Dabei sei $\theta \,\epsilon\, \Theta_m$, der zu dem Modell m gehörige Parameterraum. Bei einem bayesianischen Ansatz soll durch Herausintegrieren des Parameters θ die Wahrscheinlichkeit von x bei gegebenem Modell m

$$p(x, m) = p_m(m) \int_{\Theta_m} p_l(x|\theta, m) p_p(\theta|m) d\theta \tag{3.6}$$

herausgefunden werden und dasjenige Modell m gewählt werden, das $p(m|x)$ $\propto p(x, m)$ maximiert. Diese Kernidee von Schwarz [13] kann natürlich nicht immer technisch einwandfrei berechnet werden. Deshalb bedarf es einer adäquaten Approximation dieses Ausdrucks.

Es sei:

$$\begin{array}{lcllcl} L(x|\theta, m) & = & \ln p_l(x|\theta, m) & P(\theta|m) & = & \ln p_p(\theta|m) \\ H(x, \theta|m) & = & L(x|\theta, m) + P(\theta|m) & \hat{\theta}(x) & = & \text{argmax}_\theta H(x, \theta|m) \end{array}$$

Dabei bezeichnet $\hat{\theta}(x)$ die MAP-Schätzung (Maximum a posteriori). Die Anzahl der Freiheitsgrade werde durch d bezeichnet.

Der Integrand aus Formel (3.6) kann über die Laplace-Methode, die eine Taylor-Entwicklung um die MAP-Schätzung in Betracht zieht, approximiert werden [14]. Man erhält

$$H(x,\theta|m) \approx H(x,\hat{\theta}|m) - \frac{1}{2}(\theta-\hat{\theta})^T I_H(x:\hat{\theta}|m)(\theta-\hat{\theta}) \tag{3.7}$$

mit

$$I_H(x:\hat{\theta}|m) = \Big[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} H(x,\theta|m) \Big] ,$$

da der zweite Term der mehrdimensionalen Taylor-Entwicklung (also H') 'Null' ergibt. Durch Einsetzen von Formel 3.7 ergibt sich:

$$\begin{aligned} p(x,m) &= p_m(m) \int_\theta \exp[H(x,\theta|m)]d\theta \\ &\approx p_m(m) \cdot \exp[H(x,\hat{\theta}|m)] \\ &\quad \cdot \int \exp[-\frac{1}{2}(\theta-\hat{\theta})^T I_H(x:\hat{\theta}|m)(\theta-\hat{\theta})]dx \\ &= p_m(m) \cdot \exp[H(x,\hat{\theta}|m)] \frac{(2\pi)^{d/2}}{\sqrt{\det\, I_H(x:\hat{\theta}|m)}} . \end{aligned} \tag{3.8}$$

Der letzte Schritt der Umformung erfolgt aus der Tatsache, dass der Integrand als quadratische Form einer Gaußverteilung aufgefasst werden kann.

Logarithmieren auf beiden Seiten führt zu folgendem Ergebnis:

$$\begin{aligned} \ln p(x,m) &\approx \ln p_m(m) + H(x,\hat{\theta}|m) + \frac{d}{2}\ln 2\pi \\ &\quad -\frac{1}{2}\ln\det I_H(x:\hat{\theta}|m) \end{aligned} \tag{3.9}$$

Nimmt man nun an, dass N i.i.d. Samples $x = x_1, x_2, ..., x_N$ vorliegen, so kann die Log-likelihood wie folgt dargestellt werden:

$$L(x|\theta,m) = \sum_{i=1}^{N} L(x_i|\theta,m) .$$

Außerdem sei die empirische Fisher Information

$$I : L(x:\hat{\theta}|m) = \sum_{n=1}^{N} I_L(x_n:\hat{\theta}(x)|m) .$$

Sei nun $\hat{\theta}_{lim} = \lim_{N\to\infty} \hat{\theta}(x)$. Mit dem Gesetz der großen Zahlen kann gezeigt werden, dass

$$\frac{E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1) : \hat{\theta}(x)|m]}{(1/N)I_L(x : \hat{\theta}|m)} \to 1\,.$$

Der letzte Term von Formel (3.9) kann nun über einige grundlegende Eigenschaften der Determinante wie folgt approximiert werden:

$$\begin{aligned} & -\frac{1}{2}\ln\det I_H(x : \hat{\theta}|m) \\ \approx\ & -\frac{1}{2}\ln\det\{N E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1) : \hat{\theta}(x)|m]\} \\ =\ & -\frac{1}{2}\ln\det N I_d - \frac{1}{2}\ln\det E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1) : \hat{\theta}(x)|m] \\ =\ & -\frac{d}{2}\ln N - \frac{1}{2}E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1) : \hat{\theta}(x)|m]\,. \end{aligned}$$

Durch Einsetzen erhält man:

$$\begin{aligned} \ln p(x,m) \approx\ & \ln p_m(m) + \sum_{n=1}^{N} L(x_n|\hat{\theta}(x),m) + P(\hat{\theta}|m) \\ & + \frac{d}{2}\ln 2\pi - \frac{d}{2}\ln N \\ & - \frac{1}{2}\ln\det E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1) : \hat{\theta}(x)|m]\,. \end{aligned} \tag{3.10}$$

Für $N \to \infty$ dominieren die Funktionen von N die anderen Terme, und daher erfolgt die bekannte klassische Approximation:

$$\ln p(x,m) \approx L(x|\hat{\theta},m) - \frac{d}{2}\ln N\,.$$

Durch Multiplikation mit dem 'historischen' Term von '-2', sowie der Notation von d als p (Anzahl der zu zu schätzenden Parameter) erhält man den BIC:

$$BIC = -2L(\hat{\theta}) + p\cdot\ln(n) \tag{3.11}$$

Im Spezialfall einer kleinsten Quadrat-Schätzung mit normalverteiltem Fehlerterm, kann der BIC auch wie folgt notiert werden:

$$BIC = n\cdot ln(SSE) + ln(n)\,p - n\cdot ln(n)\,. \tag{3.12}$$

Tatsächlich unterscheiden sich AIC und BIC also nur in dem Strafterm, was dazu führt, dass das Schwarzsche Bayes-Kriterium kleinere Modelle favorisiert [4]. Der Ansatz zwischen den beiden Gütemaßen ist jedoch grundlegend verschieden.

Vereinzelt sind jedoch auch schon Versuche unternommen worden den AIC bayesianisch zu begründen und interpretieren, den BIC wiederum informationstheoretisch. Die Autoren dieser Werke [4] weisen stets darauf hin, dass eine Entscheidung zugunsten einer der beiden Gütekriterien niemals nur wegen des Ansatzes, sondern allein aufgrund einer spezifischen Problemstellung erfolgen sollte.

3.1.3 Mallows C_p

Ein weiteres Maß zur Beurteilung der Güte eines Modells ist Mallows C_p [10]. Die Grundidee dieses Maßes weicht erneut stark von den beiden anderen ab. Ziel ist es den mittleren quadratischen Vorhersagefehler (MSPE) zu minimieren. In anderen Worten bedeutet dies, dass der Erwartungswert des quadratischen Abstandes vom Mittelwert der abhängigen Variable zu deren Vorhersagewert so gering wie möglich sein sollte. Mallows C_p ist ein unverzerrter Schätzer für den MSPE und kann wie folgt notiert werden:

$$C_p \;=\; \frac{SSE}{\hat{\sigma}^2} + 2\,p - n \tag{3.13}$$

$\hat{\sigma}^2$ ist dabei die Schätzung der Varianz aus dem vollen Modell.

Für alle drei Methoden (AIC, BIC, C_p) gilt, dass beim Vergleich zweier Modelle dasjenige mit dem geringeren Wert des Gütemaßes das Bessere ist.

3.2 Maße bei unvollständigen Daten

3.2.1 Gewichtetes AIC

Die Idee des gewichteten AIC basiert auf dem Horvitz-Thompson Schätzer [9]. Jede Beobachtung wird dabei mit einer inversen 'Auswahlwahrscheinlichkeit' gewichtet. Sei

$$\delta_i = \begin{cases} 1, & \text{wenn die i-te Einheit vollständig beobachtet wurde} \\ 0, & \text{wenn die i-te Einheit nicht vollständig beobachtet wurde} \end{cases}$$

und sei $\pi_i = P(\delta_i = 1)$ die dazugehörige Wahrscheinlichkeit einer vollständigen Beobachtung, dann ergeben sich die Gewichte zu

$$w_i = \frac{\delta_i}{\pi_i}\,. \tag{3.14}$$

Unter der Annahme $Y \sim N(\mu, \sigma^2 I)$ erhält man das gewichtete AIC als

$$AIC_W \;=\; -2\sum_{i=1}^{n} w_i \, log[f(y_i; \mu(x_i; \hat{\theta}_W), \hat{\sigma}^2_W)] + 2(p+1)\,. \tag{3.15}$$

$\hat{\theta}_W$ und $\hat{\sigma}_W$ stehen dabei für die gewichteten ML-Schätzer. Die Funktion 'f' repräsentiert das zu untersuchende Modell. Für ein gewöhnliches lineares Regressionsmodell kann das Kriterium auch wie folgt geschrieben werden:

$$AIC_W = \sum_{i=1}^{n} w_i \cdot log\Big(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i}\Big) + 2(p+1) \tag{3.16}$$

Die e_i stehen dabei für die Residuen der jeweils i-ten Beobachtung.

In der Regel müssen die Gewichte w_i geschätzt werden. Dafür bieten sich viele Möglichkeiten an [6] [16]. Die zwei wichtigsten seien hier kurz erläutert.

Logit-Modell

Im Zuge einer Regressionsanalyse mit fehlenden Daten kann eine binäre Variable beschreiben, ob ein Wert beobachtet wurde oder nicht. Mit Hilfe eines Logit-Modells

$$\ln \frac{\pi}{1-\pi} \;=\; \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \tag{3.17}$$

können Wahrscheinlichkeiten π_i auch geschätzt werden [16]. Somit lassen sich also auch Auftretens- bzw. Fehlwahrscheinlichkeiten, und damit auch die notwendigen Gewichte modellieren.

Das Logit-Modell ist ein generalisiertes lineares Modell (GLM) und daher in gängigen statistischen Programmpaketen gut eingebunden. Der Vorteil des Modells liegt neben seiner Popularität sicher auch in der Einfachheit seiner Parameterinterpretation. Im Rahmen einer guten Beschreibung der Daten und dem Ziel einer möglichst genauen Prädiktion erweisen sich andere Verfahren oft als effizienter. Es empfiehlt sich, ein flexibles generalisiertes additives Modell (GAM) zu benützen.

GAM-Modelle

Ein generalisiertes additives Modell [5] ist ein generalisiertes lineares Model, bei welchem der Prädiktor durch eine Summe von Glättungsfunktionen beschrieben wird:

$$y = \sum_{i=1}^{n} f_i(x_i) \tag{3.18}$$

Durch die entsprechende Link-Funktion kann auch eine binäre Variable - analog zum Logit-Modell - modelliert werden. Für die spezifische Problemstellung der Berechnung der Gewichte des AIC_W, ergibt sich für die Schätzung der Wahrscheinlichkeiten:

$$\ln \frac{\pi}{1-\pi} = \sum_{i=1}^{n} f_i(x_i) \tag{3.19}$$

Wäre tatsächlich jede Art von Glättungsfunktion beim Fitten des Modells erlaubt, so würde die Maximum-Likelihood-Schätzung solcher Modelle ausnahmslos im 'Overfitting' enden. Daher wird für gewöhnlich eine penalisierte Likelihood maximiert um die Schätzungen zu erhalten [5].

Der Vorteil generalisierter additiver Modelle liegt eindeutig in der Prädiktion, die Interpretation der Parameter eines Modells ist dagegen nur sehr schwer möglich.

Ergänzende Bemerkungen zur Berechnung des gewichteten AIC

Zur Berechnung der gewichteten ML-Parameterschätzungen von β ist es nötig die Schätzgleichung

$$\sum_i w_i \cdot (y_i - x_i\beta)x_i = 0 \tag{3.20}$$

zu lösen [1]. Diese entspricht gerade der Schätzgleichung der gewichteten KQ-Schätzung, so dass festzuhalten ist, dass die Parameterschätzungen der beiden Methoden übereinstimmen. Diese Eigenschaft wurde in Kapitel 4 ausgenutzt, um den gewichteten ML-Schätzer $\hat{\beta}_{W,ML}$ zur Berechnung des gewichteten AIC zu bestimmen. Formel 3.16 konnte somit problemlos angewendet werden.

Wird für weiterführende Problemstellungen der gewichtete ML-Schätzer der Varianz benötigt, so kann dieser über die Residuen der gewichteten KQ-Schätzung bestimmt werden:

$$\hat{\sigma}^2_{W,ML} = \frac{\sum_i w_i \epsilon_i}{\sum_i w_i} \tag{3.21}$$

3.2.2 Gewichtetes BIC und C_p

Analog zum gewichteten AIC lassen sich auch ein gewichtetes BIC und C_p definieren [6]. Für den Spezialfall einer linearen Regression lauten diese wie folgt:

$$BIC_W = \sum_{i=1}^{n} w_i \cdot log\Big(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i}\Big) + log\Big(\sum_{i=1}^{n} w_i\Big)p\,. \tag{3.22}$$

$$C_{p_W} = \sum_{i=1}^{n} w_i \cdot \Big(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i e_i^{*2}}\Big) - \Big(\sum_{i=1}^{n} w_i - 2(p-1)\Big)\,. \tag{3.23}$$

Erneut stehen die e_i für die Residuen der jeweils i-ten Beobachtung des entsprechenden Modells, während die e_i^* die entsprechenden Residuen des vollen Modells widerspiegeln.

3.2.3 Gängige Maße bei imputierten Werten

Eine weitere Möglichkeit zur Entscheidungsfindung für ein Modell bietet sich über das Ersetzen der fehlenden Werte wie in Kapitel 2 beschrieben an. Unabhängig von der gewählten Methode können nun die in Kapitel 3.1 vorgestellten Maße auf den komplettierten Datensatz angewandt werden. Eine Entscheidung zugunsten eines Modells ist nun möglich.

4. Simulationsstudien

4.1 Erste Simulation - Zwei Einflussvariablen

In diesem ersten Szenario soll eine einfache Situation veranschaulicht werden, in der eine Variable linear von einer anderen abhängt. Die Einflussgröße soll hierbei jedoch fehlende Werte aufweisen. Eine weitere konstruierte Variable soll keinerlei Einfluss auf die anderen beiden haben.
Zu prüfen ist, wie der Umgang mit den fehlenden Daten sich auf die Modellwahl auswirkt, und ob für eine solche Situation ein gewichtetes Gütemaß oder die Imputation fehlender Werte zu bevorzugen ist.

4.1.1 Grundszenario

Zunächst wurde eine auf dem Intervall $[0, 10]$ gleichverteilte Zufallsvariable x erzeugt, sowie eine Bernoulli(0.5)-verteilte Variable z. Unter gegebenem x und z wurde dann eine normalverteilte Variable y mit Erwartungswert '$\mu_0(x, z) = -4+5x$' und Varianz $\sigma_0^2 = \exp(6)$ generiert. Anschließend wurden mit einer bedingten Wahrscheinlichkeit von

$$\pi(x, z) = \begin{cases} 1 - (1 + 0.015 \cdot y^2)^{-1} & für \quad x \leq 0 \\ 1 - (1 + 0.0005 \cdot y^2)^{-1} & für \quad x > 0 \end{cases}$$

Werte von x als fehlend deklariert. Die Größe des Samples lag bei n=100. Da die fehlenden Werte unter diesen Umständen nicht von x abhängen, kann von MAR ausgegangen werden.

Insgesamt wurden 1000 verschiedene Samples $\{(x_i, y_i, z_i), i = 1, ..., n\}$ für fixe $\{(x_i, z_i), i = 1, ..., n\}$ generiert. Für jedes Sample wurden 4 verschiedene Regressionsmodelle gefittet. Sowohl das Modell $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$, als auch alle Untermodelle davon. Tabelle 4.1 veranschaulicht diese Situation noch einmal.

Nun sollen 13 verschiedene 'Strategien' zur Modellselektion verglichen werden. Als Gütemaß zur Entscheidung für ein Modell soll jeweils der AIC dienen, jedoch immer unter einer anderen Behandlung der fehlenden Daten.

Tab. 4.1: Die vier zur Wahl stehenden Modelle für das Grundszenario der ersten Simulation

	gefittete Modelle
(1)	$y = \beta_0 + \beta_1 x$
(2)	$y = \beta_0 + \beta_1 z$
(3)	$y = \beta_0 + \beta_1 x + \beta_2 z$
(4)	$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$

Als erster Anhaltspunkt soll er sowohl für die Originaldaten (ohne fehlende x-Werte), als auch für alle vorhandenen Datentripel (Complete Cases) berechnet werden.
Des weiteren sollen Strategien für gewichtete AICs und Imputationsmethoden verglichen werden. Gewichtete AICs wurden sowohl für die Originalgewichte als auch für die geschätzten Gewichte bestimmt. Die Schätzungen wurden dabei sowohl über logistische Regression (linearer und quadratischer Einfluss von y), als auch über generalisierte additive Modelle berechnet.
Als Imputationsmethoden wurden 'Mean Imputation', 'Hot deck Imputation' (Ziehen aus der empirischen Verteilung) und 'Regression Imputation' (normal, mit Fehlerterm, Berücksichtigung der Verteilung der Schätzungen der Regressionsparameter und Varianz) herangezogen. Zwei weitere Methoden einer alternativen multiplen Imputation (siehe Kapitel 2.3.1) wurden ebenfalls betrachtet. Zum einen der Mittelwert beim 5-maligen Ziehen aus der empirischen Verteilung, sowie eine Kombination aus fünffacher 'Regression Imputation' und fünffacher 'Hot deck Imputation'.

Im Idealfall würde der jeweilige AIC also für alle 1000 Samples das 'richtige' Modell ($y = \beta_0 + \beta_1 x$) auswählen. Tabelle 4.2 veranschaulicht die Ergebnisse.

Bei Betrachtung der Originaldaten (ohne die fehlenden x-Werte) wählte das AIC 791 mal das richtige Modell aus. Wurden nur die vollständigen Fälle untersucht, so erkannte der AIC 755 mal die Situation als richtig. Dieser Wert sollte als Richtwert dienen um zu sehen, welchen Erfolg eine Gewichtung der Fälle oder die Anwendung von Imputationsmethoden verspricht. Aufgrund der hier noch recht einfach gehaltenen Situation ist der Verlust der 'Complete Case Analysis' eher gering, komplexere Situationen (siehe auch Kapitel 4.2) erwirken erwartungsgemäß etwas schlechtere Resultate.

In Abbildung 4.1 sind die Fehlwahrscheinlichkeiten, sowie Gewichte für die x-Werte dieses ersten Szenarios abgebildet. Gemäß vielen Situationen in der Praxis, führt dabei ein sehr hoher oder sehr niedriger Wert zu einer deutlich

Tab. 4.2: Grundszenario. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 35.15 % der x-Werte.

Methode	Regressionsmodell				korrekt
	x	z	x,z	x,z,xz	klassifiziert
AIC Originaldaten	791	0	124	85	791
AIC Complete Cases	755	0	123	122	755
AIC wahre Gewichte	376	0	255	369	376
AIC geschätzte Gewichte 1	596	0	190	214	596
AIC geschätzte Gewichte 2	238	1	195	566	238
AIC geschätzte Gewichte 3	456	0	242	302	456
AIC Imputation Mittelwert	858	0	118	24	858
AIC Imputation Hot deck	802	32	78	88	802
AIC Imputation Regression 1	753	0	108	139	753
AIC Imputation Regression 2	795	0	132	73	795
AIC Imputation Regression 3	799	0	121	80	799
AIC Imputation Regression 4	804	0	119	77	804
AIC Multiple Imputation 1	861	6	106	27	861
AIC Multiple Imputation 2	796	0	124	80	796

größeren Fehlwahrscheinlichkeit. Die Funktion ist jedoch asymmetrisch und verläuft linksseitig der Null deutlich steiler. Wird mit diesen wahren Gewichten gearbeitet, und der daraus resultierende gewichtete AIC berechnet, so wählt dieser nur 376 mal das 'beste' Modell aus. Eine Verbesserung innerhalb dieses sehr einfachen Szenarios gegenüber der Complete Case Analysis ist nicht zu erkennen.
In der Regel stehen die 'wahren' Gewichte jedoch nicht zur Verfügung, so dass diese geschätzt werden müssen. Drei verschiedene Ansätze wurden in dieser Simulation getestet:

1. Schätzung über folgendes Logit-Modell: $logit\,\pi = \alpha_1 + \alpha_2 y$
2. Schätzung über folgendes Logit-Modell: $logit\,\pi = \alpha_1 + \alpha_2 y + \alpha_2 y^2$
3. Schätzung über ein generalisiertes additives Modell: $logit\,\pi = f(y)$

Der Reihenfolge entsprechend wurden die Methoden in Tabelle 4.2 mit 'AIC geschätzte Gewichte 1', 'AIC geschätzte Gewichte 2' und 'AIC geschätzte

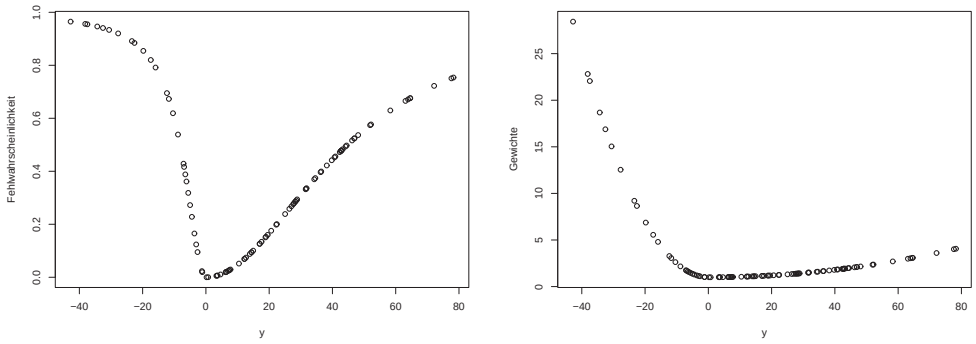

Abb. 4.1: Fehlwahrscheinlichkeit (links) und Gewichte (rechts) für das Grundszenario bei einem willkürlich ausgewählten Datentripel.

Gewichte 3' bezeichnet. Bei der ersten Methode wird man intuitiv ein sehr schlechtes Ergebnis erwarten, da ohne einen quadratischen Einflussterm das Verhalten der Fehlwahrscheinlichkeiten nicht ausreichend approximiert werden kann. Tatsächlich liefert der gewichtete AIC hier jedoch 596 mal das richtige Modell, deutlich mehr als bei der Rechnung mit den wahren Gewichten. Abbildung 4.2 liefert dafür eine Erklärung.

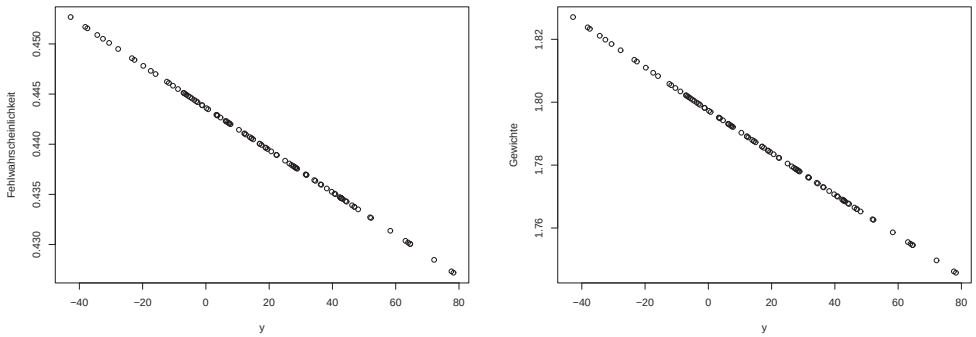

Abb. 4.2: Die über ein einfaches Logitmodell geschätzte Fehlwahrscheinlichkeit (links), sowie die Gewichte (rechts) für ein willkürliches Datentripel des Grundszenarios.

Die durch das Modell prognostizierten Wahrscheinlichkeiten, und damit auch die daraus resultierenden Gewichte, liegen für das hier gewählte Datentripel nah beieinander. Jeder Fall wird fast gleich stark gewichtet, daher ist eine

Annäherung an die 'Complete Case Analysis' zu beobachten. Für andere Datentripel wurden in dieser Form ähnliche, wenn auch nicht ganz so stark ausgeprägte, Resultate beobachtet.

Werden die Wahrscheinlichkeiten durch ein Logit-Modell mit quadratischer Einflussgröße geschätzt, so ist eine deutlich bessere Anpassung an die Originalwahrscheinlichkeiten zu beobachten (siehe Abbildung 4.3).

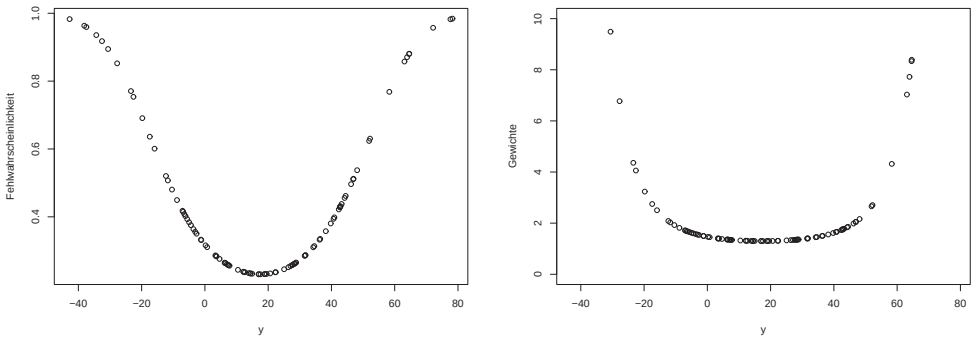

Abb. 4.3: Die über ein Logitmodell mit quadratischem Einflussterm geschätzten Fehlwahrscheinlichkeiten (links), sowie die Gewichte (rechts) für ein willkürliches Datentripel des Grundszenarios.

Trotz dieser scheinbar noch eher guten Schätzung wählt der AIC hier nur 238 mal das 'beste' Modell aus. Dies hat im wesentlichen 2 Gründe: Zum einen werden die extrem großen Werte - vor allem im Verhältnis zu den extrem kleinen Werte - zu hoch gewichtet. Zum anderen wird, bei vielen Datentripeln, eine Symmetrie um den Wert '20' konstatiert, was nicht den wahren Sachverhalt widerspiegelt.
Eine Verbesserung lässt sich über die Schätzung der Wahrscheinlichkeiten durch ein generalisiertes additives Modell finden (siehe auch Abbildung 4.4).

Sowohl Wahrscheinlichkeiten, als auch Gewichte werden hier am besten geschätzt. Mit 456 richtig erkannten Fällen liegt der Wert sogar über dem der wahren Gewichte, dennoch ist auch er deutlich unter dem Ergebnis der 'Complete Cases Analysis' anzusiedeln.

Im Allgemeinen scheinen die gewichteten Methoden in diesem Szenario keine Verbesserung darzustellen. Ein Grund hierfür ist jedoch auch die mit exp(6) eher hoch gewählte Varianz. Bei geringerer Varianz bringt auch der gewichtete AIC deutlich bessere und stabilere Ergebnisse (siehe dazu auch Kapitel

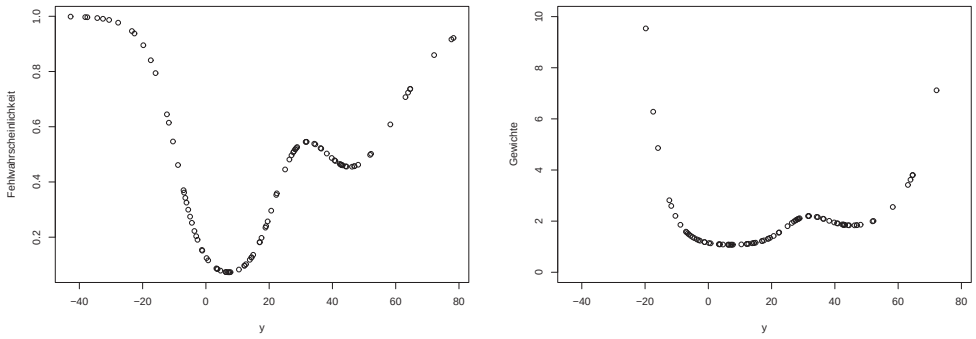

Abb. 4.4: Die über ein GAM geschätzten Fehlwahrscheinlichkeiten (links), sowie die Gewichte (rechts) für ein willkürliches Datentripel des Grundszenarios.

4.1.4). Auffällig ist auch, dass das komplexeste Modell, welches auch die Interaktion der beiden Einflussgrößen berücksichtigt, von diesen Gütemaßen sehr häufig ausgewählt wurde. Ob ein Trend in der Wahl größerer, komplexerer Modelle festzustellen ist, sollte im Verlauf dieser Arbeit noch zu klären sein.

Ein Vergleich unter den Imputationsmethoden erbringt interessante Ergebnisse. Werden die fehlenden Werte durch den Mittelwert ersetzt, so wird das 'richtige' Modell 858 mal ausgewählt. Der Grund für diesen sehr guten Wert liegt in erster Linie bei der geringen Komplexität des Modells. Für mehrere Einflussgrößen, ein komplexeres Modell oder ein höherer Anteil an fehlenden Werten (siehe dazu auch Kapitel 4.1.4) sollten weniger gute Ergebnisse erwartet werden. Auch wenn die wahren Fehlwahrscheinlichkeiten der x-Werte nicht einer kompletten Symmetrie folgen, so unterstützt der zumindestens annähernd symmetrische Verlauf die Imputation durch den Mittelwert. Für diese Datensituation erweist sich die 'Mean Imputation' als eine sehr brauchbare Methode.

Im Zuge der 'Hot deck Imputation' (also einem einmaligen Ziehen aus der empirischen Verteilung) kann zwar mit 802 'richtigen' Ergebnissen ein brauchbares Resultat erzielt werden, als Wermutstropfen erweist sich jedoch das 32-malige Auswählen des Modells mit 'z' als einziger Einflussgröße. Grund dafür ist die hier recht hohe Varianz von exp(6). Eine geringere Varianz führt zu weniger fatalen Ergebnissen (siehe auch Kapitel 4.1.4).

Weitere interessante Ergebnisse finden sich bei verschiedenen Ansätzen der

'Regression Imputation'. Mit Hilfe der vollständigen Daten wurde eine Regression von y auf x durchgeführt:

$$X_C = \gamma_0 + \gamma_1 Y_C + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Mit den daraus geschätzten Parametern und den entsprechenden y-Werten wurden die x-Werte für die fehlenden Daten berechnet:

$$X_{mis} = \hat{\gamma}_0 + \hat{\gamma}_1 Y$$

Die in Tabelle 4.2 mit 'AIC Imputation Regression 1' bezeichneten Werte zeigen, dass mit 753 'richtig' gewählten Modellen auch diese Imputationsmethode respektable Ergebnisse aufweisen kann. Der Wert liegt jedoch noch immer unter dem der 'Complete Case Analysis'.
Wird nun zu der Hilfsregression noch ein Residuum zur Berücksichtigung der Unsicherheit der vorhergesagten Werte eingebaut (unter 'AIC Imputation Regression 2' in Tabelle 4.2), also folgendes Modell zugrundegelegt,

$$X_{mis} = \hat{\gamma}_0 + \hat{\gamma}_1 Y + \epsilon, \qquad \epsilon \sim N(0, \hat{\sigma}^2)$$

so lässt sich noch eine leichte Verbesserung feststellen. 795 mal wurde das Modell richtig vorhergesagt.

Unter der Berücksichtigung der Verteilung von γ (siehe auch Kapitel 2.3.1), also

$$(\tilde{\gamma}_0, \tilde{\gamma}_1) \sim N((\hat{\gamma}_0, \hat{\gamma}_1), \hat{\sigma}^2 (Z'Z)^{-1})$$

kann eine weitere Verbesserung erzielt werden. Immerhin 799 mal fiel die Wahl auf das beste Modell (siehe auch 'AIC Imputation Regression 3' in Tabelle 4.2).
Mit 'AIC Imputation Regression 4' wird die Methode bezeichnet, die neben den bereits oben beschriebenen Sachverhalten auch noch die Unsicherheit der Schätzung der Varianz berücksichtigt (siehe Formel 2.6). Hier fiel 804 mal die Wahl auf das richtige Modell.

Sehr gute Ergebnisse lassen sich auch über die alternativen multiplen Imputationen erreichen. Die in Tabelle 4.2 mit 'AIC Multiple Imputation 1' bezeichnete Methode zieht dabei 5 mal aus der empirischen Verteilung und ersetzt die fehlenden Werte durch die entsprechenden Mittelwerte der 5 Ziehungen. Im Vergleich zu der einfachen 'Hot deck Imputation' wird nun eine deutliche Verbesserung erzielt (861 'richtige' Ergebnisse). Auch das Modell, das nur die Variable z als Einflussgröße wählt, wird deutlich weniger oft gewählt.

Eine Kombination aus 5 Ziehungen aus der empirischen Verteilung, sowie 5 Werten die über eine Hilfsregression gewonnen wurden ('AIC Multiple Imputation 2') liefert respektable Ergebnisse. 796 mal wird das 'richtige' Modell gewählt und somit eine Verbesserung gegenüber der 'Complete Case Analysis' erzielt.

Für eine einfache Datensituation wie in diesem Beispiel hat sich gezeigt, dass Imputationsmethoden deutlich bessere Ergebnisse erzielen, als das Einsetzen eines gewichteten Gütemaßes.

4.1.2 Andere Gütemaße

Im Folgenden soll nun untersucht werden, ob die Betrachtung anderer Gütemaße ähnliche Ergebnisse liefert oder nicht. So sind in Tabelle 4.3 die Resultate von AIC, BIC und C_p einander gegenübergestellt. Zu erkennen ist ein

Tab. 4.3: Grundszenario. Die Zahlen vermitteln wie oft das 'korrekte' Modell für die entsprechenden Gütemaße ausgewählt wurde.

Methode	Gütemaß		
	AIC	BIC	C_p
Originaldaten	791	965	790
Complete Cases	755	942	753
wahre Gewichte	376	646	372
geschätzte Gewichte 1	596	856	589
geschätzte Gewichte 2	238	440	233
geschätzte Gewichte 3	456	718	450
Imputation Mittelwert	858	975	856
Imputation Hot deck	802	932	797
Imputation Regression 1	753	952	751
Imputation Regression 2	795	967	795
Imputation Regression 3	799	962	799
Imputation Regression 4	804	964	803
Multiple Imputation 1	861	969	857
Multiple Imputation 2	796	965	795

im Grunde ähnliches Schema unter allen Gütemaßen. Die 'Mean Imputation', sowie die 'Multiple Imputation 1' erzielen durchweg die besten Ergebnisse.

Eine grundsätzliche Entscheidung bezüglich der Wahl der Methode scheint unabhängig von der Wahl des Maßes für dieses Szenario zu sein.

Bemerkenswert ist jedoch die Tatsache, dass der 'BIC' bei allen Methoden deutlich öfter das 'korrekte' Modell wählt, und somit eigentlich zu bevorzugen wäre. Erklärt werden kann dies durch den Strafterm, der dazu tendiert einfacheren Modellen den Vorzug zu geben (siehe Kapitel 3.1.2). Deshalb wird - im Vergleich zu AIC und C_p - das Modell mit nur einer Einflussgröße x öfter gewählt, als die Modelle mit mehreren Einflussgrößen.

4.1.3 Variation der Fehlwahrscheinlichkeit

Im Grundszenario wurde angenommen, dass ein sehr hoher bzw. ein sehr geringer Wert von y die Fehlwahrscheinlichkeit für x erhöht. Als Fragestellung bietet es sich nun an zu untersuchen, inwieweit die Wahl der Fehlwahrscheinlichkeitsfunktion die Ergebnisse beeinflusst. Aus diesem Grund soll nun dasselbe Szenario, jedoch mit einer anderen Fehlwahrscheinlichkeit, durchgespielt werden. Hierzu wurde folgende Funktion gewählt:

$$\pi(x, z) = 1 - [1 + \exp(1 - 0.09 \cdot (y + 5))]^{-1}$$

In Abbildung 4.5 sind Fehlwahrscheinlichkeit und Gewichte für ein beliebiges Datentripel abgebildet.

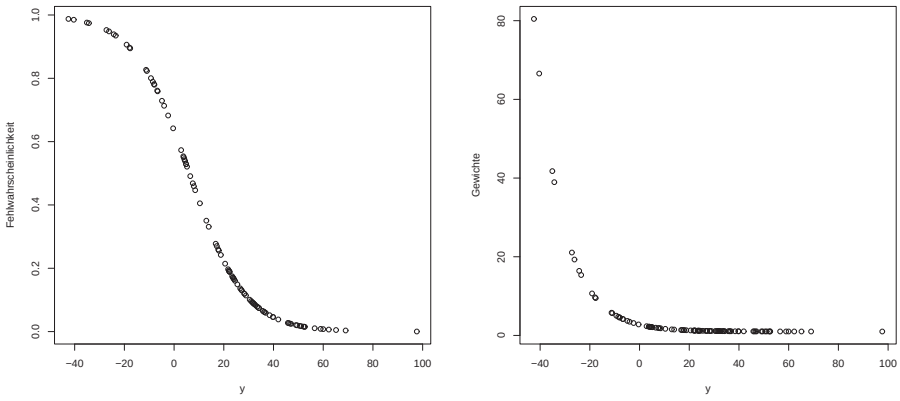

Abb. 4.5: Fehlwahrscheinlichkeit (links) und Gewichte (rechts) bei einem willkürlich ausgewählten Datentripel.

Man erkennt, dass sehr geringe Werte von y mit sehr hoher Wahrscheinlichkeit zu einem fehlenden x-Wert führen. Im Vergleich zum Grundszenario ist kein annähernd symmetrischer Verlauf zu erkennen. In Tabelle 4.4 sind die Ergebnisse des neuen Szenarios abgebildet.

Tab. 4.4: Variation des Grundszenarios. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 32.33 % der x-Werte.

Methode	Regressionsmodell				korrekt
	x	z	x,z	x,z,xz	klassifiziert
AIC Originaldaten	791	0	117	92	791
AIC Complete Cases	768	0	130	102	768
AIC wahre Gewichte	435	2	216	347	435
AIC geschätzte Gewichte 1	419	2	211	368	419
AIC geschätzte Gewichte 2	449	1	226	324	449
AIC geschätzte Gewichte 3	436	2	218	344	436
AIC Imputation Mittelwert	846	0	109	45	846
AIC Imputation Hot deck	838	6	89	67	838
AIC Imputation Regression 1	749	0	115	136	749
AIC Imputation Regression 2	799	0	113	88	799
AIC Imputation Regression 3	790	0	125	85	790
AIC Imputation Regression 4	805	0	107	88	805
AIC Multiple Imputation 1	850	2	105	43	850
AIC Multiple Imputation 2	791	0	108	101	791

Deutlich zu erkennen ist, dass die gewichteten Methoden zumindestens etwas bessere Ergebnisse liefern als im Grundszenario, jedoch immer noch deutlich schlechter abschneiden als die Imputationsmethoden. Auch liegen ihre Werte deutlich unter dem der 'Complete Case Analysis'. Um so stärker macht sich hier jedoch bemerkbar, dass die Wahl der Schätzung der Gewichte nur geringen Einfluss auf die Resultate liefert.

Des weiteren ist zu bemerken, dass ein Unterschied zwischen einer einfachen und einer 'Multiplen Hot deck Imputation' kaum noch auszumachen ist. Beide Methoden liefern ähnliche Ergebnisse.

Die besten Resultate liefern weiterhin die Mittelwertsimputation, sowie die 'Multiple Imputation 1'.

4.1.4 Variation der Varianz

Untersucht werden soll nun, inwieweit sich eine Variation der Varianz auf die verschiedenen Methoden zur Modellselektion auswirkt. Neben dem Grundszenario, bei dem die Werte von y normalverteilt mit $\sigma = \exp(3)$ waren, sollen hier nun 4 weitere Simulationen mit Werten von exp(0), exp(1), exp(2) und exp(3.5) die Fragestellung klären.

Interessanterweise beeinflusst die Wahl von Sigma ja auch die Anzahl der fehlenden Werte, da eine höhere Varianz höhere und geringere y-Werte produziert und damit auch eine höhere Fehlwahrscheinlichkeit. Dieser bei der Interpretation zu berücksichtigende Sachverhalt ist in Abbildung 4.6 noch einmal aufgeführt. Einer durchschnittlichen Fehlwahrscheinlichkeit von 18.38% der Werte bei einem Sigma von exp(0) stehen immerhin 45.94 % fehlende Werte bei Sigma = exp(3.5) gegenüber. In Tabelle 4.5 sind der Ergebnisse

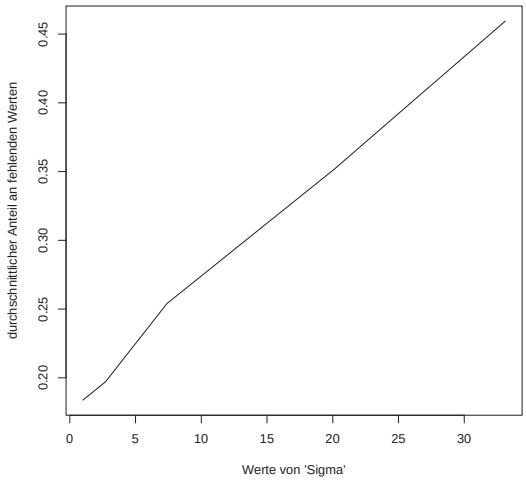

Abb. 4.6: Der von der Varianz abhängige Anteil an fehlenden Werten.

der Simulationen zu erkennen.

Zuallererst fällt auf, dass alle Ansätze mit den gewichteten AICs für eine geringe Varianz, und damit auch für einen geringen Anteil an fehlenden Werten, zwar keine überwältigenden, dafür nun aber etwas bessere Ergebnisse liefern. Der Abstand bei den richtig erkannten Modellen ist im Vergleich zu den Imputationsmethoden hier im Verhältnis deutlich geringer als bei sehr hoher Varianz.

Tab. 4.5: Variation der Varianz. Die Werte geben an, wie oft eine Methode das 'korrekte' Modell bei entsprechender Varianz gewählt hat.

Methode	Wert von 'Sigma'				
	exp(0)	exp(1)	exp(2)	exp(3)	exp(3.5)
AIC Originaldaten	765	771	776	791	770
AIC Complete Cases	766	755	774	755	691
AIC wahre Gewichte	684	661	579	376	274
AIC geschätzte Gewichte 1	673	668	640	596	502
AIC geschätzte Gewichte 2	669	656	543	238	129
AIC geschätzte Gewichte 3	677	676	632	456	345
AIC Imputation Mittelwert	786	869	863	858	489
AIC Imputation Hot deck	750	835	829	802	397
AIC Imputation Regression 1	770	775	780	753	703
AIC Imputation Regression 2	753	775	781	795	722
AIC Imputation Regression 3	771	775	789	799	729
AIC Imputation Regression 4	771	761	796	804	736
AIC Multiple Imputation 1	770	859	870	861	447
AIC Multiple Imputation 2	770	798	816	796	730

Interessant sind auch die Ergebnisse der 'Mean Imputation'. Bei sehr hoher Varianz ist der Erfolg dieser Methode bei weitem nicht so überwältigend wie bei mittlerer oder geringer Varianz. Dennoch scheint im Allgemeinen die Imputation des Mittelwertes eine sehr gute Lösung innerhalb dieses Szenarios darzustellen.

Die mit Abstand stabilsten Ergebnisse bieten jedoch die Imputationsmethoden auf Basis einer Hilfsregression. Unabhängig von der gewählten Varianz, und damit auch unabhängig vom Anteil an fehlenden Werten, versprechen diese Methoden sehr gute Ergebnisse. Bei der sehr hohen Varianz werden sogar die am deutlich besten Ergebnisse erzielt. Die multiple Imputationsmethode bei der Regressionsergebnisse das Ergebnis mitbeeinflussen, bietet dementsprechend auch akzeptable Ergebnisse.

Zu erwähnen ist noch, dass bei geringerer Varianz als exp(3) alle Methoden nahezu ausnahmslos darauf verzichten, das Modell mit 'z' als einziger Einflussgröße zu wählen. Dies kommt speziell der 'Single Hot deck Imputation' zu Gute, die für höhere Varianz sehr anfällig für dieses Modell war.

Insgesamt scheinen sowohl die Varianz, als auch der damit verbundene Anteil an fehlenden Werten, einen Einfluss auf die Wahl der Methode zu haben. Für das hier beschriebene, sehr einfach gehaltene Modell liefern die Imputationsmethoden auf Basis einer Regression die besten Ergebnisse. Ob für komplexere Zusammenhänge ähnlich gute Ergebnisse erzielt werden können, bleibt jedoch noch zu klären.

4.1.5 Variation der z-Variable

Mit der scheinbar eher willkürlichen Wahl einer 0,1-Variable für z stellt sich die Frage, ob bei einer Variation derselben ähnliche Ergebnisse zu erwarten sind, oder ob die bisherigen Resultate allein auf die Wahl als Bernoulli-Variable zurückzuführen sind.

Um zu überprüfen inwieweit eine solche Vermutung bestätigt oder widerlegt werden kann, soll das Grundszenario noch einmal simuliert werden, jedoch bei einer Variation der z-Variablen. Zuerst wurde die Situation mit einer Bernoulli(0.2) bzw. Bernoulli(0.05) Variable durchgespielt, um zu sehen inwiefern dies eine Veränderung erwirken kann. Für eine komplett neue Situation sorgt zum einen eine normalverteilte Einflussvariable ($\mu = 5$, $\sigma = 5$), und zum anderen eine exponentialverteilte ($\lambda = 3$). In Tabelle 4.6 sind die Ergebnisse der Simulationen abgebildet.

Im Wesentlichen ergeben sich ähnliche Resultate wie im Grundszenario. Eine 'Multiple Hot deck Imputation' oder ein Ersetzen durch den Mittelwert garantiert die besten Ergebnisse. Die Wahl der z-Variablen scheint generelle Aussagen nicht zu verzerren.

Bemerkenswert ist jedoch, dass bei der Extremsituation einer Bernoulli(0.05)-Variable die gewichteten Gütemaße etwas bessere Ergebnisse liefern.

4.1.6 Variation der fehlenden Werte

In der bisherigen Betrachtung wurden immer fehlende Werte für die tatsächliche Einflussvariable deklariert. Es stellt sich nun die Frage, inwiefern das Fehlen von Werten in einer anderen Variable die Ergebnisse beeinflussen kann. Dazu wurde erneut das Grundszenario durchgespielt, jedoch nun mit fehlenden Werten in der z-Variable anstelle der x-Variable.

Bei der Methodik mussten Details verändert werden, da die fehlenden Werte nun nicht mehr als stetig angenommen werden können, und nur eine Impu-

Tab. 4.6: Variation der z-Variable. Die Werte geben an, wie oft eine Methode das 'korrekte' Modell in der entsprechenden Situation gewählt hat.

Methode	Verteilung von z				
	Bern. 0.05	Bern. 0.2	Bern. 0.5	Exp. $\lambda = 3$	Normal μ, $\sigma = 5$
AIC Originaldaten	754	777	791	774	776
AIC Complete Cases	798	764	755	762	774
AIC wahre Gewichte	683	449	376	453	447
AIC geschätzte Gewichte 1	681	620	596	610	619
AIC geschätzte Gewichte 2	634	344	238	350	355
AIC geschätzte Gewichte 3	695	509	456	524	520
AIC Imputation Mittelwert	821	874	858	786	840
AIC Imputation Hot deck	757	813	802	687	763
AIC Imputation Regression 1	705	752	753	741	729
AIC Imputation Regression 2	755	794	795	786	777
AIC Imputation Regression 3	754	794	799	778	773
AIC Imputation Regression 4	750	795	804	804	779
AIC Multiple Imputation 1	812	857	818	759	826
AIC Multiple Imputation 2	740	792	787	791	780

tation von 'Nullen' und 'Einsen' möglich ist. So wurde bei der Mittelwertsimputation das Ergebnis als Wahrscheinlichkeit interpretiert und mit dieser aus den Werten 'Null' und 'Eins' gezogen. Auch wenn sich für die einfache 'Hot deck Imputation' keine Veränderung ergab, so musste doch für das multiple Vorgehen der Wert ebenfalls als Wahrscheinlichkeit aufgefasst werden. Analog wurde bei den Regressionsansätzen ein Logit-Modell gefittet und mit Hilfe der daraus resultierenden Ergebnisse erneut aus einer 0,1-Verteilung gezogen. Die Ergebnisse dieses Szenarios sind in Tabelle 4.7 aufgeführt.

Bei der Interpretation der Ergebnisse ist zu beachten, dass im wesentlichen zwei Neuerungen gegenüber dem Grundszenario zu verzeichnen sind. Zum einen sind die fehlenden Werte nicht mehr in der Einflussvariable zu finden, zum anderen werden nun Werte für eine nicht-stetige Variable geschätzt.

Bei Betrachtung der Ergebnisse fällt zuerst auf, dass nur die Mittelwertsimputation, sowie multiple Imputationen Verbesserungen bezüglich der 'Complete Case Analysis' erbringen. Weniger gut schneiden hier die Regressionsmethoden ab. Trivialerweise ändert sich an den Ergebnissen der gewichteten Gütemaße wenig.

Tab. 4.7: Szenario mit fehlenden Werten in der z-Variablen. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 34.97 % der z-Werte.

Methode	Regressionsmodell				korrekt
	x	z	x,z	x,z,xz	klassifiziert
AIC Originaldaten	758	0	137	105	758
AIC Complete Cases	767	0	136	97	767
AIC wahre Gewichte	393	0	244	363	393
AIC geschätzte Gewichte 1	603	0	204	193	603
AIC geschätzte Gewichte 2	240	4	219	537	240
AIC geschätzte Gewichte 3	478	0	253	269	478
AIC Imputation Mittelwert	775	0	153	72	775
AIC Imputation Hot deck	761	0	149	90	761
AIC Imputation Regression 1	579	0	333	88	579
AIC Imputation Regression 2	615	0	293	92	615
AIC Imputation Regression 3	606	0	294	100	606
AIC Imputation Regression 4	616	0	286	98	616
AIC Multiple Imputation 1	781	0	155	64	781
AIC Multiple Imputation 2	776	0	148	76	776

Wie oben bereits erwähnt, kann es zwei mögliche Ursachen für das Zustandekommen der Ergebnisse geben. Weitere, hier jedoch nicht im Detail aufgelistete, Untersuchungen bestätigen jedoch, dass die Änderungen tatsächlich allein auf das Vertauschen der fehlenden Werte von der x-Variable auf die z-Variable zurückzuführen sind. Die nun größere Stochastizität durch das Schätzen der Fehlenden 0,1-Werte hat keinen bemerkenswerten Einfluss.

Erneut scheinen also die 'Mean Imputation', sowie die multiple 'Hot deck Imputation' die stabilsten und besten Ergebnisse zu liefern.

4.1.7 Korrelation unter den möglichen Einflussgrößen

Als weitere Fragestellung bietet es sich an zu untersuchen, inwiefern Korrelationsstrukturen unter den Einflussgrößen sich auf die Resultate auswirken können. Hierzu wurde erneut das Grundszenario durchgespielt, jedoch mit einer Varianz von $\sigma^2 = \exp(5)$. Als weitere Veränderung wurden die z-Werte über die x-Werte generiert (Normalverteilte Zufallsgröße mit '$\mu = 5x+2$' und

'$\sigma = \exp(2.5)$'). Es ergab sich eine Korrelation von 0.70. Die Ergebnisse des Szenarios sind in Tabelle 4.8 abgebildet.

Tab. 4.8: Szenario für eine hohe Korrelation unter den Einflussgrößen. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 34.31 % der x-Werte. Die Korrelation zwischen x und z betrug 0.70.

Methode	Regressionsmodell			
	x	z	x,z	x,z,xz
AIC Originaldaten	759	0	143	98
AIC Complete Cases	750	30	110	110
AIC wahre Gewichte	420	28	208	344
AIC geschätzte Gewichte 1	599	22	165	214
AIC geschätzte Gewichte 2	325	25	162	488
AIC geschätzte Gewichte 3	492	28	203	277
AIC Imputation Mittelwert	41	683	221	55
AIC Imputation Hot deck	16	688	211	85
AIC Imputation Regression 1	696	7	172	125
AIC Imputation Regression 2	614	8	261	117
AIC Imputation Regression 3	630	10	248	112
AIC Imputation Regression 4	615	13	267	105
AIC Multiple Imputation 1	33	692	216	59
AIC Multiple Imputation 2	461	53	383	103

Sofort fällt auf, dass die bisher stabilen und guten Methoden der 'Hot deck Imputation' und Mittelwertsimputation sehr schlechte Ergebnisse liefern. Selbst unter der Annahme, dass das Modell mit Interaktion als richtig angesehen werden kann, wird nur in verheerend wenig Fällen ein sinnvolles Modell gefunden. Beim multiplen Ziehen aus der empirischen Verteilung wird sogar 692 Mal das schlechteste Modell mit z als einziger Einflussgröße ausgewählt, auf das Modell mit 'x' fiel die Wahl nur 33 mal.

Alle anderen Methoden bieten den Umständen entsprechend immer noch sehr respektable Ergebnisse. Speziell die Regressionsansätze versprechen gute Resultate. Eine Verbesserung gegenüber der 'Complete Case Analysis' kann jedoch nie erreicht werden.

4.1.8 Resultate

Im Allgemeinen bleibt festzuhalten, dass für dieses Szenario die Imputationsmethoden den gewichteten Kriterien vorzuziehen sind. Unter fast allen Variationen des Grundszenarios erweisen sich die Imputationsmethoden als stabil und zuverlässig. Oft können Gewinne gegenüber der 'Complete Case Analysis' verbucht werden. Speziell die 'Mean Imputation' und die 'Multiple Hot deck Imputation' liefern sehr gute Ergebnisse.
Eine Ausnahme stellt jedoch der Fall hoher Korrelationsstrukturen unter den Einflussgrößen dar. Hierfür eignet es sich besonders auf Basis einer Hilfsregression Werte zu imputieren.

4.2 Zweite Simulation - Drei Einflussvariablen

In diesem zweiten Szenario soll nun geklärt werden, inwieweit sich eine weitere Einflussgröße auf die bisherigen Ergebnisse auswirken kann. Es stellt sich die Frage, ob eine höhere Anzahl an Einflussvariablen oder fehlenden Werten die Entscheidung für ein Modell beeinflusst.

4.2.1 Grundszenario

Erneut wurde eine auf dem Intervall [0,10] gleichverteilte Zufallsvariable x, sowie eine Bernoulli(0.5)-verteilte Variable z erzeugt. Des weiteren wurde eine neue exponentialverteilte Einflussgröße v ($\lambda = 3$) generiert. Unter gegebenem x,z und v wurde schließlich eine normalverteilte Variable y mit Erwartungswert '$\mu_0(x, z, v) = -4 + 5x + 3z$' und einer moderateren Varianz von exp(4) erzeugt. Von diesen drei Einflussgrößen haben nun also zwei (nämlich x und z) einen Einfluss auf die Werte von y, eine jedoch nicht (nämlich v). Die Funktion der Fehlwahrscheinlichkeit

$$\pi(x, z) = \begin{cases} 1 - (1 + 0.015 \cdot y^2)^{-1} & für \quad x \leq 0 \\ 1 - (1 + 0.0005 \cdot y^2)^{-1} & für \quad x > 0 \end{cases}$$

deklariert - wie im Grundszenario der ersten Simulation - Werte von x als fehlend. Erneut lag die Sample-Größe bei $n = 100$.

Insgesamt wurden 1000 verschiedene Samples $\{(v_i, x_i, y_i, z_i), i = 1, ..., n\}$ für fixe $\{(v_i, x_i, z_i), i = 1, ..., n\}$ generiert. Für jedes Sample wurden 18 verschiedene Regressionsmodelle gefittet. Sowohl das Modell mit allen 3 Einflussgrößen und Interaktionen, als auch alle Untermodelle davon. Tabelle 4.9 veranschaulicht diese Situation noch einmal.

Tab. 4.9: Die 18 zur Wahl stehenden Modelle für das Grundszenario der zweiten Simulation

	gefittete Modelle		
(1)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xz + \beta_5 zv + \beta_6 xz + \beta_7 xzv$		
(2)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xz + \beta_5 zv + \beta_6 xz$		
(3)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xz + \beta_5 xv$		
(4)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xz + \beta_5 zv$		
(5)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xv + \beta_5 zx$		
(6)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 zv$		
(7)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xz$		
(8)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z + \beta_4 xv$		
(9)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 z$	(10)	$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$
(11)	$y = \beta_0 + \beta_1 v + \beta_2 x + \beta_3 xv$	(12)	$y = \beta_0 + \beta_1 v + \beta_2 z + \beta_3 vz$
(13)	$y = \beta_0 + \beta_1 x + \beta_2 z$	(14)	$y = \beta_0 + \beta_1 x + \beta_2 v$
(15)	$y = \beta_0 + \beta_1 v + \beta_2 z$	(16)	$y = \beta_0 + \beta_1 z$
(17)	$y = \beta_0 + \beta_1 v$	(18)	$y = \beta_0 + \beta_1 x$

Nun sollen 8 verschiedene 'Strategien' zur Modellselektion verglichen werden. Neben der 'Complete Case Analysis' wurden erneut gewichtete Gütemaße zur Entscheidungsfindung herangezogen. Neben den wahren Gewichten wurden über generalisierte additive Modelle Gewichte geschätzt.
Des weiteren wurden 'Mittelwertsimputation', 'Single Hot deck Imputation' und 'Multiple Hot deck Imputation' betrachtet. Imputationen auf Basis einer Hilfsregression wurden nur noch über einen Ansatz bestimmt:

$$X_{mis} \quad = \quad \hat{\gamma}_0 + \hat{\gamma}_1 Y_{mis} + \epsilon, \qquad \epsilon \sim N(0, \hat{\sigma}^2)$$

Die komplexeren Modelle, bei denen noch die Verteilung der Schätzungen der Regressionsparameter und Varianz berücksichtigt wurden, sind hier nicht mehr betrachtet worden. Die bisherigen heuristischen Ergebnisse lassen darauf schließen, dass obiger einfacher Ansatz den Ergebnissen komplexerer Modelle in nichts nachsteht.
Außerdem wurde erneut eine alternative multiple Imputation, die 'Regression-' und 'Hot deck Imputation' kombiniert, berücksichtigt.

Wiederum sollte im Optimalfall ein Gütemaß 1000 mal das richtige Modell ($y = \beta_0 + \beta_1 x + \beta_2 z$) erkennen. Diskussionswürdig ist sicherlich, ob das Modell $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$, das die Interaktion berücksichtigt, auch noch als 'richtig' eingestuft werden kann. In Tabelle 4.10 sind die Ergebnisse des Grundszenarios dargestellt.

Tab. 4.10: Grundszenario. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 26.10 % der x-Werte. M1 bezeichnet die Anzahl der 'richtig' ausgewählten Modelle unter der Annahme, dass nur Modell 13 als korrekt angesehen werden kann, M2 erlaubt darüber hinaus noch Modell 10 als richtig anzusehen.

	Regressionsmodell																		korrekt	
Methode	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	M1	M2
AIC	17	7	14	0	9	57	19	44	75	102	27	0	426	38	0	0	0	165	426	528
AIC CC	22	2	17	7	13	49	15	35	57	74	36	0	375	42	0	0	0	256	375	449
AIC Gewichte wahr	89	16	51	34	40	65	23	59	65	105	45	0	230	44	0	0	0	134	230	335
AIC Gewichte GAM	104	14	46	30	36	64	31	69	49	100	38	0	249	43	0	0	0	127	249	349
AIC Mittelwert	1	0	1	2	8	94	6	5	51	7	4	0	454	34	0	0	0	333	454	501
AIC Hot deck	30	5	6	27	33	100	12	22	32	81	14	0	304	29	0	0	0	305	304	385
AIC Regression	16	6	14	11	17	42	14	30	65	86	33	0	376	42	0	0	0	248	376	462
AIC Multiple Imp. 1	4	1	0	6	18	93	6	8	49	19	7	0	426	28	0	0	0	248	426	445
AIC Multiple Imp. 2	11	3	7	7	13	48	18	20	51	84	30	0	431	50	0	0	0	227	431	515
BIC	0	0	0	0	0	4	0	4	12	19	5	0	436	14	0	0	0	506	436	455
BIC CC	0	0	0	1	0	5	1	4	9	24	8	0	358	27	0	0	0	563	358	382
BIC Gewichte wahr	5	0	6	5	4	13	9	11	25	57	23	0	359	40	0	0	0	443	359	416
BIC Gewichte GAM	36	0	4	3	2	13	6	9	29	49	22	0	365	34	0	0	0	443	365	414
BIC Mittelwert	0	0	0	0	0	5	0	0	1	0	1	0	291	11	0	0	0	691	291	291
BIC Hot deck	0	1	0	1	3	7	1	1	3	19	4	0	198	10	0	0	0	752	198	217
BIC Regression	0	0	0	1	0	4	2	2	6	13	3	0	317	20	0	0	0	632	261	261
BIC Multiple Imp. 1	0	0	0	0	1	8	1	0	2	0	1	0	261	11	0	0	0	715	317	330
BIC Multiple Imp. 2	0	0	0	0	0	1	1	3	10	7	0	0	343	15	0	0	0	620	343	350
C_p	19	7	14	1	8	60	20	48	75	103	28	0	423	37	0	0	0	157	423	526
C_p CC	25	2	17	8	15	53	17	36	60	84	37	0	366	50	0	0	0	230	366	450
C_p Gewichte wahr	96	17	53	36	42	68	27	64	62	105	47	0	221	41	0	0	0	121	221	326
C_p Gewichte GAM	111	15	47	30	40	68	32	73	52	104	35	0	235	39	0	0	0	119	235	104
C_p Mittelwert	1	0	1	2	9	100	6	6	53	8	5	0	456	35	0	0	0	318	456	464
C_p Hot deck	34	5	7	29	34	103	15	25	33	81	13	0	299	31	0	0	0	291	299	380
C_p Regression	18	6	14	12	19	44	14	32	67	88	35	0	368	42	0	0	0	241	368	456
C_p Multiple Imp. 1	5	1	0	6	18	97	6	8	50	22	8	0	425	29	0	0	0	325	425	447
C_p Multiple Imp. 2	15	3	7	8	14	48	19	20	52	90	31	0	424	52	0	0	0	217	424	514

In den Zeilen ist die jeweilige Methode für ein Gütemaß (AIC,BIC,C_p) dargestellt, in den Spalten sind die einzelnen Regressionsmodelle zu erkennen (Die Aufschlüsselung der Zahlen ist Tabelle 4.9 zu entnehmen).
Die letzten beiden Spalten lassen erkennen, wie oft die entsprechende Methode sich für das 'richtige' Modell entschieden hat. Zum einen unter der Annahme, dass nur Modell (13) als korrekt angesehen werden kann (M1), zum anderen unter der Annahme, dass auch das Modell welches die Interaktion zwischen x und z berücksichtigt die Tatsachen korrekt widerspiegelt (M2).

Betrachtet man zuerst die Ergebnisse der verschiedenen Methoden für den AIC, so ist deutlich zu erkennen, dass das Grundmuster der Ergebnisse ähnlich dem von Kapitel 4.1 ist. Für die gewichteten Gütemaße kann erneut ein eher negatives Resümee gezogen werden. Sie schneiden noch immer etwas schlechter ab, als alle anderen Imputationsmethoden. Der Wert von maximal nur 249 richtig erkannten Modellen ist immer noch sehr gering. Nur in etwa einem Viertel der Fälle versprechen diese Methoden gute Ergebnisse.

Entgegen den Erwartungen aus Kapitel 4.1, kann die Mittelwertsimputation der Komplexität des jetzigen Modells noch immer Rechnung tragen. Der Hauptgrund hierfür könnte erneut in dem annähernd symmetrischen Verlauf der Fehlwahrscheinlichkeitsfunktion zu finden sein. Mit 454 richtig erkannten Modellen kann eine deutliche Verbesserung gegenüber der 'Complete Case Analysis' erzielt werden.

Stark herauszuheben ist des weiteren, dass die allgemeine geringe Anzahl an korrekten Entscheidungen vor allem zu Gunsten des Modells (18), mit x als einziger Einflussgröße, geht. So wählt die 'Mean Imputation' beispielsweise ganze 333 mal dieses Modell aus. Hier stellt sich nun die Frage, ob das Nichterkennen von 'z' als Einflussgröße vor allem auf dessen Wahl als Bernoulli-Variable zurückzuführen ist, oder ob generell mit größerer Komplexität Abstriche hinsichtlich des Erfolges der Methodik gemacht werden müssen. Im weiteren Verlauf der Untersuchungen bleibt dies noch zu klären.

Als weiterer Punkt bleibt festzuhalten, dass bei den gewichteten Gütemaßen, im Gegensatz zu den Imputationen, sehr oft ein komplexeres Modell vorgeschlagen wird. Allein 104 mal schlägt der AIC mit den geschätzten Gewichten das volle Modell mit allen Einflussgrößen und Interaktionen vor. Dies sollte nicht passieren. Die Hypothese aus Kapitel 4.1.1, dass gewichtete Gütemaße eher größere und komplexere Modell favorisieren, kann hier zumindestens nicht widerlegt werden.

Positiv zu bemerken ist, dass alle von der Variablen 'v' stark geprägten Modelle, bei allen Methoden, sehr selten ausgewählt wurden. Immerhin wurde

so sehr oft erkannt, dass diese Variable keinen Einfluss besitzt.

Ergebnisse des BIC und Mallows C_p

Vergleicht man die Ergebnisse der Methoden beim BIC mit denen des AIC, so fällt zu allererst auf, dass Modell (18) durchgängig am stärksten favorisiert wird. Wiederum kann dies in erster Linie auf den Strafterm zurückgeführt werden, der kleinere Modelle bevorzugt. Ergebnisse von bis zu 752 gewählten Modellen, mit x als einziger Einflussgröße, sind jedoch mehr als inakzeptabel.

Den gewichteten Methoden kann hier ein zumindestens eingeschränkt positives Zeugnis ausgestellt werden. Sie scheinen weniger anfällig für das falsche Modell zu sein, und liefern unter diesen Umständen teilweise bessere Ergebnisse als dies bei der 'Complete Case Analysis' zu beobachten ist. Es bleibt festzustellen, dass die generelle Aussage der ersten Simulation, Imputationsmethoden zu bevorzugen, zumindestens in Frage gestellt werden kann.
Alle Imputationsmethoden liefern hier eher dürftige Ergebnisse und sind kaum zu empfehlen.

Die Ergebnisse, die man bei Mallows C_p erhält, haben eine gewisse Ähnlichkeit mit denen des AIC. Das Grundmuster ist dasselbe, nur in Details unterscheiden sich die beiden Gütemaße. So lässt sich beispielsweise eine geringe Verbesserung bei der Mittelwertsimputation feststellen.

4.2.2 Variation der z-Variable

Wie in Kapitel 4.2.1 bereits angedeutet, stellt sich die Frage wie stark das teilweise recht unbefriedigende Ergebnis auf die Wahl einer Bernoulli-Variable zurückzuführen ist. Sehr oft wurde nur die Variable x, jedoch nicht die binäre Variable z von den Gütemaßen in das Modell aufgenommen. Für die erste Simulation mit nur zwei Einflussvariablen hat sich herausgestellt, dass die Wahl der Verteilung der Einflussvariablen eigentlich keinen Ausschlag für das Ergebnis geben sollte (siehe auch Kapitel 4.1.5). Ob dies auch für die Situation von drei Einflussgrößen gilt, soll in diesem Abschnitt erörtert werden.

Im Wesentlichen wurde erneut das Grundszenario simuliert, jedoch mit der Ausnahme, dass die z-Variable nun nicht mehr als bernoulliverteilt deklariert wurde, sondern als normalverteilt mit einem Erwartungswert von 5 und einer Standardabweichung von ebenfalls 5. Die Ergebnisse der Simulation sind in Tabelle 4.11 aufgeführt.

Tab. 4.11: Variation der z-Variable. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 38.06 % der x-Werte. M1 bezeichnet die Anzahl der 'richtig' ausgewählten Modelle unter der Annahme, dass nur Modell 13 als korrekt angesehen werden kann, M2 erlaubt darüber hinaus noch Modell 10 als richtig anzusehen.

	Regressionsmodell																		korrekt	
Methode	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	M1	M2
AIC	26	10	14	0	9	62	14	64	108	123	0	0	570	0	0	0	0	0	570	693
AIC CC	35	6	22	13	17	47	28	65	107	141	0	0	519	0	0	0	0	0	519	660
AIC Gewichte wahr	168	29	59	52	40	70	46	91	86	132	0	0	227	0	0	0	0	0	227	359
AIC Gewichte GAM	157	26	43	47	48	63	44	94	105	119	0	0	254	0	0	0	0	0	254	373
AIC Mittelwert	1	1	2	7	8	34	13	5	85	72	0	0	772	0	0	0	0	0	772	844
AIC Hot deck	22	1	12	3	7	18	7	38	58	193	0	0	641	0	0	0	0	0	641	734
AIC Regression	15	2	18	13	15	71	23	42	105	116	0	0	580	0	0	0	0	0	580	696
AIC Multiple Imp. 1	15	0	1	5	2	32	13	16	87	80	0	0	749	0	0	0	0	0	749	829
AIC Multiple Imp. 2	9	3	15	26	12	51	38	21	75	209	0	0	541	0	0	0	0	0	541	650
BIC	0	0	0	0	0	4	1	6	29	39	0	0	921	0	0	0	0	0	921	960
BIC CC	1	0	2	0	3	14	7	10	41	71	0	0	851	0	0	0	0	0	851	922
BIC Gewichte wahr	13	3	11	9	11	42	22	56	85	141	0	0	607	0	0	0	0	0	607	751
BIC Gewichte GAM	32	2	11	8	11	36	13	53	89	118	0	0	627	0	0	0	0	0	627	745
BIC Mittelwert	0	0	0	0	0	3	1	0	9	7	0	0	980	0	0	0	0	0	980	987
BIC Hot deck	0	0	0	1	0	2	1	4	5	70	0	0	916	0	0	1	0	0	916	986
BIC Regression	0	0	0	0	0	7	1	4	43	34	0	0	911	0	0	0	0	0	911	945
BIC Multiple Imp. 1	0	0	0	0	0	2	0	0	12	16	0	0	970	0	0	0	0	0	970	986
BIC Multiple Imp. 2	0	0	0	0	0	10	6	1	25	67	0	0	891	0	0	0	0	0	891	958
C_p	29	10	16	2	10	67	16	67	110	119	0	0	554	0	0	0	0	0	554	673
C_p CC	39	6	23	16	17	49	31	70	108	147	0	0	494	0	0	0	0	0	494	647
C_p Gewichte wahr	175	29	58	55	43	73	48	93	85	131	0	0	210	0	0	0	0	0	210	341
C_p Gewichte GAM	166	28	44	48	48	67	44	92	106	119	0	0	238	0	0	0	0	0	238	357
C_p Mittelwert	1	2	2	7	8	38	13	7	84	74	0	0	764	0	0	0	0	0	764	838
C_p Hot deck	27	1	14	3	8	20	7	39	59	193	0	0	629	0	0	0	0	0	629	722
C_p Regression	15	2	19	15	17	72	23	49	106	116	0	0	566	0	0	0	0	0	566	682
C_p Multiple Imp. 1	15	0	1	5	4	37	13	17	87	81	0	0	740	0	0	0	0	0	740	821
C_p Multiple Imp. 2	11	3	15	27	13	51	40	22	77	210	0	0	531	0	0	0	0	0	531	741

Ein erster Blick auf die Resultate lässt sofort erkennen, dass die Hauptursache für das verhältnismäßig schlechte Abschneiden aller Gütemaße und Methoden in Kapitel 4.1.1, im Wesentlichen die Wahl der z-Variablen war. Modell (18) wird nun bei keiner Methode und keinem Gütemaß als richtig eingestuft. Ansonsten wird zu großen Teilen das richtige Modell, oder zumindestens das mit Interaktionsterm, ausgewählt.

Beim Vergleich der verschiedenen Methoden hat sich das Bild kaum verändert. Die Imputationsmethoden schneiden durchweg besser ab als die gewichteten Gütemaße. Erneut liefern die 'Mean Imputation' und 'Multiple hot deck Imputation 1' sehr gute Ergebnisse.

Zu erkennen ist erneut, dass bei Verwendung des klassischen AIC die Präferenz eindeutig auf kleineren Modellen liegt, die großen Modelle mit vielen Interaktionstermen und Einflussvariablen werden dabei kaum berücksichtigt. Ein genau gegenteiliges Bild ergibt sich bei der Betrachtung des gewichteten AIC. Sowohl bei Verwendung der wahren Gewichte, als auch der geschätzten, werden sehr oft, sehr große Modelle favorisiert. Modell (1) beispielsweise wurde 168 bzw. 157 mal ausgewählt, bei anderen Methoden lag der Wert bei maximal 35. Die in Kapitel 4.1 herausgearbeitete Hypothese, dass die gewichteten Gütemaße zu großen Modellen tendieren, kann erneut bestätigt werden.

Es erweist sich als sehr erstaunlich, wie stark sich die Wahl der z-Variablen auf die Ergebnisse auswirkt. Kleinste Änderungen können gesamte Grundaussagen ändern.
Generell bleibt festzuhalten, dass meistens eine einfache 'Mean Imputation' oder eine multiple 'Hot deck Imputation' sehr gute Ergebnisse liefern, für Extremsituationen sind diese Methoden jedoch auch meist sehr anfällig.

Ergebnisse des BIC und Mallows C_p

Wird das BIC als Gütemaß verwendet, so bestätigt sich erneut das Bild, dass deutlich mehr Modelle richtig klassifiziert werden. Alle Imputationsmethoden erbringen eine Verbesserung gegenüber der 'Complete Case Analysis'. Für die 'Mean Imputation' ergab sich der beste Wert. 980 mal wurde das Modell richtig erkannt, unter der Annahme eines korrekten Modells (10) sogar ganze 987 mal. Diese Anzahl kann als überwältigend hoch eingestuft werden.

Für Mallows C_p ergeben sich ähnliche Resultate wie beim AIC. Das Grundmuster ist dasselbe, in der Tendenz scheinen alle Methoden jedoch weniger oft das richtige Modell zu erkennen.

4.2.3 Variation für die Schätzung eines GAM

Bisher wurde bei der Schätzung der Gewichte für die Gütemaße durch ein generalisiertes additives Modell immer der gesamte Kenntnisstand der Simulation miteinbezogen. Das heißt, die Fehlwahrscheinlichkeit wurde richtigerweise immer unter der Annahme geschätzt, dass das Fehlen eines Wertes als Funktion von y aufzufassen ist.
In der Realität kann diese Situation jedoch anders aussehen. Selbst unter einer MAR-Annahme, also unter der Voraussetzung, dass die fehlenden Werte nicht von x selbst abhängen, könnten noch einige andere GAMs gefittet werden.

Zu untersuchen ist nun

1. wie eindeutig die Wahl für das Modell der Fehlwahrscheinlichkeit in Abhängigkeit einer Funktion von y ist,

2. wie eine falsche Wahl das Ergebnis beeinflussen kann.

Zur Untersuchung des ersten Punktes wurde für sechs verschiedene Modelle (siehe auch Tabelle 4.12) das Grundszenario für 1000 verschiedene Samples simuliert um zu testen, ob die p-Werte der gefitteten Modelle tatsächlich nur den Term 'f(y)' als statistisch signifikant einstufen oder auch andere Terme. Als einzige Änderung wurde die z-Variable als normalverteilt mit Erwartungswert und Standardabweichung von jeweils 5 vorgegeben um eine einwandfreie Schätzung des GAM zu gewährleisten.

Tab. 4.12: Variation des generalisierten additiven Modells zur Schätzung der Fehlwahrscheinlichkeit. Sechs mögliche Modelle wurden dabei betrachtet. $P(\delta)$ bezeichnet die Wahrscheinlichkeit für einen Wert zu fehlen.

(1)	$P(\delta) = f(y)$	(2)	$P(\delta) = f(z)$
(3)	$P(\delta) = f(v)$	(4)	$P(\delta) = f(y) + f(z)$
(5)	$P(\delta) = f(y) + f(v)$	(6)	$P(\delta) = f(z) + f(v)$

Die Ergebnisse der Simulationen sind sehr vielversprechend. Für das erste Modell ist der Term 'f(y)' stets hochsignifikant. Bei den Modellen (4) und (5) ist 'f(y)' ebenfalls stets hoch signifikant, die anderen Terme 'f(z)' und 'f(v)' dagegen nicht. Für Modell (6) sind keine signifkanten Einflüsse bei

beiden Termen zu beobachten, auch in Modell (3) wird 'f(v)' nie als signifikant eingestuft. Einzig bei Modell (2) wird unter den 1000 Samples 'f(z)' des öfteren als signifikant zum 5% Niveau eingestuft.

Dies zeigt, dass die Variable y einen Einfluss auf die Fehlwahrscheinlichkeit haben sollte, eine eindeutige Entscheidung *gegen* den Term 'f(z)' konnte jedoch nicht konstatiert werden. Daher soll nun untersucht werden, ob sich eine falsche Entscheidung beim Fitten eines GAM-Modells auf die Ergebnisse entscheidend auswirken kann. Tabelle 4.13 zeigt die Resultate der Simulationen.

Tab. 4.13: Anzahl an richtig erkannten Modellen (Modell 13) je GAM und Gütemaß.

	Anzahl an richtig erkannten Modellen					
gefittete	AIC		BIC		C_p	
Modelle	wahr	GAM	wahr	GAM	wahr	GAM
$P(\delta) = f(y)$	195	217	558	580	183	211
$P(\delta) = f(z)$	241	247	640	650	224	235
$P(\delta) = f(y) + f(z)$	261	282	621	636	237	262

Für die Wahl des Modells $P(\delta) = f(z)$, sowie $P(\delta) = f(y) + f(z)$ sollte man nun im Vergleich zu $P(\delta) = f(y)$ eigentlich schlechtere Ergebnisse erwarten, die Wahrscheinlichkeiten sollten schlechter geschätzt werden, und damit auch zu einer geringeren Anzahl an richtig gewählten Modellen führen. Das Gegenteil ist jedoch der Fall. Sogar eine Verbesserung in beiden Fällen ist zu erkennen. Der Grund hierfür ist jedoch simpler als vielleicht gemeinhin anzunehmen. Abbildung 4.7 verdeutlicht dies.

Während beim Modell $P(\delta) = f(y)$ die Gewichte annähernd gut geschätzt werden, ist bei den Modellen $P(\delta) = f(z)$ und $P(\delta) = f(y)+f(z)$ nur eine geringe Variabilität unter den Gewichten zu erkennen. Eine Verbesserung wird nur dadurch erzielt, dass eine Annäherung an die 'Complete Case Analysis' erfolgt. Die auf den ersten Blick verwunderliche Anordnung von Punkten als Gewichtsschätzung für das Modell $P(\delta) = f(z)$ ist durch eine hohe Anzahl an stark gewichteten hochgradigen Polynomen zu erklären.

Um auch ohne eine detaillierte grafische Betrachtung zu obigem Schluss zu kommen kann die Summe des Absolutbetrages der Differenz von wahren und geschätzten Auswahlwahrscheinlichkeiten betrachtet werden:

$$\xi = \sum_{i=1}^{n} |P(\delta_i)_{wahr} - P(\delta_i)_{GAM}| \,. \tag{4.1}$$

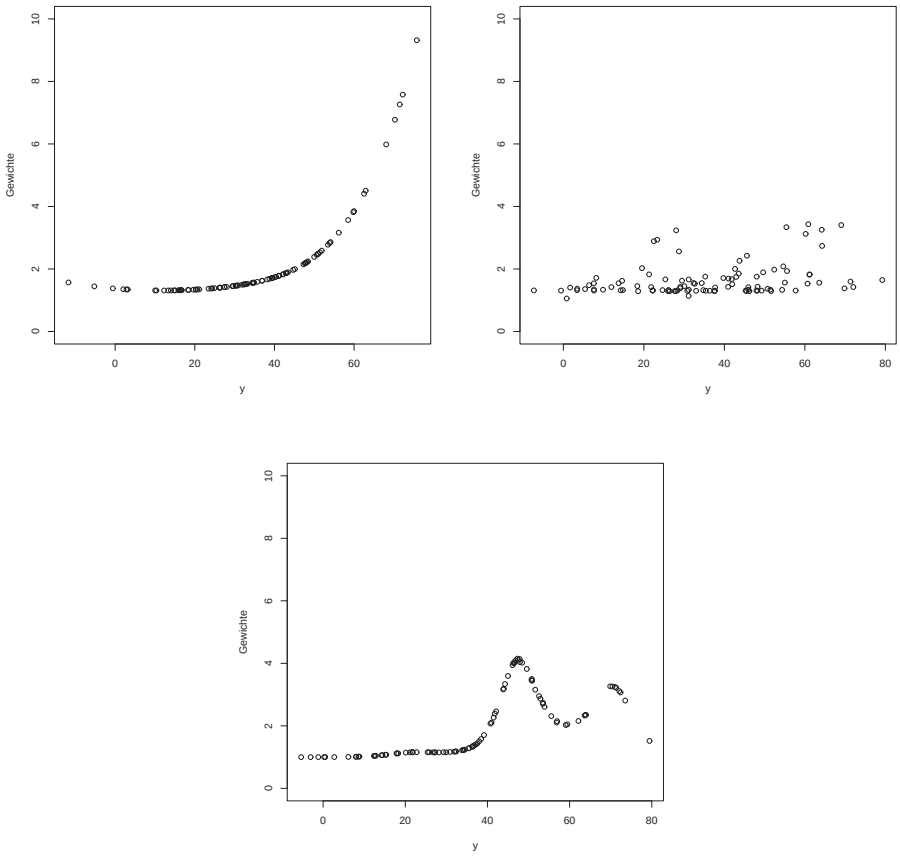

Abb. 4.7: geschätzte Gewichte für drei willkürlich ausgewählte Datenquartupel: $P(\delta) = f(y)$ (oben links), $P(\delta) = f(z)$ (oben rechts), $P(\delta) = f(y) + f(z)$ (unten) .

Für 1000 Samples wäre es also sinnvoll das arithmetische Mittel der einzelnen ξ zu betrachten. Für die drei oben betrachteten Modelle lauten die Werte für $\bar{\xi}$ wie folgt:

Modell	$P(\delta) = f(y)$	$P(\delta) = f(z)$	$P(\delta) = f(y) + f(z)$
$\bar{\xi}$	9.77	15.87	11.21

Die Werte bestätigen die Vermutung. Der geringste Wert ist beim Modell $P(\delta) = f(y)$ zu erkennen und bestätigt die Hypothese, dass das gute Abschneiden der Modelle $P(\delta) = f(z)$ und $P(\delta) = f(y) + f(z)$ allein durch die Annäherung an die 'Complete Case Analysis' zu erklären ist. Die $\bar{\xi}$ können

für n=100 auch als mittlere prozentuale Abweichung von geschätzten und wahren Auswahlwahrscheinlichkeiten interpretiert werden.

4.2.4 Variation der Variablen mit fehlenden Werten

In den bisherigen Variationen der Simulation mit drei Einflussgrößen wurde stets nur der Fall *einer* Variable mit fehlenden Werten betrachtet. Interessant ist, auch unter Aspekten der Praxisrelevanz, inwieweit sich Resultate ändern, wenn mehrere Einflussgrößen fehlende Werte aufweisen.
Entscheidender Unterschied ist hier auch die Tatsache, dass mit einer Zunahme an Variablen mit fehlenden Werten die relevanten Fälle für die 'Complete Case Analysis' abnehmen sollten. In einer solchen Simulation sollte also zu zeigen sein, welche Gütemaße wirklich eine Verbesserung erwirken können.

Gegenüber dem Grundszenario wurde eine wesentliche Änderung vorgenommen. Unter den drei Einflussvariablen wiesen nun zwei fehlende Werte auf. Werte von x (also der gleichverteilten Zufallsvariable) wurden gewohnterweise mit einer Wahrscheinlichkeit von

$$\pi(v,x,z)_x = \begin{cases} 1-(1+0.015\cdot y^2)^{-1} & für \quad x \leq 0 \\ 1-(1+0.0005\cdot y^2)^{-1} & für \quad x > 0 \end{cases}$$

als fehlend deklariert. Im Unterschied dazu fehlten Werte bei v (also der exponentialverteilten Zufallsvariable) mit einer Wahrscheinlichkeit von

$$\pi(v,x,z)_v = 1 - \frac{1}{0.0015(y-20)^2+1}\,.$$

Die fehlenden Werte in den beiden Variablen unterlagen also nicht dem gleichen Zufallsmechanismus, waren aufgrund ihrer Abhängigkeit von y aber noch immer jeweils MAR. Die jeweiligen Fehlwahrscheinlichkeits-Funktionen, sowie die Fehlwahrscheinlichkeitsfunktion für das Fehlen eines Werte insgesamt ($\pi(v,x,z)_{v\cup x}$) sind noch einmal in Abbildung 4.8 aufgeführt. Deutlich zu erkennen ist hier, dass die Wahrscheinlichkeit für das Fehlen eines Wertes insgesamt linksseitig des Wertes 20 stärker ansteigt, als für Werte größer 20. Die über generalisierte additive Modelle geschätzten Gewichte wurden für die Fehlwahrscheinlichkeiten $\pi(v,x,z)_{v\cup x}$ geschätzt.

In Tabelle 4.14 sind die Ergebnisse der Simulation aufgelistet.

Zuallererst fällt auf, dass der Unterschied an richtig erkannten Modellen für vollständige Daten und 'Complete Case Analysis' deutlich größer ist als in

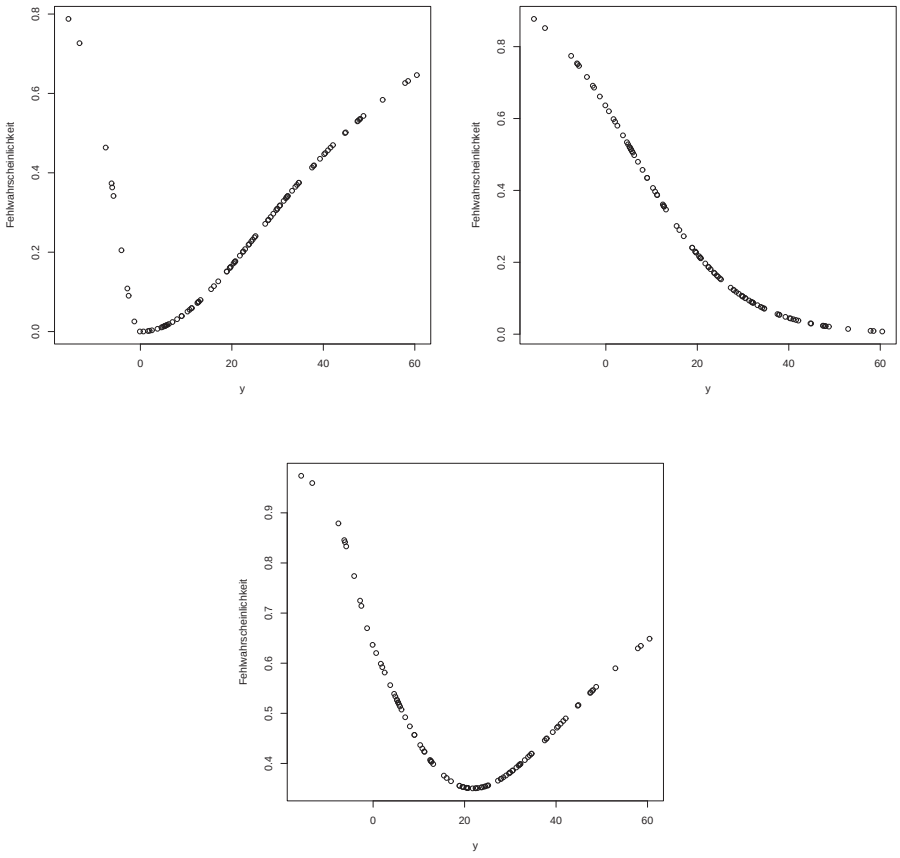

Abb. 4.8: Fehlwahrscheinlichkeiten für die x-Variable (oben links), für die v-Variable (oben rechts), sowie insgesamt (unten). Beispiel eines willkürlichen Datenquartupels.

den bisherigen Simulationen. 420 richtig erkannten Modellen beim Betrachten aller Daten stehen nur 301 für die 'Complete Case Analysis' gegenüber. Durch die Hinzunahme einer weiteren Variable mit fehlenden Werten stellt sich die Gesamtsituation nun komplexer dar. Ein gutes Gütemaß sollte nun also in der Lage sein eine Verbesserung gegenüber der 'Complete Case Analysis' zu erwirken.

Betrachtet man nun die gewichteten AICs, so kann keine Verbesserung festgestellt werden. Die Werte liegen hier mit 145 (für die wahren Gewichte), sowie 163 (für die geschätzten Gewichte) deutlich unter den Erwartungen die man an ein zuverlässiges Gütemaß stellen würde.

Tab. 4.14: Variation der Variablen mit fehlenden Werten. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 47.25 % der Werte (x oder v). M1 bezeichnet die Anzahl der 'richtig' ausgewählten Modelle unter der Annahme, dass nur Modell 13 als korrekt angesehen werden kann, M2 erlaubt darüber hinaus noch Modell 10 als richtig anzusehen.

	Regressionsmodell																		korrekt	
Methode	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	M1	M2
AIC	18	11	15	0	13	50	20	43	85	84	20	0	420	25	0	0	0	196	420	504
AIC CC	19	6	13	15	15	51	16	20	49	75	45	0	301	59	0	0	0	316	301	376
AIC Gewichte wahr	162	39	48	45	46	89	45	63	59	105	48	0	145	24	0	0	0	82	145	250
AIC Gewichte GAM	139	35	44	44	44	85	42	53	76	100	47	0	163	30	0	0	0	98	163	263
AIC Mittelwert	1	0	1	1	5	98	3	8	77	24	1	0	546	21	0	0	0	214	546	570
AIC Hot deck	26	5	10	10	25	72	19	37	49	97	6	0	432	24	0	0	0	188	432	529
AIC Regression	11	2	9	15	11	59	11	40	81	79	32	0	371	49	0	0	0	230	371	450
AIC Multiple Imp. 1	5	3	0	3	3	87	6	19	73	47	3	0	524	23	0	0	0	204	524	571
AIC Multiple Imp. 2	17	5	8	10	12	54	24	43	71	91	21	0	409	37	0	0	0	198	409	500
BIC	0	0	1	0	3	4	3	1	13	11	6	0	433	20	0	0	0	505	433	444
BIC CC	1	0	0	3	0	6	4	0	16	20	12	0	256	37	0	0	0	645	256	276
BIC Gewichte wahr	25	7	15	9	14	41	21	28	51	92	44	0	282	55	0	0	0	316	282	374
BIC Gewichte GAM	18	2	13	13	19	39	14	19	34	77	38	0	287	54	0	0	0	373	287	364
BIC Mittelwert	0	0	0	0	0	2	0	0	8	1	0	0	428	18	0	0	0	543	428	429
BIC Hot deck	1	0	0	1	1	6	3	0	9	23	6	0	377	15	0	0	0	558	377	400
BIC Regression	0	0	0	0	0	7	0	3	25	11	4	0	342	36	0	0	0	572	342	353
BIC Multiple Imp. 1	0	0	0	0	1	3	0	0	12	1	0	0	411	13	0	0	0	559	411	412
BIC Multiple Imp. 2	0	0	0	0	1	3	1	3	14	21	0	0	384	20	0	0	0	553	384	405
C_p	22	12	17	3	10	55	20	42	86	86	21	0	414	27	0	0	0	185	414	500
C_p CC	22	6	15	18	20	55	17	24	51	84	50	0	295	62	0	0	0	281	295	379
C_p Gewichte wahr	177	41	49	42	45	92	47	65	53	104	45	0	140	22	0	0	0	78	140	244
C_p Gewichte GAM	147	36	44	47	46	90	42	58	73	105	46	0	148	29	0	0	0	89	148	253
C_p Mittelwert	2	0	1	3	5	102	3	8	77	27	1	0	545	23	0	0	0	89	545	582
C_p Hot deck	29	6	11	12	25	73	21	38	50	98	7	0	428	25	0	0	0	177	428	526
C_p Regression	13	2	9	15	12	60	12	42	84	82	33	0	368	49	0	0	0	219	368	450
C_p Multiple Imp. 1	6	4	0	4	3	89	6	22	77	45	3	0	522	23	0	0	0	196	522	567
C_p Multiple Imp. 2	19	5	10	12	14	52	26	42	71	98	22	0	405	37	0	0	0	187	405	503

Verhältnismäßig oft wird das Modell (18), dass nur die Variable x als signifikant erkennt, ausgewählt. Wesentlicher Unterschied ist aber vor allem die starke Anfälligkeit für das Auswählen von Modellen mit vielen Einflussgrößen. Das Modell mit allen Variablen und Interaktionstermen wird 106 bzw. 109 mal als am besten eingestuft. Es zeigt sich nun auch in dieser Simulation, welch große Schwächen ein gewichtetes Gütemaß mit sich bringen kann, und dass die Tendenz zu größeren Modellen unübersehbar ist.

Sehr gute Ergebnisse liefert erneut die einfache Mittelwertsimputation. 546 mal wurde das richtige Modell ausgewählt. Die wesentlichen Fehlklassifikationen beliefen sich auf Modell (18), bei dem die Bernoulli-Variable als nicht signifikant eingestuft wird. Große Modelle, auch mit vielen Interaktionstermen werden bei der 'Mean Imputation' nur äußerst selten ausgewählt.

Eine einfache 'Hot deck Imputation' lieferte ebenfalls eine Verbesserung gegenüber der 'Complete Case Analysis'. 432 mal wurde hier das richtige Modell ausgewählt. Unter der Annahme, dass auch Modell (10) mit dem Interaktionsterm als richtig angesehen werden kann, wird in über der Hälfte der Fälle (529) ein adäquates Modell ausgewählt. Dennoch sind die Ergebnisse nicht ganz so gut einzustufen wie bei einer Mittelwertsimputation.

Am schlechtesten schneidet bei den Imputationsmethoden die Regressionsimputation ab. Auch wenn eine Verbesserung erzielt werden konnte, liegt der Wert von 371 richtig erkannten Modellen weit unter denen der anderen Methoden. Dies ist insofern erstaunlich, als dass die Regressionsimputation bisher recht stabile Ergebnisse aufweisen konnte. Dennoch ist auch hier eine Verbesserung gegenüber der 'Complete Case Analysis' festzustellen.
Das Hauptproblem dieser Methode zeigt sich hier in seiner ganzen Bandbreite. Es stellt sich die Frage wie mit zunehmender Anzahl an Einflussgrößen ein sinnvolles, jedoch auch nicht willkürliches, Regressionsmodell ausgewählt werden kann, um eine erfolgreiche Imputation durchzuführen. Eine genauere Analyse dieser Problemstellung könnte hierzu Hinweise geben.

Die alternativen multiplen Imputationsverfahren liefern erneut beständig gute Ergebnisse (524 bzw. 409 richtig erkannte Modelle). Beim Imputationsverfahren mit dem Einfluss der Regressionsimputation können verständlicherweise nicht ganz so überzeugende Resultate erzielt werden.

Ergebnisse des BIC und Mallows C_p

Betrachtet man die Ergebnisse zur Entscheidungsfindung anhand des Schwarzschen Bayeskriteriums, so ergeben sich interessante Ergebnisse.

Die gewichteten BICs können nun endlich eine geringe Verbesserung gegenüber der 'Complete Case Analysis' erzielen. Im Vergleich zu den Imputationsmethoden ist diese zwar gering, in Anbetracht der bisher erzielten Ergebnisse jedoch durchaus respektabel. Zu vermuten ist, dass die Verbesserung vor allem auf die Eigenschaft des BIC kleine Modelle zu bevorzugen zurückzuführen ist. Die sehr komplexen Modelle werden daher weit weniger oft ausgewählt, was dem gewichteten BIC in dieser spezifischen Situation zu Gute kommt. Dennoch bleibt festzuhalten, dass nur etwa in einem Viertel der Fälle das richtige Modell ausgewählt wurde, was immer noch sehr unbefriedigend ist.

Im Bereich der Imputationsmethoden erwies sich als geeignetste Methode die 'Mean Imputation'. Sie lag in 428 der 1000 Fälle mit ihrer Modellwahl richtig. Im Grundmuster ähneln sich die Ergebnisse des AIC und des BIC erneut, entscheidender Unterschied ist hier jedoch erneut die häufige Auswahl des Modells (18), dass nur x als Einflussgröße besitzt. Dies ist in Anbetracht der Ergebnisse der vorhergehenden Simulationen zwar nicht erstaunlich, verdeutlicht aber erneut die erstaunlich große Problematik der Bernoulli-Variable in das Modell mitaufgenommen zu werden.

Im Allgemeinen lässt sich festhalten, dass der Gewinn bezüglich der 'Complete Case Analysis' deutlich geringer ist, als bei der Entscheidungsfindung auf der Basis des AIC. Der Grund hierfür ist sehr einfach. Aufgrund der Präferenz des BIC sparsame Modelle zu bevorzugen gibt es, wie bereits erwähnt, sehr viele Fehlklassifikationen zugunsten des Modells (18) mit nur einer Einflussgröße. Daraus resultieren auch die vergleichsweise schlechten Ergebnisse des BIC.

Die Simulation scheint kein Spiegelbild der bisherigen Simulationen zu sein. Aufgrund der größeren Komplexität bei den fehlenden Werten, sind hier die Ergebnisse differenzierter zu betrachten und zu interpretieren.

Beispielhaft zeigt diese Simulation die Komplexität der Wahl einer geeigneten Imputation und eines geeigneten Gütemaßes. Zuviele Faktoren, im Wesentlichen jedoch vor allem die Anzahl und die Eigenschaften der Einflussgrößen, bestimmen die Güte eines Maßes in einer spezifischen Situation. So kann in erster Linie die Stabilität, auch in Extremsituationen, ein Anhaltspunkt für eine Auswahl darstellen.

Mallows C_p zeigt erneut starke Ähnlichkeit mit Akaikes Informationskriterium. Die Ergebnisse sind im Kern gleich und unterstreichen noch einmal die Aussage, dass die beiden Gütemaße trotz eines grundverschiedenen Ansatzes, und einer unterschiedlichen Berechnung, konstante Gemeinsamkeiten aufweisen.

4.2.5 Variation der Variablen mit fehlenden Werten 2

In diesem Abschnitt soll untersucht werden, inwiefern sich ein Fehlen bei *allen* drei Einflussgrößen auswirken kann. Die bernoulliverteilte Zufallsvariable z soll daher nun ebenfalls fehlende Werte aufweisen. Die zugehörige Fehlwahrscheinlichkeitsfunktion lautet

$$\pi(v,x,z)_z = 1 - \frac{1}{\exp(1 - 0.09 \cdot (y+5))}\,.$$

Analog zu Abschnitt 4.2.4 fehlten v-Werte mit einer Wahrscheinlichkeit von

$$\pi(v,x,z)_v = 1 - \frac{1}{0.0015(y-20)^2 + 1}\,,$$

sowie die Werte von x mit einer Wahrscheinlichkeit von

$$\pi(v,x,z)_x = \begin{cases} 1 - (1 + 0.015 \cdot y^2)^{-1} & für \quad x \leq 0 \\ 1 - (1 + 0.0005 \cdot y^2)^{-1} & für \quad x > 0 \end{cases}\,.$$

In Abbildung 4.9 sind noch einmal alle Fehlwahrscheinlichkeitsfunktionen abgebildet. Eindeutig ist zu erkennen, dass der Verlauf der neuen Funktion sich von den beiden anderen unterscheidet. Je geringer ein Wert, desto höher die Wahrscheinlichkeit, dass er fehlt.
Ebenfalls aufgeführt ist die Fehlwahrscheinlichkeitsfunktion für das Fehlen eines Wertes insgesamt ($\pi(v,x,z)_{v \cup x \cup z}$). Da sich die Fehlwahrscheinlichkeitsfunktionen der Variablen v und z in etwa egalisieren, ist die Wahrscheinlichkeit für das Fehlen eines Wertes allgemein, näherungsweise gleich der Wahrscheinlichkeit der Variablen x.

Bei diesem Szenario fehlten im Durchschnitt 22.87% der x-Werte, 18.88% der v-Werte und 23.90% der z-Werte, insgesamt mussten 53.52% der Fälle als nicht vollständig eingestuft werden.

Für die Imputation von fehlenden z-Werten war es nötig, einige Veränderungen durchzuführen. Bei der 'Mean Imputation' wurde der Mittelwert als Wahrscheinlichkeit interpretiert und mit dieser aus den Werten Null und Eins gezogen.
Für eine einfache 'Hot deck Imputation' ergab sich keine Veränderung. Es wurde wie üblich aus der empirischen Verteilung ein Wert gezogen. Für die alternative 'Multiple Hot deck Imputation' dagegen wurden 5 Werte aus der empirischen Verteilung gezogen und der Mittelwert erneut als Wahrscheinlichkeit interpretiert.
Für die Regressionsimputation wurde ein Logit-Modell gefittet und mit den

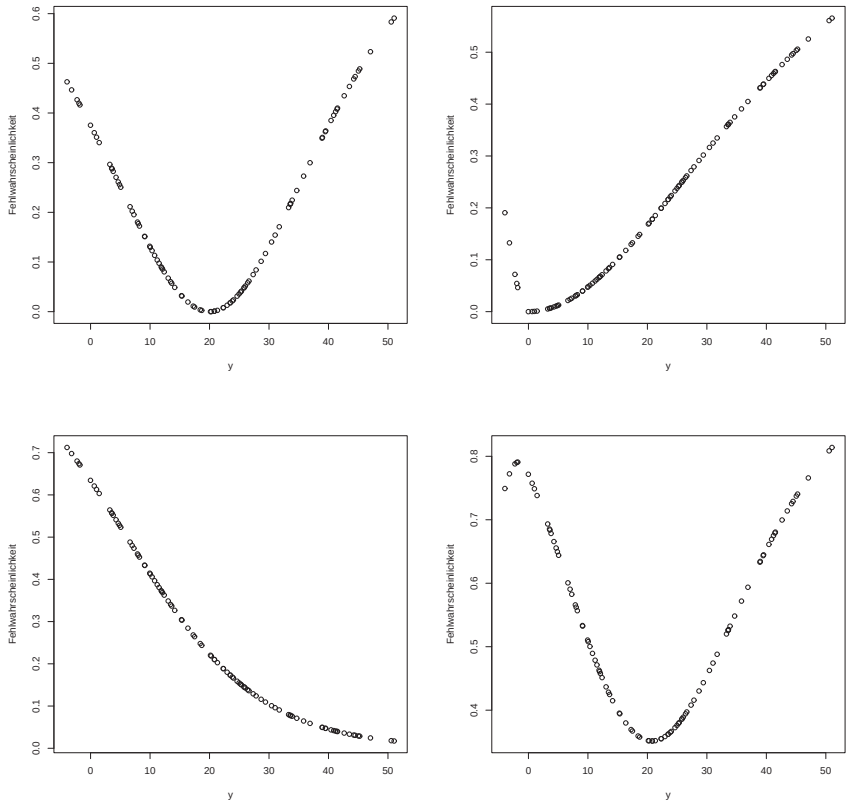

Abb. 4.9: Fehlwahrscheinlichkeiten für die x-Variable (oben links), Fehlwahrscheinlichkeiten für die v-Variable (oben rechts), Fehlwahrscheinlichkeiten für die z-Variable (unten links) sowie insgesamt (unten rechts). Beispiel eines willkürlichen Datenquartupels.

geschätzten Wahrscheinlichkeiten entsprechend aus den Werten Null und Eins gezogen.
Bei der Kombination aus 'Hot deck-' und 'Regressionsimputation' wurde erneut der Mittelwert der 10 Werte als Wahrscheinlichkeit interpretiert.

In Tabelle 4.15 sind die Ergebnisse der Simulation aufgelistet.

Betrachtet man die Resultate innerhalb des AIC, so fällt auf, dass beim Erkennen des richtigen Modells erneut dasselbe Grundmuster wie in den vorherigen Simulationen zu finden ist.

Tab. 4.15: Variation der Variablen mit fehlenden Werten 2. Die Zahlen vermitteln wie oft ein Modell für jede Strategie ausgewählt wurde. Im Schnitt fehlten 55.52 % der Werte (x,z oder v). M1 bezeichnet die Anzahl der 'richtig' ausgewählten Modelle unter der Annahme, dass nur Modell 13 als korrekt angesehen werden kann, M2 erlaubt darüber hinaus noch Modell 10 als richtig anzusehen.

	Regressionsmodell																		korrekt	
Methode	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	M1	M2
AIC	23	6	15	0	20	69	18	66	79	116	0	0	588	0	0	0	0	0	588	704
AIC CC	29	10	10	16	21	77	23	62	77	127	0	0	539	2	0	0	0	7	539	666
AIC Gewichte wahr	206	48	64	56	49	83	43	96	66	129	1	0	156	1	0	0	0	2	156	285
AIC Gewichte GAM	198	44	60	51	57	85	49	95	60	124	0	0	174	1	0	0	0	2	174	296
AIC Mittelwert	2	0	0	3	0	15	3	3	25	47	3	0	375	40	0	0	0	484	375	422
AIC Hot deck	15	4	6	6	11	28	6	13	21	90	38	0	210	52	0	0	0	500	210	300
AIC Regression	12	3	5	11	21	69	21	53	72	129	3	0	562	6	0	0	0	33	562	691
AIC Multiple Imp. 1	2	0	0	1	2	30	7	2	30	33	7	0	337	42	0	0	0	507	337	370
AIC Multiple Imp. 2	14	0	4	7	9	39	17	28	52	134	10	0	482	19	0	0	0	185	482	616
BIC	0	0	0	0	0	5	3	6	30	42	0	0	913	0	0	0	0	1	913	955
BIC CC	2	1	1	2	2	24	4	21	32	58	0	0	805	1	0	0	0	47	805	863
BIC Gewichte wahr	42	10	22	20	13	61	23	68	67	158	1	0	505	1	0	0	0	9	505	663
BIC Gewichte GAM	39	6	17	18	20	55	22	62	73	155	1	0	517	1	0	0	0	14	517	672
BIC Mittelwert	0	0	0	0	0	1	0	0	1	4	0	0	205	12	0	0	0	777	205	209
BIC Hot deck	0	0	0	2	0	2	0	2	4	13	5	0	76	14	0	0	0	882	76	89
BIC Regression	0	0	0	0	2	11	0	6	22	32	0	0	741	7	0	0	0	179	741	773
BIC Multiple Imp. 1	0	0	0	0	0	2	0	0	3	3	0	0	164	9	0	0	0	819	164	167
BIC Multiple Imp. 2	0	0	0	1	0	0	0	2	9	31	0	0	448	8	0	0	0	501	448	479
C_p	24	7	15	2	18	73	19	68	80	116	0	0	578	0	0	0	0	0	578	694
C_p CC	32	10	13	18	24	85	22	65	85	132	0	0	506	2	0	0	0	6	506	638
C_p Gewichte wahr	216	49	66	58	49	82	43	98	66	126	1	0	143	1	0	0	0	2	143	269
C_p Gewichte GAM	209	45	63	52	58	88	49	97	57	120	0	0	159	1	0	0	0	2	159	279
C_p Mittelwert	2	0	0	3	0	17	3	3	25	50	3	0	374	42	0	0	0	478	374	424
C_p Hot deck	18	4	7	7	14	30	7	14	24	93	39	0	203	50	0	0	0	490	203	296
C_p Regression	15	3	5	11	21	72	21	55	75	133	3	0	549	6	0	0	0	31	549	682
C_p Multiple Imp. 1	3	0	0	1	2	34	8	2	30	34	8	0	338	41	0	0	0	499	338	372
C_p Multiple Imp. 2	14	0	8	7	10	39	18	29	55	140	9	0	476	17	0	0	0	178	476	616

Interessanterweise kann fast keine der Imputations- oder Gewichtungsmethoden wirklich eine Verbesserung gegenüber der 'Complete Case Analysis' erwirken. Die Werte der Gewichtungsmethoden liegen dabei mit 123 und 107 weit von einem zufriedenstellenden Ergebnis entfernt. Eine eingeschränkte Betrachtung der gewichteten AICs könnte keine eindeutige Entscheidung zu Gunsten eines Modells liefern. Fast jedes Modell wurde hier mehrmals ausgewählt. Wiederum wurde das größte Modell mit allen Variablen und Interaktionen am häufigsten vorgeschlagen.

Eher dürftige Ergebnisse liefern auch die meisten Imputationsmethoden. Am besten schneidet hier noch die Imputation auf Basis einer Hilfsregression ab. Insgesamt 562 mal wurde das richtige Modell dabei erkannt, und konnte damit im Vergleich zu den anderen Methoden immerhin eine Verbesserung gegenüber der 'Complete Case Analysis' erzielen.

Alle anderen Imputationsmethoden erzielen unter dem Gesichtspunkt richtig erkannter Modelle nicht nur schlechte Ergebnisse, sondern tendieren erneut dazu die bernoulliverteilte Variable nicht zu erkennen, und daher Modell (18) als richtig einzustufen. Dass sich diese Problematik sehr konsequent durch fast alle Simulationen zieht, konnte sicherlich nicht erwartet werden.

Da für die z-Variable ein β von 3 geschätzt werden sollte, und die Standardabweichung in den y-Werten bei exp(2) liegt, kann von einem guten Gütemaß durchaus erwartet werden, das richtige Modell vorherzusagen. Leider geschieht dies hier zu selten.

Ergebnisse des BIC und Mallows C_p

Beim Betrachten des BIC können leicht differenzierte Ergebnisse konstatiert werden.

Die Anfälligkeit für das Modell (18) liegt erneut bei den Imputationsmethoden, der generelle Erfolg dieser Methoden in dieser Simulation variiert jedoch weit stärker als bisher beobachtet. Auch wenn eine einfache 'Hot deck Imputation' sehr schlechte Ergebnisse liefert (76 richtig erkannte Modelle), so ist der Abstand zur besten Imputationsmethode (Regression Imputation, 741 richtige Modelle) extrem hoch. Eine Verbesserung gegenüber der 'Complete Case Analysis' wird in diesem Szenario nie erreicht.

Immerhin bleibt festzuhalten, dass eine Regressionsimputation einigermaßen stabile Ergebnisse liefert, und bei einer Grundkenntnis der Ursache des Fehlens sehr zu empfehlen ist.

Das gewichtete BIC schneidet im Wesentlichen besser ab als die Imputationsmethoden. Im Verhältnis zu den Ergebnissen der 'Complete Case Analysis', sind diese Werte jedoch immer noch äußerst gering und stellen diesem Maß ein schlechtes Zeugnis aus.

Die Resultate von Mallows C_p sind im Muster wieder ähnlich zu denen des AIC. Erneut kann konstatiert werden, dass nur etwas weniger oft das richtige Modell erkannt wird.

4.2.6 Weitere Variationen

Ein Szenario mit drei Einflussgrößen und einer Zielgröße bietet eine Vielzahl an möglichen Variationen. Diese alle detailliert aufzulisten und zu analysieren sprengt mit Sicherheit den Rahmen einer solchen Arbeit.

Auf den ein oder anderen interessanten Punkt soll hier jedoch noch einmal eingegangen werden, wenn auch nicht in der Tiefe, so jedoch zumindest in den Grundaussagen.

Korrelation unter den Einflussgrößen

Wie bereits in Kapitel 4.1.7 angedeutet, kann eine Korrelation unter den Einflussgrößen Ergebnisse stark beeinflussen und verändern. Deshalb wurde die Variable v nicht mehr als exponentialverteilt generiert, sondern als normalverteilte Zufallsvariable mit Mittelwert '$10x + 2$' und Varianzen von 'exp(2.5)', 'exp(3.75)' und 'exp(6)'. So konnte eine Korrelation unter den Einflussgrößen x und v simuliert werden. Die (vereinfachten) Ergebnisse sind in Tabelle 4.16 zu erkennen.

Betrachtet man zunächst die Ergebnisse, so erlaubt sich sowohl ein Vergleich mit dem Grundszenario, als auch mit Kapitel 4.1.7. Schon dort wirkte sich eine hohe Korrelation vor allem auf die Mittelwertsimputation und die 'Hot Deck'-Imputationen aus. Die Ergebnisse in diesem Szenario lassen auf Ähnliches schließen.

Im Fall einer nur geringen Korrelation (0.08) zwischen v und x lieferten alle Gütemaße in etwa die Qualität an Ergebnissen, wie sie auch schon in den vorhergegangenen Simulationen zu finden war. Gegenüber dem Grundszenario ergaben sich in Betrag und Verhältnis der Resultate kleinere Unterschiede, die mit der Veränderung der Variable v erklärt werden können.
Eine höhere Korrelation, unabhängig ob 0.58 oder 0.92, erwirkt vor allem

Tab. 4.16: Grundszenario. Die Zahlen vermitteln wie oft das 'korrekte' Modell mit dem AIC für die entsprechende empirische Korrelation ausgewählt wurde.

Methode	Korrelation		
	0.08	0.52	0.92
Originaldaten	441	420	469
Complete Cases	371	361	351
wahre Gewichte	246	216	269
geschätzte Gewichte GAM	227	249	284
Imputation Mittelwert	182	70	0
Imputation Hot deck	72	16	0
Imputation Regression	352	356	222
Multiple Imputation 1	160	51	0
Multiple Imputation 2	295	310	8

eine Schwächung der 'Mean Imputation' und 'Hot deck Imputation'. So gut wie nie wurden in diesem Fall die richtigen Modelle vorhergesagt.

Variation der Varianz

Schon in Kapitel 4.1.4 wurde untersucht, inwieweit sich eine Variation der Varianz bei den y-Werten auswirken kann. Das für die Variable z zu schätzende β besitzt den Wert drei. Eine hohe Varianz unter den y-Werten kann eine Entscheidung für ein Gütemaß erschweren.

Untersucht werden soll nun, ob eine Variation der Varianz Einfluss auf die Ergebnisse hat, und welche Methoden am stabilsten darauf reagieren.
Die normalverteilte Zufallsvariable y mit dem Mittelwert '$-4 + 5x + 3z$' soll nun bezüglich ihrer Varianz variiert werden. Getestet wurden die Werte exp(0), exp(1), exp(2) und exp(3) für σ. In Tabelle 4.17 sind die Ergebnisse der Simulationen aufgeführt.

Die Ergebnisse lassen erkennen, dass die Kernaussagen aus Kapitel 4.1.4 nicht komplett übernommen werden können. Dort erwies sich vor allem die Imputation auf Basis einer Regression als stabil und zuverlässig – auch bei einer sehr hohen Varianz. Andere Methoden, wie beispielsweise die Mittelwertsimputation, waren dagegen für eine Varianzvariation sehr anfällig und mussten bei höherer Varianz Qualitätsverluste in Kauf nehmen.

Tab. 4.17: Grundszenario. Die Zahlen vermitteln wie oft das 'korrekte' Modell mit dem AIC für die entsprechende Varianz bei den y-Werten ausgewählt wurde

Methode	Sigma			
	exp(0)	exp(1)	exp(2)	exp(3)
Originaldaten	577	589	426	139
Complete Cases	573	581	375	102
wahre Gewichte	437	408	230	85
geschätzte Gewichte GAM	433	388	249	88
Imputation Mittelwert	715	321	454	162
Imputation Hot deck	478	148	304	115
Imputation Regression	277	479	376	145
Multiple Imputation 1	661	263	426	159
Multiple Imputation 2	701	558	431	141

In diesem Szenario ist durchweg eine Abnahme an Erfolg und Qualität mit steigender Varianz zu beobachten – unabhängig von der gewählten Methode. Dass eine bestimmte Methode für geringe beziehungsweise hohe Varianz größere Besonderheiten aufweist, kann nicht konstatiert werden.

Weitere Überlegungen

Analog zu den ersten Simulationen aus Kapitel 4.1 können verschiedenste Variationen untersucht werden. Im Wesentlichen sind Änderungen bezüglich der Verteilung der einzelnen Variablen und deren Parameter möglich. Detaillierte und spezifische Problemstellungen bezüglich dieser Sachlage wären noch möglich gewesen.

Auch hätten Variationen von Fehlwahrscheinlichkeitsfunktionen, sowie neue Kombinationen an Variablen mit fehlenden und nicht fehlenden Werten noch interessante Ergebnisse erbringen können. Ein Ausschöpfen aller Möglichkeiten wäre sicher zu Lasten der Übersichtlichkeit gegangen und hätte grundlegende Aussagen erschwert.

4.2.7 Resultate

Insgesamt lassen sich die Resultate aus Kapitel 4.2 differenzierter und detaillierter darstellen als in Kapitel 4.1.
Es hat sich bestätigt, dass in einer Vielzahl an Situationen die Imputationsmethoden weit bessere Ergebnisse als die gewichteten Methoden liefern. Oft war eine Verbesserung gegenüber der 'Complete Case Analysis' möglich.
Konnte kein Gütemaß eine richtige Verbesserung erzielen, so lag dies im Wesentlichen an einer Extremsituation, oder am Nichterkennen der Bernoulli-Variable.

Unter den Imputationsmethoden war je nach Situation eine andere am empfehlenswertesten. Oft wurden sehr gute Ergebnisse mit multiplen Imputationen erreicht, dagegen sehr stabile bei den Regressionsmethoden. Eine Mittelwertsimputation stellt in vielen Fällen eine sinnvolle Alternative dar.

Im Gegensatz zu Kapitel 4.1 ergaben sich nun verschiedene Muster bei den Resultaten von AIC und BIC. War anfangs das BIC dem AIC noch weit überlegen, so ergab erwartungsgemäß eine Zunahme an Variablen komplexere Ergebnisse.

4.3 Resümee bezüglich der Simulationen

Als Fragestellung zu Beginn des Kapitels wurde ausgegeben zu prüfen, ob der Umgang mit den fehlenden Daten sich auf die Modellwahl auswirkt, und ob für eine solche Situation ein gewichtetes Gütemaß oder die Imputation fehlender Werte zu bevorzugen ist.

Nach der Simulation verschiedenster Situationen kann nun zumindest eine teilweise eindeutige Aussage getroffen werden: Verglichen mit Aufwand und Ertrag, scheint vom Gebrauch gewichteter Gütemaße eher abzuraten zu sein.

In der Regel sollte eine Imputation bei fehlenden Werten eine Verbesserung gegenüber der 'Complete Case Analysis' erwirken können. Welche Methode hierfür jedoch verwendet werden sollte, kann nur vom jeweiligen Sachverhalt abhängig gemacht werden.

Fast alle Ergebnisse verdeutlichen noch einmal die Schwierigkeit und die Problematik einer Modellselektion. Selbst bei vollständigen Daten werden oft nur unbefriedigende Ergebnisse erzielt.
Alternative Methoden sollten im Rahmen komplexer Problemstellungen zumindest diskutiert werden.

5. Vergleich mit vorhergehenden Studien

Auch wenn es eine vergleichsweise hohe Anzahl an Veröffentlichungen bezüglich fehlender Daten und Modellselektion gibt, so ist Literatur für die Kombination beider Themen sehr dünn gesät. Eine Ausnahme bilden hier Niels Hens, Marc Aerts und Geert Molenberghs, die sich in einer ihrer Arbeiten [6] mit dieser Thematik beschäftigen. Ausschlaggebend für eine Analyse der Modellselektion unter der Problematik fehlender Daten war ein spezifisches Problem bei der Auswertung einer Studie zu Gebärmutterkrebs. Dabei wurden sowohl konkrete Daten und Modelle zur Analyse herangezogen, als auch versucht mit Hilfe einer Simulationsstudie die auftretenden Probleme zu klären.

5.1 Aufbau und Ergebnis der Simulation von Hens, Aerts und Molenberghs

Im Unterschied zu den Ansätzen dieser Arbeit ging es in den Studien von Hens et al. im Wesentlichen um die Effektivität eines gewichteten Gütemaßes. Dementsprechend diente als Entscheidungshilfe stets das gewichtete AIC, auch hier sowohl für bekannte, als auch geschätzte Gewichte.

In diesem Abschnitt sollen nun exemplarisch an einer ausgewählten Simulation die Ergebnisse der Studie erläutert werden. Sie ist in ihrer Grundkonzeption ähnlich zu denen dieser Arbeit.

Generiert wurde eine gleichverteilte Zufallsvariable x, eine Bernoulli(0.5)-verteilte Variable z, sowie eine Zielvariable y mit Mittelwert '$-3 + 3x + 5x^2$' und Varianz exp(5). Grund für die sehr spezielle Wahl an Variablen und Zusammenhängen stellt der Bezug zu der Gebärmutterkrebsstudie dar.

Im nächsten Schritt wurden Werte von x mit einer Wahrscheinlichkeit von

$$\pi(x, z) = 1 - [1 + \exp\{1 - 0.009(y - 300)\}]^{-1}$$

als fehlend deklariert. Dies führte zu einer sehr hohen Fehlwahrscheinlichkeit

bei hohen Werten und zu einer niedrigen Fehlwahrscheinlichkeit bei sehr kleinen Werten. Die geschätzten Gewichte lagen in der Regel im Intervall$[1, 4]$.

Anschließend wurden sowohl das Modell '$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz$' als auch alle Submodelle davon gefittet. Für 1000 verschiedene Samples $\{(x_i, z_i, y_i), i = 1, ..., n\}$ mit fixen $\{(x_i, z_i), i = 1, ..., n\}$ und einer Samplegröße von $n = 100$ wurden 4 verschiedene Strategien miteinander verglichen:

1. Gebrauch des AIC bei den Originaldaten
2. Gebrauch des AIC bei einer 'Complete Case Analysis'
3. Gebrauch des gewichteten AIC (wahre Gewichte)
4. Gebrauch des gewichteten AIC (geschätzte Gewichte mittels GAM)

Als 'richtiges' Modell bewerteten die Autoren dabei nicht nur das Modell mit x und x^2 als Einflussgröße, sondern auch alle Untermodelle davon, das heißt auch die Modelle die neben den bereits erwähnten Termen auch einen Interaktionseffekt enthalten.

In Tabelle 5.1 sind die Ergebnisse der Simulation zu erkennen.

Tab. 5.1: Ergebnisse einer Simulation von Hens et al.. Die Zahlen zeigen wie oft das jeweilige Gütekriterium ein entsprechendes Modell auswählt.

	1	x	z	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz	richtig erkannt
Originaldaten	0	114	0	666	31	18	106	65	837
Complete Cases	0	312	0	452	65	35	91	45	588
wahre Gewichte	0	199	0	371	67	61	129	173	673
gesch. Gewichte	0	228	0	416	70	56	110	121	647

Die Autoren wählten neben diesem Grundszenario noch eine Vielzahl an Variationen. So wurde sowohl die Sample-Größe variiert, als auch Varianz, quadratischer Effekt und fehlende Werte. Alle Variationen erbrachten jedoch in der Grundaussage ähnliche Ergebnisse.

Als interessanten Unterpunkt betrachten die Autoren des weiteren noch die Situation, bei der das 'richtige' Modell nicht in der Menge der ausgewählten Modelle enthalten ist. Für die detaillierten Ergebnisse sei auf [6] verwiesen.

5.2 Bewertung der Ergebnisse

Auf den ersten Blick scheinen die Ergebnisse den gewichteten Gütemaßen ein gutes Zeugnis auszustellen. Betrachtet man die korrekt klassifizierten Modelle, so erkennt erwartungsgemäß die 'Complete Case Analysis' gegenüber den wahren Daten deutlich weniger oft das richtige Modell. Das gewichtete AIC dagegen scheint tatsächlich eine Verbesserung erwirken zu können.

In der Interpretation ihrer Ergebnisse bewerten die Autoren die Resultate als äußerst positiv. So sei, wenn auch mit erheblichem Aufwand, eine Verbesserung gegenüber der 'Complete Case Analysis' zu erkennen. Diese sei zwar nicht sehr groß, in Anbetracht der großen Herausforderung der Thematik, jedoch sehr respektabel.

Bei der detaillierteren Betrachtung zeigt sich jedoch sofort der Schwachpunkt in der Argumentation der Autoren. Als Grundlage der Bewertung wird ausnahmslos auf die letzte Spalte der korrekt klassifizierten Modelle verwiesen. Diese wiederum setzt sich aus den drei Modellen zusammen, die die Variablen x und x^2 enthalten, aber auch entsprechende Interaktionsterme. Mit Sicherheit ist diese Vorgehensweise diskussionswürdig, eigentlich sollten die Parameterschätzungen für β_3 und β_4 Null sein, und damit die entsprechenden Variablen in das Modell nicht eingehen. Die Selbstverständlichkeit mit der diese Modelle als 'korrekt' gewertet werden ist nicht nachzuvollziehen.

Eine genauere Betrachtung der Ergebnisse liefert (für alle Szenarien) ein anderes Bild. Wird tatsächlich nur Modell (4), also '$y = \beta_0 + \beta_1 x + \beta_2 x^2$'als richtig gewertet, so kann das gewichtete AIC niemals eine Verbesserung gegenüber der 'Complete Case Analysis' erbringen. Das zunächst positive Resümee muss also revidiert werden.

Es scheint, als würde das gewichtete AIC seine Gewinne vor allem durch das Wählen der größeren, komplexeren Modelle mit den Interaktionstermen verbuchen. Vergleiche mit Kapitel 4 lassen dies als wenig verwunderlich erscheinen. Schon dort wählten die gewichteten Gütemaße sehr gern, und sehr oft die komplexesten der angebotenen Varianten aus.

Im Großen und Ganzen scheinen sich hier also noch einmal die Aussagen aus Kapitel 4 zu bestätigen: Es ist sehr schwer mit gewichtetem AIC (o.ä.) eine Verbesserung gegenüber der 'Complete Case Analysis' zu erwirken. Die Ergebnisse beim Verwenden eines gewichteten Gütemaßes sind weniger eindeutig, eine Entscheidung zugunsten eines Modells fällt schwerer. Daher werden oft auch falsche Entscheidungen getroffen.

6. Diskussion der Modellselektion

6.1 Grundüberlegungen

Aus statistischer Sicht kann eine Entscheidungsfindung mit Modellselektion zugunsten eines einzigen Modells durchaus als ambivalent angesehen werden. Stets ist zu beachten, dass Modelle für sich nie Anspruch nehmen eine Realität vollständig wiederzugeben, sondern eine Approximation an diese zu liefern. Eine Gütemaß ist hilfreich, um ein bestes Modell aus einer Menge von Kandidatenmodellen auszuwählen. Beschreiben alle Modelle aus dieser Menge die Daten jedoch verhältnismäßig schlecht, so wird dennoch 'das beste unter den schlechten' ausgewählt. Daher sollte sichergestellt werden, dass sich in den zur Verfügung stehenden Modellen eine adäquate Auswahl befindet.

Es stellt sich die Frage, ob ein Modell in erster Linie zur Beschreibung und Interpretation eines Vorgangs oder Prozesses betrachtet wird, oder ob die Prädiktion unbekannter Untersuchungseinheiten im Vordergrund stehen soll.

Für den ersten Fall der Modellierung und Interpretation eines Zusammenhangs, ist die Reduzierung auf ein fixes Modell durchaus sinnvoll. Nur so lassen sich generelle Grundaussagen treffen. Maße zur Güte eines Modells, sowie weitere Kennziffern, die Schwierigkeiten innerhalb eines Modells beschreiben, bieten dem Statistiker die Möglichkeit einer Abwägung seiner Entscheidung und Interpretation. Oft sollen Zusammenhänge 'im Trend' oder 'im Mittel' erkannt werden, eine detailliertere Aussage ist hier nicht von Nöten.

Um Parameter zu interpretieren, und Zusammenhänge zu quantifizieren, bietet es sich also an auf ein fixes Modell festzulegen. Um auf dieses eine, gewünschte Modell zu kommen, bedient man sich in der Regel der Modellselektion. Diese besitzt jedoch den Schwachpunkt, dass die Relation der Unterschiede für die einzelnen Modelle in keinster Weise gewürdigt wird. Besitzt beispielsweise ein Modell A einen AIC von 610 und ein anderes Modell B einen AIC von 610.1, so wird Modell A verworfen, ähnlich als hätte es einen verschwindend geringen AIC. Eine Abstufung unter den einzelnen Modellen findet also nur in vereinfachter Form, ohne Beachtung der Relation statt. Soll eine Entscheidung zugunsten eines einzigen Modells getroffen werden, so

ist diese Problematik natürlich nicht zu verdrängen, man hat jedoch keine andere Wahl, als dies zu akzeptieren.

Anders stellt sich der Fall der Prädiktion dar. Nicht die Interpretierbarkeit einzelner Parameter steht dabei im Vordergrund, sondern die möglichst genau Vorhersage. Somit könnten Alternativen unter dieser Zielsetzung durchaus ihre Anwendung finden. Eine stupide Selektion zu Gunsten eines (möglicherweise falschen) Modells könnte vermieden werden, wenn die Erklärungskraft der einzelnen Modelle berücksichtigt wird.

6.2 Multimodel Inference

Die Werte des AIC an sich sind als solche nicht zu interpretieren, da sie willkürliche Konstanten enthalten und stark von der Stichprobengröße beeinflusst werden [4]. Es bietet sich daher an, R verschiedene Modelle an der Differenz

$$\Delta_i = AIC_i - AIC_{min}, \qquad i = 1, ..., R$$

zu beurteilen. Dabei bezeichnet AIC_{min} das Minimum unter den R verschiedenen AIC_i-Werten. Diese Transformation bewirkt in erster Linie, dass das beste unter den R Modellen einen Wert von $\Delta = 0$ annimmt, während die restlichen Modelle positive Werte besitzen.

Die Δ_i sind nun aussagekräftig und einfach zu interpretieren: Je größer Δ_i ist, desto geringer ist die Plausibilität, dass das Modell i die beste Annäherung unter den R Modellen erbringt.

Burnham und Anderson schlagen in einer ihrer Arbeiten [4] vor, sich an folgenden Daumenregeln zu orientieren:

$\Delta_i \leq 2$	$\rightarrow$	erhebliche Unterstützung des Modells
$4 \leq \Delta_i \leq 7$	$\rightarrow$	wenig Unterstützung des Modells
$\Delta_i > 10$	$\rightarrow$	im Wesentlichen keine Unterstützung des Modells

Akaikes Gewichte

Um den Einfluss eines Modells zu quantifizieren, können die so genannten 'Akaike-Gewichte' definiert werden [4]:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)}$$

Der Zähler kann dabei als Likelihood des Modells i (unter gegebenen Daten) interpretiert werden, da

$$L(g_i|\text{Daten}) \propto \exp\Big(-\frac{1}{2}\Delta_i\Big)$$

gilt. Somit kann das Gewicht eines jeden Modells als die Wahrscheinlichkeit interpretiert werden, mit der dieses Modell unter Kullback-Leibler als das Beste eingestuft wird. Die Summe aller Gewichte ergibt trivialerweise Eins.

Der Gebrauch der Gewichte ermöglicht nun ein 'Model averageing'. D.h. unter der Zielsetzung der Prädiktion (siehe auch Kapitel 6.1) kann für den Schätzer des Vorhersagewertes $\hat{y}$ folgende Alternative vorgeschlagen werden:

$$\hat{\bar{y}} = \sum_{i=1}^{R} w_i \hat{y}_i$$

Der Vorhersagewert eines jeden Modells i wird also gewichtet in die Gesamtschätzung der Vorhersage miteinbezogen.

Wichtig ist hier die Unterscheidung, ob unter den R Kandidatenmodellen sämtliche möglichen Modelle zu finden sind, oder ob bereits eine Vorauswahl getroffen wurde. Jeder Sachverhalt lässt hier eine andere Vorgehensweise oder Philosophie zu.

Ergänzend sei auch noch darauf hingewiesen, dass die Idee eines (Akaike-) gewichteten Schätzers selbstverständlich auch außerhalb des Spezialfalls der Prädiktion angewandt werden kann. Für jeden Parameter θ, der in allen R Modellen vorkommt, kann der Schätzer

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

definiert werden. Jedes θ_i steht dabei für den Parameterschätzer basierend auf dem Modell i.

6.3 Eine Simulation

Analog zu Kapitel 4.2.5 soll die Situation betrachtet werden, bei der zwei von drei Einflussgrößen einen Einfluss besitzen (nämlich x und z) und bei allen drei Variablen fehlende Werte vorliegen. Alle Einstellungen aus Kapitel 4.2.5 wurden übernommen, die Werte für Varianz, Fehlwahrscheinlichkeitsfunktionen etc. stimmen überein.

Zusätzlich soll nun der Ansatz der 'Multimodel Inference' untersucht werden. Dazu wurden für alle 1000 Samples jeweils zwei Möglichkeiten verglichen:

1. Vorhersagewert über *ein* Modell, also das mit dem geringsten AIC
2. Vorhersagewert über die gewichtete Schätzung *aller* 18 Modelle nach Akaike

Als Gütemaß für den Vorhersagewert wurde die durchschnittliche betragliche Abweichung zwischen wahren und vorhergesagten Werten betrachtet, also:

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Der Wert von $\bar{\xi}$ kann dabei als durchschnittliche Abweichung von wahren und vorhergesagten y-Werten interpretiert werden. Angewandt auf alle N Samples, lässt sich die Idee wie folgt verallgemeinern:

$$\bar{\bar{\xi}} = \frac{1}{N}\sum_{i=1}^{N}\bar{\xi}_i$$

Je geringer also ein Wert von $\bar{\bar{\xi}}$ ist, desto besser ist die Güte der Vorhersagemethode.

Als Fragestellung bietet es sich nun also an zu untersuchen, ob eine Vorhersage nach Akaike eine Verbesserung erwirken kann, und ob der Umgang mit den fehlenden Daten ebenfalls in diese Entscheidung mit hineinspielt.

Um einen ersten Eindruck über die Ergebnisse der Simulation zu erhalten, kann ein Blick auf die durchschnittlichen Gewichte nach Akaike erste Rückschlüsse liefern. In Tabelle 6.1 sind die Ergebnisse der Simulation zu erkennen. Je Modell (siehe auch Tabelle 4.9) und Methode wurde das durchschnittliche Gewicht nach Akaike aufgelistet, aufgrund von Rundungsfehlern addieren sich die Gewichte nicht genau zu Eins auf.

Der Trend aus Kapitel 4 ist auch hier deutlich zu erkennen. Die Relation unter Akaikes Gewichten kann durchaus ein Hinweis für die Evidenz eines Modells sein. Abgesehen von den gewichteten Gütemaßen wird das richtige Modell (13), sowie das falsche Modell (18) mit x als einziger Einflussgröße durchgehend stark gewichtet, in der Regel mit einem Wert zwischen 0.15 und 0.20. Insofern spiegeln die Gewichte hier die Ergebnisse aus Kapitel 4 wieder. Auch die von v und z gemeinsam dominierten Modelle werden gering bis gar nicht gewichtet.

Tab. 6.1: Die Zahlen vermitteln die einzelnen Akaike Gewichte (in%) je Modell und Methode.

Methode	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AIC	3.10	3.18	4.54	4.61	4.61	7.44	6.63	7.36	10.67
AIC CC	3.70	3.33	4.16	4.42	4.51	6.28	5.87	5.76	8.28
AIC Gewichte wahr	22.84	8.84	7.74	7.32	4.24	5.89	6.20	5.89	4.55
AIC Gewichte geschätzt	21.42	8.65	7.59	7.43	6.99	6.17	6.28	5.96	4.91
AIC Mittelwert	1.00	1.64	2.68	2.93	3.19	5.72	4.86	5.14	9.36
AIC Hot deck	2.69	2.61	3.84	3.79	3.62	5.55	5.57	5.32	8.10
AIC Regression	2.32	2.69	3.89	4.25	4.11	6.59	6.05	5.86	9.44
AIC Multiple Imp. 1	1.13	1.78	2.70	3.17	3.28	5.81	5.03	4.93	9.22
AIC Multiple Imp. 2	2.57	2.64	4.07	4.22	3.79	5.93	6.26	5.76	8.76
Methode	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
AIC	11.50	3.46	0.00	18.40	5.35	0.00	0.00	0.00	9.14
AIC CC	10.13	5.90	0.00	14.38	8.52	0.00	0.00	0.00	14.77
AIC Gewichte wahr	7.38	4.03	0.00	5.33	3.12	0.00	0.00	0.00	3.62
AIC Gewichte geschätzt	7.64	4.04	0.00	5.99	3.16	0.00	0.00	0.00	3.77
AIC Mittelwert	8.83	6.04	0.00	17.12	10.90	0.00	0.00	0.00	20.59
AIC Hot deck	9.34	6.97	0.00	13.62	10.62	0.00	0.00	0.00	18.37
AIC Regression	9.31	6.03	0.00	14.74	9.75	0.00	0.00	0.00	14.96
AIC Multiple Imp. 1	8.99	5.88	0.00	16.72	11.10	0.00	0.00	0.00	20.27
AIC Multiple Imp. 2	10.61	5.88	0.00	14.85	9.01	0.00	0.00	0.00	15.63

Die gewichteten Gütemaße erweisen sich erneut als Schwachpunkt. Die Tendenz zu großen Modellen ist an den Gewichten gut zu erkennen. Das Modell mit allen Einflussgrößen und Interaktionen wird am stärksten gewichtet ($w_1 = 0.2284$). Das richtige Modell (13) dagegen unterscheidet sich im Einfluss kaum von den anderen.

In Abbildung 6.1 wird die Situation noch einmal grafisch verdeutlicht. Für die vollständigen Daten, einer multiplen Hot Deck Imputation und dem gewichteten AIC sind hier noch einmal die Gewichte nach Akaike zu erkennen.

Entscheidend sind nun aber die Vorhersagewerte abhängig von der Methodik. Sollte ein gewichteter Vorhersagewert tatsächlich eine Verbesserung bezüglich der Prädiktion darstellen, so sollte eine signifikante Abweichung der $\bar{\bar{\xi}}$ im Verhältnis zu einer gewöhnlichen Voraussage nachzuweisen sein. In Tabelle 6.2 sind die Ergebnisse der Simulation im Hinblick auf diese Fragestellung aufgelistet.

Sofort zu erkennen ist, dass unabhängig vom Umgang mit den fehlenden Daten der Ansatz der 'Multimodel Inference' eine Verbesserung in der Prädiktion erwirkt. Die letzte Spalte verdeutlicht, dass diese Unterschiede auch signifikant zum 1%-Niveau sind. Zumindest in diesem Szenario scheinen die Auswirkungen von fehlenden Daten auf die Aussage der Güte einer gewichteten Vorhersage keinen Einfluss zu nehmen. Eine Verbesserung ist in allen Fällen zu konstatieren.

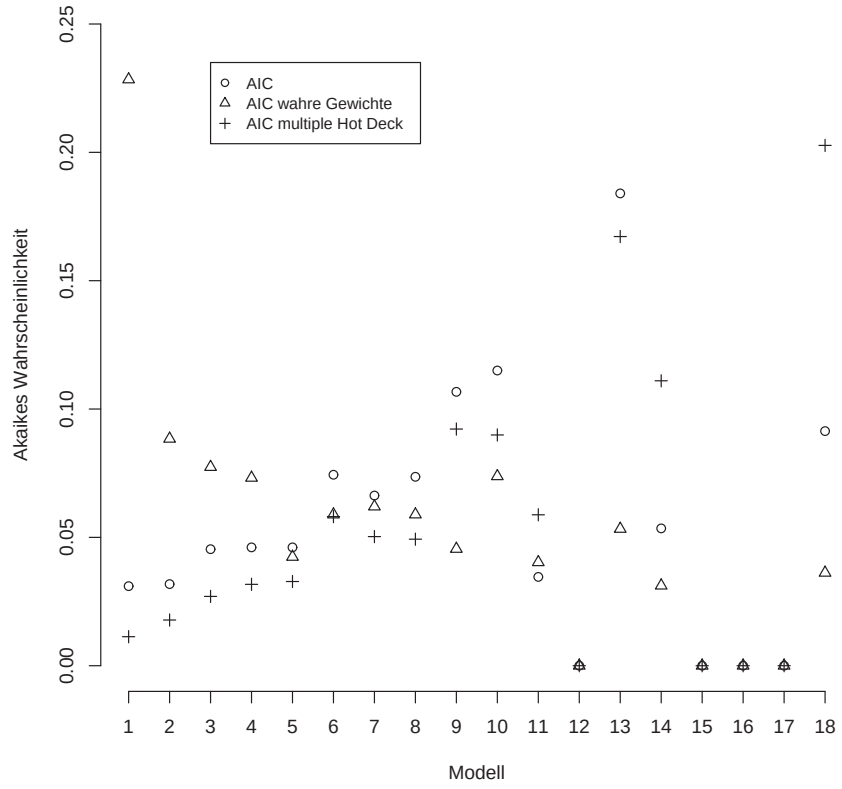

Abb. 6.1: Akaikes Fehlwahrscheinlichkeiten je Modell und Methode.

Als interessant erweist sich eine genauere Betrachtung der Werte. Führt man sich die Tatsache zu Augen, dass die y-Werte in der Regel im Intervall $[-10, 60]$ streuen, so lassen sich die Zahlen gut einordnen. Eine durchschnittliche mittlere Abweichung von 5.72 ($\bar{\bar{\xi}}_{18}$) bei den vollständigen Daten kann durchaus als akzeptabel eingestuft werden. Mit Sicherheit kann in spezifischen Problemstellungen aber auch dieser Wert noch zu hoch sein.

Interessant ist auch die Tatsache, dass die durchschnittliche Abweichung zwischen wahren und vorhergesagten Werten bei den gewichteten Gütemaßen sehr gering ist. Dies ist mit der Tatsache zu erklären, dass die Anzahl der in den Wert von $\bar{\bar{\xi}}$ eingehenden Beobachtungen geringer ist als bei den Imputationsmethoden, und daher die betragsmässig höchsten Anteile des Wertes – nämlich diejenigen zwischen imputierten Werten und deren Vorhersagen –

Tab. 6.2: Werte für $\bar{\bar{\xi}}$ je Methode - einmal für die Vorhersage mit einem Modell ($\bar{\bar{\xi}}_1$) und einmal für die Vorhersage über die Gewichtung aller 18 Modelle ($\bar{\bar{\xi}}_{18}$).

Methode	$\bar{\bar{\xi}}_{18}$	$\bar{\bar{\xi}}_1$	Differenz	Sig.
AIC	5.71648	5.94985	0.23337	$***$
AIC CC	4.93048	5.00577	0.07169	$***$
AIC Gewichte wahr	4.82147	4.83712	0.01565	$***$
AIC Gewichte gesch.	4.81597	4.82936	0.01339	$***$
AIC Mittelwert	8.21684	8.27912	0.06228	$***$
AIC Hot deck	8.95528	9.02297	0.06769	$***$
AIC Regression	5.21642	5.25031	0.03389	$***$
AIC multiple Imp. 1	8.36841	8.43233	0.06392	$***$
AIC multiple Imp. 2	6.43444	6.47399	0.03955	$***$

komplett fehlen.

In diesem Szenario wurde der Extremfall modelliert, dass *keine* Vorauswahl unter den Modellen getroffen wurde. Trotzdem konnten bereits signifikante Unterschiede festgestellt werden. Es lässt sich erahnen, dass eine Reduzierung der Modellkandidaten erneut eine Verbesserung erbracht hätte. Eine solche Vorauswahl hätte mit Sicherheit auch den Werten von Δ_i getroffen werden können. Ein Grenzwert trifft dann die Entscheidung, ob die Vorhersage eines Modells in die Gewichtsschätzung eingeht oder nicht.

6.4 Schlussfolgerungen

Die Grundaussagen des vierten Kapitels haben sich auch hier bestätigt. Imputationsmethoden erbringen eine Verbesserung gegenüber der 'Complete Case Analysis' und sind gewichteten Gütekriterien vorzuziehen.

Auch erscheinen unter der Zielsetzung der Prädiktion gewichtete Vorhersageschätzungen eine sinnvolle Alternative zu sein. Unabhängig vom Umgang mit den fehlenden Daten konnten im beobachteten Szenario Verbesserungen in der Vorhersage erzielt werden. Diese waren klein, jedoch signifikant. Ob der Aufwand einer solchen Schätzung mit dem Erfolg in Relation steht ist jedoch fragwürdig und eine endgültige Entscheidung muss der Anwender selbst treffen.

7. Ein Datenbeispiel

7.1 Problemstellung

In einer Schweizer Studie sollten die Ausgaben (in Schweizer Franken) der Bürger der Stadt Basel für Kulturelles im Allgemeinen, sowie Theaterbesuche im Speziellen, genauer untersucht werden. Es wurden folgende Merkmale erhoben:

'Ausgaben Theater'	→	Ausgaben für Theaterbesuche im laufenden Jahr
'Ausgaben Kultur'	→	Ausgaben für Kultur im laufenden Jahr
'Geschlecht'	→	Das Geschlecht der Befragten
'Alter'	→	Das Alter der Befragten
'Einkommen'	→	Das aktuelle Gehalt der Befragten

Mit Hilfe einer Regressionsanalyse sollte in erster Linie erörtert werden, welche Komponenten die jährlichen Theaterausgaben der Basler Bürger beeinflussen. Der Stichprobenumfang lag bei $n = 699$.

Untersucht werden soll nun, welche der drei möglichen Einflussgrößen ('Geschlecht', 'Alter', 'Gehalt') einen Einfluss auf die Theaterausgaben besitzen. Passenderweise ist die Situation der Variablen fast analog zu denen des Szenarios aus Kapitel 4.2. Auch dort gab es drei Einflussvariablen, und auch dort waren zwei davon metrisch und eine binär.

Um die Problematik der Imputation fehlender Daten, sowie wichtige Aspekte der Modellselektion adäquat erörtern zu können, wurden bei allen drei Einflussgrößen Werte künstlich als fehlend deklariert. Die Konstruktion der Fehlwahrscheinlichkeitsfunktionen unterlag dabei keinen Annahmen aus der Realität, im Wesentlichen sollten Ideen der ersten sechs Kapitel adaptiert werden.

Im folgenden soll die Variable 'Geschlecht' auch mit 'G' bezeichnet werden, die Variable 'Einkommen' mit 'E', das Alter mit 'A', sowie die Theaterausgaben mit 'T'. Die Fehlwahrscheinlichkeitsfunktionen der drei Variablen lauten

wie folgt:

$$\begin{aligned}
\pi(G,E,A)_G &= 1 - \frac{1}{1+\exp(1-0.08\cdot(y-100))} \\
\pi(G,E,A)_E &= 1 - \frac{1}{0.0005\cdot(y-140)^2+1} \\
\pi(G,E,A)_A &= 1 - \frac{1}{0.00005\cdot(y-80)^2+1}
\end{aligned}$$

In Abbildung 7.1 sind die entsprechenden Fehlwahrscheinlichkeitsfunktionen noch einmal abgebildet.

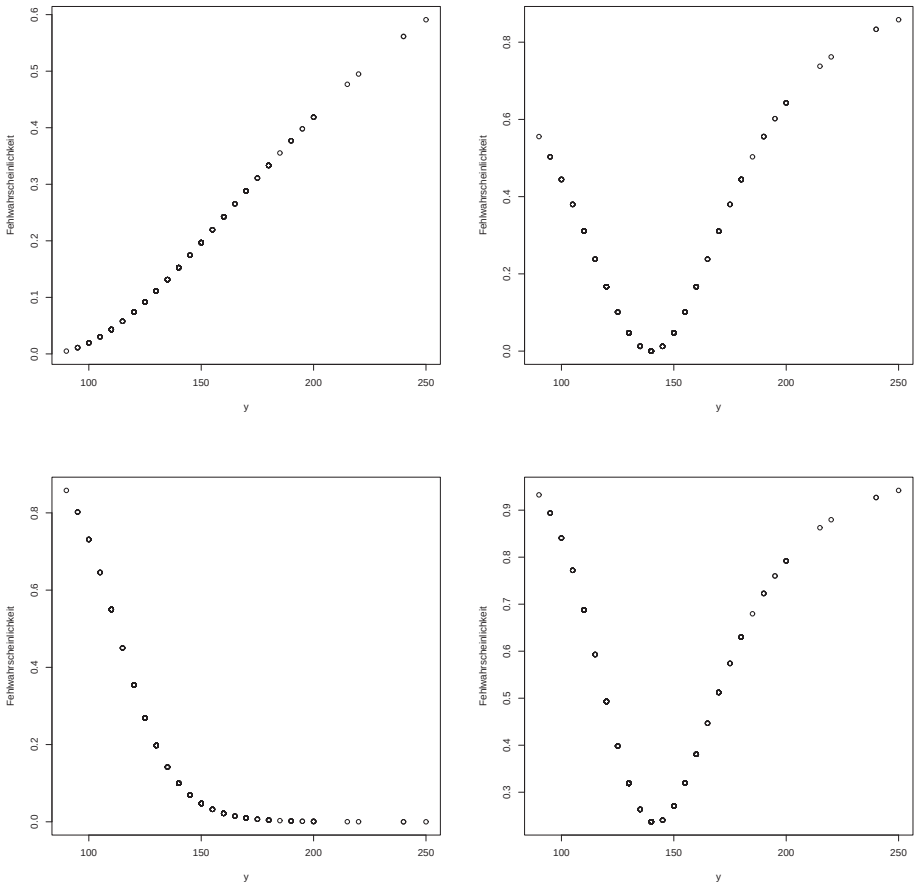

Abb. 7.1: Fehlwahrscheinlichkeiten für das Alter (oben links), Einkommen (oben rechts), Geschlecht (unten links) sowie insgesamt (unten rechts)

Auch die Funktion für das Fehlen eines Wertes in der X-Matrix insgesamt ($\pi(G, E, A)_{G \cup E \cup A}$) ist dort zu erkennen. Die Anzahl der Punkte aller Grafiken in Abbildung 7.1 wirkt im Verhältnis zum Stichprobenumfang insofern gering, als dass alle Merkmale der Untersuchung quasi-stetig erfasst und notiert wurden. Die Theaterausgaben – jeweils zu sehen auf der x-Achse – wurden nur in 5 SFR-Schritten erfasst, daher ist die Anzahl der verschiedenartigen Ausprägungen entsprechend gering.

Analog zum bisherigen Vorgehen sollen 18 Modelle gefittet werden (siehe auch Tabelle 7.1), und für jedes Modell und jede Methode jeweils der AIC und das zugehörige Akaikegewicht berechnet werden. Zu untersuchen ist, ob die Wahl der Imputationsmethode die Entscheidung zugunsten eines Modells beeinflusst, und ob diese Entscheidung eindeutig ausfällt oder nicht.

Tab. 7.1: Die 18 zur Wahl stehenden Modelle für die Daten aus Basel

	gefittete Modelle		
(1)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 GE + \beta_6 AG + \beta_7 AGE$		
(2)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 GE + \beta_6 AG$		
(3)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 AE$		
(4)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 GE$		
(5)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AE + \beta_5 GA$		
(6)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 GE$		
(7)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG$		
(8)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AE$		
(9)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G$	(10)	$y = \beta_0 + \beta_1 A + \beta_2 G + \beta_3 AG$
(11)	$y = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 AE$	(12)	$y = \beta_0 + \beta_1 E + \beta_2 G + \beta_3 EG$
(13)	$y = \beta_0 + \beta_1 A + \beta_2 G$	(14)	$y = \beta_0 + \beta_1 A + \beta_2 E$
(15)	$y = \beta_0 + \beta_1 E + \beta_2 G$	(16)	$y = \beta_0 + \beta_1 G$
(17)	$y = \beta_0 + \beta_1 E$	(18)	$y = \beta_0 + \beta_1 A$

7.2 Ergebnisse der Auswertung

Die Ergebnisse der Auswertung sind in Tabelle 7.2 zusammengefasst. Die Zeilen beschreiben dabei die verschiedenen Methoden im Umgang mit den fehlenden Daten (analog zu Kapitel 4.2), in den Spalten sind die drei favorisierten Modelle der jeweiligen Methode (also diejenigen mit dem geringsten AIC) und deren Gewicht nach Akaike angegeben.

Tab. 7.2: Favorisierte Modelle für die Daten, in Klammern stehen die dazugehörigen Gewichte nach Akaike

	Präferenz unter den Modellen					
Methode	Rang 1		Rang 2		Rang 3	
AIC	13	(0.1772)	18	(0.1614)	10	(0.1028)
AIC CC	13	(0.1524)	18	(0.1327)	14	(0.0931)
AIC Gew. wahr	1	(0.1819)	3	(0.1368)	8	(0.1113)
AIC Gew. gesch.	1	(0.1845)	3	(0.1363)	8	(0.1086)
AIC Mittelwert	18	(0.2460)	13	(0.1540)	14	(0.1184)
AIC Hot deck	18	(0.2059)	13	(0.1445)	14	(0.1252)
AIC Regression	18	(0.1914)	13	(0.1413)	14	(0.1351)
AIC Multipel 1	18	(0.2148)	13	(0.1447)	14	(0.1245)
AIC Multipel 2	18	(0.1869)	13	(0.1782)	8	(0.0977)

Betrachtet man nun die Resultate für den vollständigen Datensatz, so ist zu erkennen, dass Modell (13) ausgewählt wird. Dies ist das Modell mit 'Alter' und 'Geschlecht' als Einflussgrößen – ohne Interaktionen. Ebenfalls zu sehen ist jedoch auch, dass die Wahl für dieses Modell keinesfalls eindeutig ausfiel. Die Gewichte nach Akaike liegen unter den ersten drei Modellen sehr nah beisammen, speziell bei den ersten beiden. Das Modell mit dem zweitgeringsten AIC, nämlich Modell (18), wählt nur das Alter als Einflussgröße aus, Modell (10) – welches an dritter Stelle steht – erkennt dagegen noch die Interaktion von Geschlecht und Alter als signifikant an.

Burnham und Anderson [4] schlagen vor, nur Modelle mit einem Gewicht von etwa 0.9 als das eindeutig beste anzusehen, jedoch immer unter der Prämisse einer bereits getroffenen Vorauswahl der Modelle. Da hier alle möglichen Modelle in Erwägung gezogen wurden, ist diese Zahl sicherlich abzuschwächen, die Grundaussage, dass die Wahl für ein Modell uneindeutig ist, kann aber sicherlich nicht in Frage gestellt werden.

Beim Betrachten der 'Complete Cases' ist im Resultat keine große Veränderung festzustellen. Es ist jedoch anzumerken, dass insgesamt 43.92% der Fälle verworfen wurden. Dass unter diesen Umständen immer noch sehr ähnliche Aussagen im Vergleich zum vollen Datensatz getroffen werden, ist als positiv einzustufen.

Ein erster Blick auf die Methoden im Umgang mit fehlenden Daten scheint erneut die Kernthesen der ersten sechs Kapitel zu bestätigen.

Wählt man den Ansatz eines gewichteten Gütemaßes, so ist erneut der Hang zu komplexen, großen Modellen zu erkennen. Modell (1), welches alle drei Einflussgrößen und Interaktionen enthält, wird eindeutig als das Beste eingestuft. Dies kann mit Sicherheit als Fehlentscheidung bezeichnet werden. Dass des weiteren noch Modell (3), das ebenfalls sehr groß ist, als sehr evident bezeichnet wird, bestätigt die bisher erarbeiteten Hypothesen noch einmal.

Alle Imputationsmethoden dagegen liefern nicht nur nahezu identische Ergebnisse, sondern auch solche, die in etwa denen der vollständigen Daten entsprechen. In letzter Konsequenz wird hier zwar stets Modell (18) dem Modell (13) vorgezogen, dass das Alter einen Einfluss auf die Theaterausgaben haben sollte, kann aber auch hier zweifelsohne festgestellt werden. Dass beim Betrachten eines Endmodells dem 'Geschlecht' der Einfluss aberkannt wird, muss natürlich trotzdem kritisch bemerkt werden.

Es hat sich also gezeigt, dass die in dieser Arbeit anhand von Simulationen erarbeiteten Hypothesen durchaus auch auf einen realen Datensatz übertragen werden können. Fehlen im Zuge einer Datenanalyse viele Daten, so können häufig mit Imputationen die besten Ergebnisse erzielt werden; welche Methodik in welcher Situation zu empfehlen ist, hängt jedoch stark von der Struktur der Daten ab.

8. Abschließende Bewertung und Ausblick

Abschließend lässt sich festhalten, dass für die Modellselektion unter der Problematik fehlender Daten eine Reihe an Möglichkeiten bereit steht. Ein Großteil der in dieser Arbeit durchgeführten Simulationen stärkt die These, dass Imputationsmethoden gegenüber einer einfachen 'Complete Case Analysis' Verbesserungen erzielen können. Wie stark diese ausfallen und welche Methoden welchen Gewinn versprechen hängt stark vom jeweiligen Sachverhalt ab. Mitunter kann es in Extremsituationen sogar vorkommen, dass überhaupt keine Verbesserung festzustellen ist. In der Regel ist dies auf hohe Korrelationen, hohe Varianzen oder spezifische Problemstellungen zurückzuführen.

Die multiplen Imputationsmethoden lieferten im Wesentlichen recht gute und stabile Ergebnisse. Da Sie nur für Extremsituationen anfällig waren, kann zu Ihrem Gebrauch geraten werden. Auch Imputationen auf Basis einer Hilfsregression lieferten in der Regel sehr stabile Resultate, es bleibt jedoch darauf zu verweisen, dass sich in der Praxis die Wahl der Einflussgrößen als sehr schwierig gestalten kann und Abweichungen von den überwiegend guten Ergebnissen der Simulationen zu erwarten sind.

Die recht neue Methodik der gewichteten Gütemaße besticht durch ihre Originalität, ihre Ergebnisse erweisen sich im Allgemeinen jedoch als sehr dürftig. Nur sehr selten konnten Verbesserungen gegenüber einer 'Complete Case Analysis' beobachtet werden. Ein Hang zum Auswählen komplexer und großer Modelle konnte nahezu ausnahmslos beobachtet werden und erklärt diesbezüglich die teilweise positiven Ergebnisse von Hens et. al [6]. Auch ist die Schätzung der hierfür nötigen Fehlwahrscheinlichkeiten nicht ohne Probleme zu bewältigen, eine Grundkenntnis über die Datensituation sollte vorhanden sein.

Im Vergleich der einzelnen Gütemaße (AIC, BIC, C_p) lässt sich festhalten, dass unabhängig von der gewählten Situation Akaikes Informationskriterium nahezu identische Resultate im Vergleich zu Mallows C_p liefert. Aufgrund

ihrer grundsätzlich verschiedenen Konzeption war dies in diesem Ausmaß sicherlich nicht zu erwarten.
Anders gestaltet sich die Situation des BIC. In einem Großteil der betrachteten Situationen ergab sich ein ähnliches Grundmuster in Bezug auf den AIC, ob die Tendenz des BIC kleinere Modelle zu bevorzugen jedoch Erfolg versprach, hing selbstverständlich stark von der Datensituation ab. In den Kapiteln 4.2.4 und 4.2.5 war Diskrepanz der Ergebnisse von AIC und BIC sogar sehr hoch. Eine Entscheidung zugunsten eines Gütemaßes hängt sicher stark von der philosophischen Betrachtung des Anwenders und der speziellen Problematik eines Datensatzes ab.

Ausblick

Im Zuge dieser Arbeit wurden insofern recht einfach gehaltene Situationen betrachtet, als dass die Anzahl der Einflussgrößen bei maximal drei lag und komplexeren Situationen transformierter Variablen komplett die Betrachtung verweigert wurde. Eine nächste Herausforderung, auch unter Berücksichtigung der Programmierung, stellt sicherlich die Erörterung von Situationen mit einer Vielzahl an Einflussgrößen dar. Es stellt sich dann die Frage, ob hierfür tatsächlich eine Entscheidung unter *allen* möglichen Modellen gesucht werden soll, oder ob bereits eine Vorauswahl getroffen wird. In diese Thematik spielt mit Sicherheit auch die Betrachtung quadratischer, logarithmierter oder anders transformierter Größen hinein. Als interessant erweist sich die Betrachtung des Aspekts, bei der das 'wahre' Modell nicht unter den Kandidatenmodellen zu finden ist. Alternative Lösungsansätze könnten hier sicherlich noch entwickelt werden.

Der Entscheidungsfindungsprozess bei einer Modellwahl unter Berücksichtigung der Problematik fehlender Daten wurde bisher nur anhand linearer Modelle betrachtet. Ein nächster Schwerpunkt könnte die Überprüfung der bisher erarbeiteten Hypothesen bei generalisierten linearen Modellen sein. Ein weites Spektrum neuer Gesichtspunkte sollte sich hierbei eröffnen. Ob ähnliche Resultate zu erwarten sind müssen zukünftige Arbeiten zeigen.

Anhang

A. Simulationsübersicht

Kapitel	Anzahl Variablen	Zielgröße	Einflussgrößen	Werte fehlen bei	Einfluss auf y haben	gefittete Modelle	Besonderheiten
4.1.1	3	y	x,z	x	x	4	Grundszenario
4.1.2	3	y	x,z	x	x	4	Betrachtung BIC, C_p
4.1.3	3	y	x,z	x	x	4	Variation der Fehlwahrscheinlichkeitsfunktion bei x
4.1.4	3	y	x,z	x	x	4	Variation der Varianz bei y
4.1.5	3	y	x,z	x	x	4	Variation bei der Verteilung von z
4.1.6	3	y	x,z	z	x	4	Werte fehlen nun bei z
4.1.7	3	y	x,z	x	x	4	Hohe Korrelation unter x und z
4.2.1	4	y	v,x,z	x	x,z	18	Grundszenario
4.2.2	4	y	v,x,z	x	x,z	18	Variation bei der Verteilung von z
4.2.3	4	y	v,x,z	x	x,z	18	Untersuchung der GAM
4.2.4	4	y	v,x,z	x,v	x,z	18	Werte fehlen auch bei v
4.2.5	4	y	v,x,z	x,v,z	x,z	18	Werte fehlen auch bei v und z
4.2.6	4	y	v,x,z	x	x,z	18	1) Korrelationen unter x und v 2) Variation der Varianz bei y
6.3	4	y	v,x,z	x,v,z	x,z	18	Untersuchung von 'Multimodel Inference'

B. Das griechische Alphabet

α	A	Alpha	ν	N	Ny
β	B	Beta	ξ	Ξ	Xi
γ	Γ	Gamma	o	O	Omikron
δ	Δ	Delta	π	Π	Pi
ϵ	E	Epsilon	ρ	P	Rho
ζ	Z	Zeta	σ	Σ	Sigma
η	H	Eta	τ	T	Tau
θ	Θ	Theta	υ	Υ	Ypsilon
ι	I	Iota	ϕ	Φ	Phi
κ	K	Kappa	χ	X	Chi
λ	Λ	Lambda	ψ	Ψ	Psi
μ	M	My	ω	Ω	Omega

Literaturverzeichnis

[1] Agostinelli, C., 2002, *Robust model selection in regression via weighted likelihood methdology*: Statistics and Probability Letters, 64(4):583-639

[2] Akaike, H., 1974, *A new look at the statistical model identification*: IEEE Trans. on Automatic Control, 19(6):716– 723

[3] Amemiya, T., 1985, *Advanced Econometrics*: Blackwell Verlag

[4] Burnham, K., Anderson, D., 2002, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.*: Springer-Verlag

[5] Hastie, T., Tibsharani, R., 1990, *Generalised additive models*: Chapman and Hall, London

[6] Hens, N., Aerts, M., Molenberghs G., *Model selection for incomplete and design based samples*: IAP Statistics Network

[7] Kobayashi, M., Sakata, S., 1990, *Mallows C_p Criterion and Unbiasedness of Model selection*: Journal of Econometrics, 45:385-395

[8] Kullback, S., Leibler, R., 1951, *On information and sufficiency*: Annals of Mathematical Statistics, 22(1):79-86

[9] Little R., Rubin D., 1976, *Statistical Analysis with missing Data*: John Wiley & Sons

[10] Mallows, C.L., 1975, *Some comments on Cp*: Technometrics, 15:661

[11] Rueger, B., 1999, *Test- und Schaetztheorie (Band I)*: Oldenbourg Verlag

[12] Shannon, E., 1948, *A mathematical theory of evidence*: Bellsyt. Techn. Journal, 27:379-423, 623-653

[13] Schwarz, G., 1978, *Estimating the dimension of a model*: Annals of Statistics, 6(2):461-464

[14] Schwarz, Wallace and Rissanen, 2000, *Intertwining Themes in Theories of Model selection*: Available at University of Illinois

[15] Toutenburg, H., 2003, *Lineare Modelle, 2. Auflage*: Physica Verlag

[16] Tutz, G., 2000, *Die Analyse kategorialer Daten*: Oldenbourg Verlag

[17] Wiener, N., 1948, *Cybernetics or control and communication in the animal and the machine*, Technical Report 1053, Act. Sci. Indust., Hermann et Cie.

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 4. Oktober 2006

(Michael Schomaker)