Behandlung fehlender Daten

Diplomarbeit

zur Erlangung des Grades
eines Diplom-Volkswirtes

an der Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

vorgelegt von

Lars Rohrschneider

(Matrikel-Nr. 176715)

Betreuer:	**Dr. Sigbert Klinke**
Erstgutachter:	**Prof. Dr. Wolfgang Härdle**
Zweitgutachter:	**PD Dr. Marlene Müller**

Berlin, 23. Juli 2007

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Lars Rohrschneider

Berlin, 23. Juli 2007

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Bei der Analyse von empirischen Daten steht man häufig vor dem Problem von fehlenden Daten. Die meisten Methoden zur Datenanalyse wurden für Datensätze ohne fehlende Werte entwickelt. Man unterscheidet zwischen unerwarteten fehlenden Werten und den sogenannten erwarteten oder intentionalen fehlenden Werten. Dies kann z.B eine Frage nach der Anzahl der Ehejahre einer unverheirateten Person sein. In den meisten Lehrbüchern wird der Missing-Data-Thematik keine, oder nur eine Beachtung am Rande geschenkt. Die zu Lehrveranstaltungen oder in Publikationen bereitgestellten Daten sind oft auf die vollständigen Beobachtungen reduziert. Traditionelle Ansätze wie Mittelwertersetzung und Listenweiser Fallausschluss können zu erheblichen Verzerrungen bei der Schätzung von wichtigen Parametern führen. In dieser Arbeit soll ein Überblick über die Entstehungsmöglichkeiten und die gängigen Muster von fehlenden Daten gegeben werden. Neben der methodischen Beschreibung sollen in einer Monte-Carlo Simulationsstudie die Stärken und Schwächen der Verfahren getestet werden. Es wird ebenfalls ein Augenmerk auf kategorielle Daten gelegt werden. Zur Thematik gibt es einige grundlegende Arbeiten. In den frühen 1920er Jahren gab es von Wilk und Fisher erste Ratschläge, fehlende Daten durch Mittelwerte zu ersetzen. Die historische Literatur zur Missing Value Thematik, ist jedoch erst durch die grundlegende Arbeit von Rubin (1976a) gewachsen. Der heute populäre EM-Algorithmus wurde erstmal von Dempster, Laird und Rubin (1977) vorgestellt. Eines der umfangreichsten Bücher zur Thematik wurde von Little und Rubin (1987) veröffentlicht. Das Konzept der multiplen Imputation und wichtige Kenngrößen wurde ebenfalls von Rubin (1987) entwickelt. Der baysianisch ausgerichtete Joseph Schafer hat mit der Implementierung der Algorithmen in die freie Software R, ebenfalls einen grossen Beitrag geleistet. In den neunziger Jahren entwickelten vor allem viele Biostatistiker und Forscher aus dem klinischen Bereich die Algorithmen weiter. Allison (2001) hat die Thematik in den Sozialwissenschaften populär gemacht. Viele Statistikprogramme wie SPSS, SAS, Stata und Splus verfügen heute standardmäßig über Methoden zur Ersetzung von fehlenden Werten. Ein kontroverser Fall, der bis zum Obersten Bundesgericht der USA ging, ergab sich nach der letzten US Volkszählung im Jahre 2000. Nachdem die US Zensusbehörde nach einigen Zweifeln an der Richtigkeit der Haushaltsgrößen fragwürdige Angaben durch Werte von ähnlichen Haushalten nach einem sogenannten Nächsten-Nachbar-Verfahren ersetzte, verlor der Bundesstaat Utah einen Sitz im Repräsentantenhaus, klagte dagegen aber verlor letztendlich.

2 Daten

2.1 Datengrundlage

Die für die nachfolgenden Analysen und Simulation vorliegenden Daten wurden vom Institut für Statistik der Wirtschaftswissenschaftlichen Fakultät der Humboldt Universität zu Berlin erhoben. Die Daten wurden im Rahmen der Klausuren Statistik 1 und Statistik 2 im Sommersemester 2005 bzw. Wintersemester 2005/2006 erhoben. Die Studierenden werden regelmäßig vor der Klausur gebeten, einige Fragen zur Person zu beantworten.Die Vorlesung Statistik 1 wird jedes Wintersemester und die Vorlesung Statistik 2 jedes Sommersemester angeboten. Die beiden Vorlesungen sind für alle wirtschaftswissenschaftlichen Studiengänge im Grundstudium bzw. Bachelor Pflichtveranstaltungen und müssen bestanden werden. Bei Studenten der alten Diplomstudiengänge BWL und VWL gibt es die Möglichkeit, eine nicht - oder endgültig nicht bestandene Statistik 1 oder 2 Note auszugleichen. Die Bachelorstudiengänge wurden laut Prüfungsamt der Wirtschaftswissenschaftlichen Fakultät der Humboldt Universität zu Berlin zum Wintesemester 04/05 eingeführt.

2.2 Datenstruktur

Der Datensatz besteht aus 195 Beobachtungen und 9 Variablen. Der Datensatz weist numerische sowie Textvariablen auf.

- Variable Nr. 1 - Laufende Nr. (1-195)
- Variable Nr. 2 - Termin Statistik 1
- Variable Nr. 3 - Termin Statistik 2
- Variable Nr. 4 - Geschlecht (Text)
- Variable Nr. 5 - Fach (Text)
- Variable Nr. 6 - Punkte Statistik 1
- Variable Nr. 7 - Punkte Statistik 2
- Variable Nr. 8 - Note Statistik 1
- Variable Nr. 9 - Note Statistik 2

Nominalskaliert sind die Variablen zum „Termin Statistik 1“ und „Termin Statistik 2“ sowie „Geschlecht“ und „Fach“. Ordinalskaliert sind die Variablen „Note Statistik 1“ und „Note Statistik 2“. Die Variablen „Punkte Statistik 1“ und „Punkte Statistik 2“ sind metrisch skaliert. Im folgenden sollen die Variablen näher beschrieben werden.

- Variable Nr. 1 - Laufende Nr. (1-195)

Die erste Variable stellt die Identifikationsnummer der Beobachtungen dar.

Die Variablen Nr. 2-5 besitzen folgende Ausprägungen sowie Häufigkeiten.

- Variable Nr. 2 - Termin Statistik 1

	Häufigkeit	Prozent
1	112	57,4
2	83	42,6
Total	195	100

Tabelle 1: Häufigkeiten „Termin Statistik 1“

Aus Tabelle 1 ist zu erkennen, dass 57.4 Prozent der Studierenden die Klausur Statistik 1 zum ersten möglichen Termin zum Ende des Sommersemesters geschrieben haben. 42.6 Prozent haben sich dafür entschieden, die Klausur zum Ende der vorlesungsfreien Zeit zu schreiben.

- Variable Nr. 3 - Termin Statistik 2

	Häufigkeit	Prozent
3	150	76,9
4	45	23,1
Total	195	100

Tabelle 2: Häufigkeiten „Termin Statistik 2“

Für die Klausur Statistik 2 haben sich nach Tabelle 2, bereits 76.9 Prozent der Studierenden für den ersten möglichen Termin zum Ende des Wintersemesters entschieden. Nur 23.1 Prozent haben den zweiten möglichen Termin gewählt.

Um etwas über die Abweichungen der gewählten Termine herauszufinden, soll dazu eine Kontingenztabelle zwischen dem ersten und dem zweiten Termin betrachtet werden.

		Termin2		
		3	4	Total
Termin 1	1	96	16	112
	2	54	29	83
Total		150	45	195

Tabelle 3: Kontingenztabelle „Termin Statistik 1“ vs. „Termin Statistik 2“

In Tabelle 3 ist zu erkennen, dass Studierende die die Klausur Statistik 1 zum ersten Termin geschrieben haben, auch in der Mehrheit die Klausur Statistik 2 zum ersten Termin wählen. Studierende, die sich bei der Statistik 1 Prüfung für den zweiten Termin entschieden haben, sind dann auch bei der Statistik 2 Prüfung zum ersten möglichen Termin erschienen.

- Variable Nr. 4 - Geschlecht (Text)

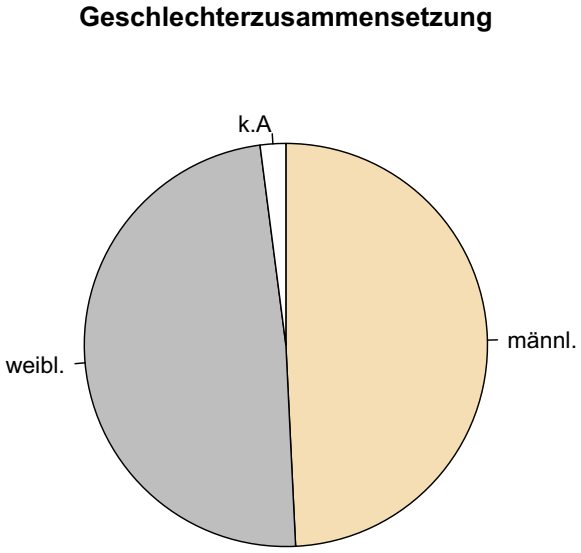

Abbildung 1: Geschlechterzusammensetzung der Studierenden

Aus Abbildung 2 und Tabelle 4 kann man entnehmen, dass das Verhältnis der Geschlechter nahezu ausgeglichen ist. Von den insgesamt 195 Studierenden haben 191 Angaben zum Geschlecht gemacht. Davon sind 95 weiblich und 96 männlich. Vier Studierende haben die Antwort verweigert. Da mir die Originaldaten mit Namen und Matrikelnummern aus Datenschutzgründen nicht vorliegen, kann an dieser Stelle keine Vollständigkeit erzeugt werden.

	Häufigkeit	Prozent	Kumuliert
k.A.	4	2.1	2.1
w	95	48.7	50.8
m	96	49.2	100
Total	195	100	100

Tabelle 4: Häufigkeiten „Geschlecht“

- Variable Nr. 5 - Fach (Text)

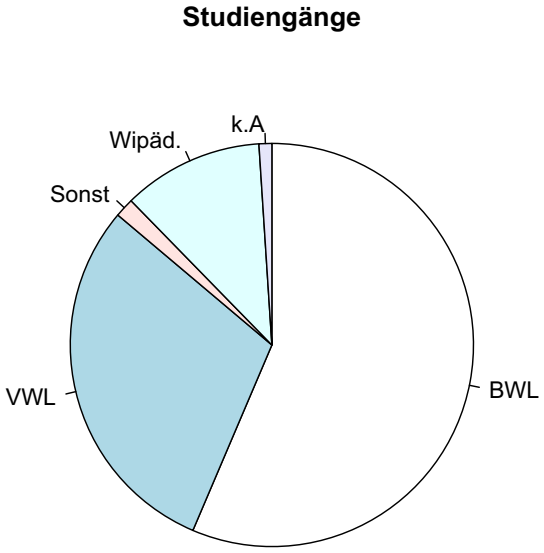

Abbildung 2: Zusammensetzung nach Studiengängen

	Häufigkeit	Prozent	Kumuliert
BWL BA.	1	0.5	0.5
k.A.	2	1	1.5
VWL BA.	2	1	2.5
Wipäd	22	11.3	13.8
VWL	58	29.7	43.5
BWL	110	56.4	100
Total	195	100	100

Tabelle 5: Häufigkeiten Studiengänge

Von den 195 Studierenden haben 193 Angaben zu ihrem Studienhauptfach gemacht. 57 Prozent studieren BWL, ca 31 Prozent studieren VWL und 11 Prozent Wirtschaftspädagogik. Ein Prozent der Studierende hat keine Angaben gemacht.

Im folgenden soll die Fächeraufteilung nach Geschlecht untersucht werden. In Tabelle 6 kann man den Anteil der Geschlechter an den Studiengängen ersehen. 61 von 95 Studentinnen studieren BWL. Auch in der Wirtschaftspädagogik sind Studentinnen stärker vertreten. Die Studenten bevorzugen zwar ebenfalls BWL, jedoch ist ihr relativer Anteil in der VWL doppelt so hoch wie bei den Studentinnen.

		Fach				
		BWL	VWL	Wipäd	Sonst.	Total
Geschlecht	M	49	39	7	1	96
	W	61	17	15	2	95
Total		110	56	22	3	191

Tabelle 6: Kontingenztabelle Geschlecht vs. Fach

- Variable Nr. 6 - Punkte Statistik 1
- Variable Nr. 7 - Punkte Statistik 2

Für die Schätzung und Visualisierung der Dichtefunktion der „Punkte Statistik 1“ und „Punkte Statistik 2“ soll die sogenannte **Kerndichteschätzung** verwendet werden. Die Kerndichteschätzung ist ein nichtparametrisches Verfahren welches es erlaubt, die Wahrscheinlichkeitsdichte zu schätzen. Die Dichtefunktion charakterisiert nahezu das komplette Verhalten einer Zufallsvariablen, wichtige Parameter wie Mittelwert und Varianz können direkt aus ihr abgeleitet werden.

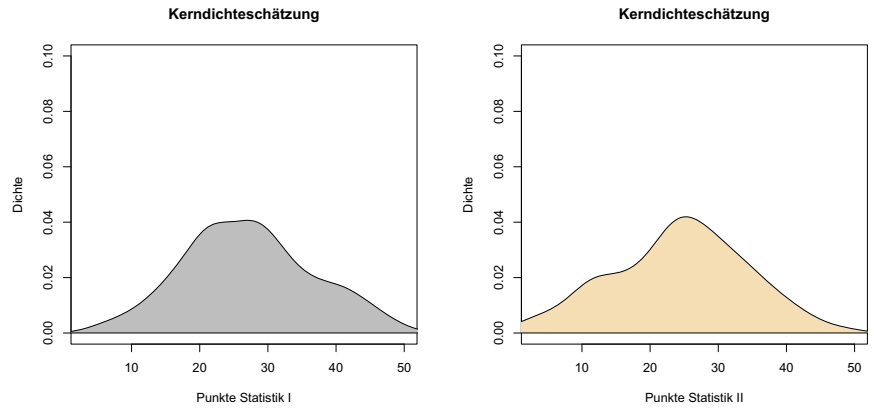

Abbildung 3: Kerndichte

Um eine stetige Schätzung der Dichtefunktion zu erhalten, benötigt man eine Bandbreite h und die sogenannte Kernfunktion die Beobachtungen um x_i

nach bestimmten Kriterien gewichtet. Wobei der Abstand zwischen X und x_i nicht größer als h sein darf. Anhand der Bandbreite und des benutzten Kernes kann man festlegen, wie stark geglättet die Dichtefunktion sein soll. Für die Berechnung der optimalen Bandbreite und der Benutzung eines optimalen Kernes wurde zahlreiche wissenschafliche Arbeiten veröffentlicht. Für weitere Informationen wird auf Härdle et al. (2004) oder Silverman (1986) verwiesen. Für die Grafik wurde die Software R mit der Gausschen Kernfunktion, sowie Silverman's Daumenregel für die optimale Bandweite h_{opt} benutzt.

- Variable Nr. 8 - Note Statistik 1
- Variable Nr. 9 - Note Statistik 2

In Abbildung 4 sind die Histogramme der Statistiknoten dargestellt. Das Histogramm ist eine einfache Möglichkeit zu erkennen, wie eine Variable verteilt ist. Man kann eine linksschiefe Verteilung der Statistiknoten erkennen. Die Durchfallrate (Note = 5.0) liegt bei 20 Prozent in Statistik 1 und bei 15 Prozent in Statistik 2.

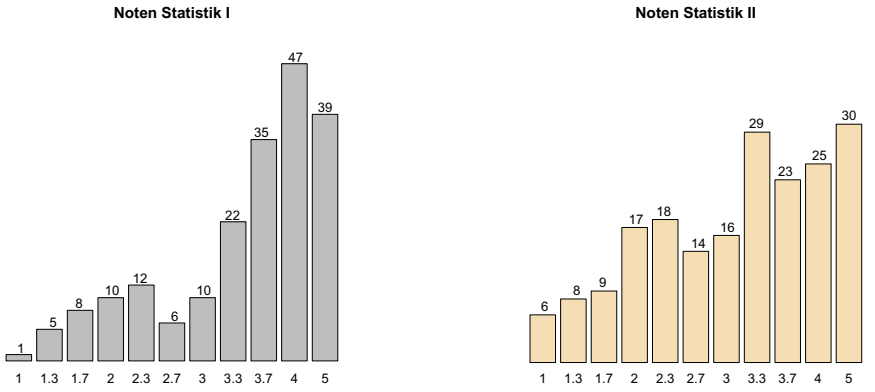

Abbildung 4: Notenverteilung

Auffallend ist die große Anzahl der Noten im Bereich von 3.3 bis 4.0. Wobei man mit einer 4.0 die Klausur gerade noch bestanden hat. Der Mittelwert liegt in Statistik 1 bei 3.6 und bei der Statistik 2 mit 3.2 etwas besser (Median 3.7 bzw 3.3).

2.3 Kurze Tests und Analysen des Datensatzes

In Vorbereitung der späteren Simulation sollen die Punkte sowie Noten auf Normalverteilung getestet werden, sowie auf die Ergebnisse der Diplomarbeit von Frau Brandes (2005) eingegangen werden. Frau Brandes hatte in ihrer Abschlussarbeit Leistungen der Studierenden mit Hilfe ähnlicher Daten aus vom Wintersemester 2000/2001 bis zum Wintersemester 2002/2003 evaluiert. Zu jener Zeit waren die Leistungen von männlichen Studierenden noch signifikant besser als die von weiblichen Studierenden. Des weiteren wird kurz auf eventuelle Leistungsunterschiede zwischen den Studiengängen eingegangen.

2.3.1 Kolmogorov-Smirnov Test

Die „Punkte Statistik 1“ und „Punkte Statistik 2“ sehen wie aus Abbildung 3 zu erkennen ist, bereits recht normalverteilt aus. Da die Normalverteilung für spätere Simulationen hilfreich sein könnte, soll nun getestet werden, ob die Daten mit einer gewissen Sicherheit einer Normalverteilung folgen. Dies soll mit Hilfe des KS-Tests geschehen. Der Kolmogorov-Smirnov Test ist ein nichtparametrischer Anpassungstest zur Überprüfung der Hypothese, ob die Verteilung $F_n(x)$ der Stichprobe $(X_1, ..., X_n)$ aus einer Grundgesamtheit mit der Verteilung $F_0(x)$ stammt vgl. Rönz (2001). Wenn die vorgegebene Verteilung nun auch die wahre Verteilung in der Grundgesamtheit ist, so sollte sie auch in unseren Daten zu beobachten sein. Die Nullhypothese und die Alternativhypothese lautet wie folgt.

$$H_0 : F_n(x) = F_0(x) \text{ vs. } H_1 : F_n(x) \neq F_0(x)$$

Man überprüft insbesondere, ob die unbekannte Verteilungsfunktion $F_n(x)$ einer Normalverteilung mit den Parametern $\phi = \left(\frac{x-\mu}{\sigma}\right)$ folgt.

Als Teststatistik wird die Formel 2.1

$$D_n = \|F_n - F_0\| = \sup |F_n(x) - \phi(\frac{x - \bar{x}}{s})| \tag{2.1}$$

verwendet, wobei μ durch $\bar{x}$ und σ durch s ersetzt wird. Es werden so für jeden Wert die relativen Summenhäufigkeiten verglichen und die maximale Differenz wird als Prüfgröße verwendet. Diese Teststatistik ist unabhängig von der Verteilung. Wenn diese Teststatistik kleiner ist, als der tabellierte kritische Wert $d_{n,1-\alpha}$, so kann die Nullhypothese auf einem vorgegebenem Signifikanzniveau α nicht verworfen werden.

2.3.2 T-Test

Bei diesem Test handelt es sich um einen Hypothesentest mit t-verteilter Prüfgröße. Wir testen, ob es signifikante Unterschiede zwischen den Mittelwerten aus zwei

verschiedenen Gruppen gibt. Es wird die Nullhypothese $H_0 : \mu_1 = \mu_2$ überprüft.

Die Voraussetzungen für die Durchführung sind nach Rönz (2001) wie folgt:

1. Gegeben seien zwei Grundgesamtheiten mit $E(X_1) = \mu_1$ mit $Var(X_1) = \sigma_1^2$ und $E(X_2) = \mu_2$ mit $Var(X_2) = \sigma_2^2$. μ_1 und μ_2 sind unbekannt.
2. Es wird unterstellt, dass der Umfang der beiden Grundgesamtheiten N_1 und N_2 hinreichend groß ist, so dass man von der Realisierung einfacher Zufallsstichproben ausgehen kann. Der Stichprobenumfang ist dann n_1 und n_2.
3. Die Zufallsvariablen X_1 und X_2 in den Grundgesamtheiten sind normalverteilt oder n_1 und n_2 sind groß genug, so dass der zentrale Grenzwertsatz gilt.
4. Die Zufallsstichproben $X_{1,1}, ..., X_{1,n1}$ und $X_{2,1}, ..., X_{2,n2}$ sind unabhängig.
5. Die beiden Grundgesamtheiten haben eine gleiche, aber unbekannte Varianz $\sigma_1^2 = \sigma_2^2$

Es wird nun mit einem zweiseitigen t-Test geprüft, ob die Mittelwerte in beiden Stichproben gleich sind. Die Nullhypothese und die Alternativhypothese lauten wie folgt.

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

Für diesen zweiseitigen t-Test gilt folgende Teststatistik.

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2} \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-1}}} \tag{2.2}$$

Diese Teststatistik folgt unter Gültigkeit der Nullhypothese einer t-Verteilung mit $f = n_1 + n_2 + 2$ Freiheitsgraden. Die kritischen Werte $t_{(f,1-\frac{\alpha}{2})}$ und $t_{(f,\frac{\alpha}{2})}$ ergeben sich als $1 - \frac{\alpha}{2}$ und $\frac{\alpha}{2}$ Quantile der t-Verteilung. Ist der Wert der Teststatistik größer als der kritische Wert, so wird die Nullhypothese verworfen.

2.3.3 Welch Test

Oft ist es der Fall, dass die Annahme der Gleichheit der Gruppenvarianzen $\sigma_1^2 \neq \sigma_2^2$ beim t-test nicht haltbar ist. In diesem Fall folgt die Teststatistik

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.3}$$

einer t-Verteilung mit

$$f = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}} \tag{2.4}$$

Freiheitsgraden. Vgl. auch Welch (1947) und Rönz (2001).

2.3.4 F-Test

Einige Verfahren, wie der t-Test oder die lineare Diskriminanzanalyse, verlangen explizit nach Gleichheit der Gruppenvarianzen $\sigma_1^2 = \sigma_2^2$. Mit dem F-Test nach Ronald Aylmer Fisher (1890-1962) lässt sich folgenden Hypothese überprüfen.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2$$

Die Voraussetzungen für die Durchführung sind nach Rönz (2001) wie folgt:

1. Gegeben seien zwei normalverteilte Grundgesamtheiten $E(X_1) = \mu_1$ mit $Var(X_1) = \sigma_1^2$ und $E(X_2) = \mu_2$ mit $Var(X_2) = \sigma_2^2$. μ_1 und μ_2 sind unbekannt.

2. Es wird unterstellt, dass der Umfang der beiden Grundgesamtheiten N_1 und N_2 hinreichend groß ist, so dass man von der Realisierung einfacher Zufallsstichproben ausgehen kann. Die Stichprobenumfänge sind dann n_1 und n_2.

3. Die Zufallsvariablen X_1 und X_2 in den Grundgesamtheiten sind normalverteilt oder n_1 und n_2 sind gross genug, so dass der zentrale Grenzwertsatz gilt.

4. Die Zufallsstichproben $X_{1,1}, ..., X_{1,n1}$ und $X_{2,1}, ..., X_{2,n2}$ sind unabhängig und haben die gleiche Verteilung wie die Grundgesamtheit.

Es gilt folgende Teststatistik.

$$V = \frac{S_1^2}{S_1^2} \tag{2.5}$$

Diese folgt unter Gültigkeit der Nullhypothese einer F-Verteilung. Dabei sind $F_{(f_1,f_2,\frac{\alpha}{2})}$ bzw. $F_{(f_1,f_2,1-\frac{\alpha}{2})}$ die $\frac{\alpha}{2}$ bzw. $1 - \frac{\alpha}{2}$ Quantile der F-Verteilung mit $f_1 = n_1 - 1$ und $f_2 = n_1 - 1$ Freiheitsgraden. Die Nullhypothese wird auf dem α Signifikanzniveau verworfen wenn $V \leq \mathrm{F}_{(f_1,f_2,\frac{\alpha}{2})}$ bzw. $V \geq \mathrm{F}_{(f_1,f_2,1-\frac{\alpha}{2})}$.

2.4 Ergebnisse

Alle folgenden Ergebnisse beziehen sich auf Analysen der Klausurpunkte aus Statistik 1 und Statistik 2. Es werden die Verfahren verwendet die im vorigen Abschnitt erläutert wurden.

2.4.1 Test auf Normalverteilung

Die Nullhypothese des Kolmogorov-Tests auf Normalverteilung kann auf dem 5 Prozent Niveau nicht abgelehnt werden. Die Verteilungen der „Punkte Statistik 1“ und „Punkte Statistik 1“ sind somit mit hoher Sicherheit normalverteilt (p-Werte: „Punkte Statistik 1“ 0.41 und „Punkte Statistik 2“ 0.27).

2.4.2 Varianzanalyse

Die Nullhypothese des F-Tests, Gleichheit der Varianzen zwischen den Geschlechtern, kann auf dem 5 Prozent Signifikanzniveau nicht abgelehnt werden. (p-Wert: 0.16) Die p-Werte des Tests zwischen den Studiengängen sind wie in Tabelle 7 und 8 zu erkennen angegeben.

	BWL	VWL	Wipäd.
BWL	1		
VWL	0.97	1	
Wipäd	0.98	0.96	1

Tabelle 7: p-Werte F-Test „Punkte Statistik 1“

	BWL	VWL	Wipäd.
BWL	1		
VWL	0.21	1	
Wipäd	0.80	0.62	1

Tabelle 8: p-Werte F-Test „Punkte Statistik 2“

Die Varianzen zwischen den Studiengängen können auf dem 5 Prozent Signifikanzniveau ebenfalls als gleich angesehen werden.

2.4.3 Welch Test

Eine interessante Frage ist nun, ob sich wie in der Arbeit von Frau Brandes (2005) noch signifikante Leistungsunterschied zwischen männlichen und weiblichen Studierenden feststellen lassen. Da die Varianzen zwischen Geschlechtern und zwischen den Studiengängen gleich sind, kann entweder der t-Test oder der Welch-Test angewandt werden. Für die erreichten Punkte in Statistik 1 und 2,

lässt sich in Gegensatz zur Arbeit von 2005 kein signifikanter Unterschied zwischen den Geschlechtern mehr feststellen. (p-Werte: „Punkte Statistik 1“ 0.40 und „Punkte Statistik 2“ 0.28) Die weiblichen Studierende sind sogar leicht besser in beiden Veranstaltungen, was allerdings noch zufallsbedingt ist.

	M	W	BWL	VWL	Wipäd.
Statistik I.	26.63	27.76	27.24	27.64	25.32
Statistik II.	23.72	25.23	23.35	27.26	22.86

Tabelle 9: Punkteverteilung nach Geschlecht und Studiengang

Zwischen den Studiengängen gibt es in Statistik 1, wie man aus Tabelle 10 ersehen kann, keine signifikanten Unterschiede. Die Ergebnisse des Mittelwertvergleichs für Statistik 2 sind in Tabelle 11 zu finden. Unter anderem erreichen Studierende der VWL eine signifikant höhere Punktzahl als Studierende der BWL (p-Wert: 0.01).

	M	W	BWL	VWL	Wipäd.
M	1				
W	0.40	1			
BWL	0.65	0.68	1		
VWL	0.52	0.94	0.79	1	
Wipäd	0.56	0.27	0.39	0.33	1

Tabelle 10: p-Werte Mittelwertvergleich Statistik I

Die Vergleiche zwischen BWL und Wirtschaftspädagogik sowie VWL und Wirtschaftspädagogik sind nicht signifikant (p-Wert > 0.05).

	M	W	BWL	VWL	Wipäd.
M	1				
W	0.29	1			
BWL	0.80	0.16	1		
VWL	0.03	0.18	0.01	1	
Wipäd	0.71	0.30	0.83	0.07	1

Tabelle 11: p-Werte Mittelwertvergleich Statistik II

3 Muster und Mechanismen fehlender Daten

3.1 Muster fehlender Daten

In einer Datenmatrix Y repräsentieren die n Zeilen die Beobachtungen oder Fälle und die p Spalten die Variablen. In der Literatur wird zwischen dem sogenannten Datenmuster der fehlenden Werte und dem Mechanismus der fehlenden Daten unterschieden. Das Datenmuster beschreibt welcher Wert in den Daten beobachtet wird und welcher Wert fehlt. Der Mechanismus beschreibt einen möglichen Zusammenhang, der hinter den fehlenden Werten steht. Auf die Mechanismen soll im nächsten Abschnitt eingegangen werden.

$$Y = \begin{pmatrix} y_{11} & \cdots & \cdots & y_{1p} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ y_{n1} & \cdots & \cdots & y_{np} \end{pmatrix}$$

Einige Verfahren benötigen eine spezielle Datenstruktur um die fehlenden Werte Handhaben zu können, andere Verfahren stellen keinerlei Ansprüche. Es ist oft sinnvoll Zeilen und Spalten der Matrix mit den fehlenden Werten so umzustellen, dass sich bestimmte Muster ergeben. Little und Rubin (2002) schlagen vor durch Bildung einer Indikatormatrix $M = (m_{ij})$ der Datenmatrix $Y = (y_{ij})$ dieses Muster abzubilden. Die Indikatormatrix bekommt dann den Wert $m_{ij} = 0$ wenn y_{ij} keinen fehlenden Wert aufweist und $m_{ij} = 1$ falls y_{ij} fehlt. Diese Matrix bestimmt dann die Muster fehlender Daten. In einer Datenmatrix mit p Variablen kann es bis zu 2^p verschiedene Muster geben. So kann es bei fünf Variablen bereits bis zu 32 verschiedene Muster geben. Ein wichtiger Aspekt kommt monotonen und zufälligen Mustern, sowie nichtmonotonen Mustern zu. Eine besondere Bedeutung haben monotone Datenmustern vor allem, da diese durch Imputationsverfahren oft einfacher zu handhaben sind. Die wichtigen Muster sollen nun vorgestellt werden.

Univariate Missing Data: Bei diesem Muster konzentrieren sich die fehlenden Werte in einer einzelnen Variablen. Dieses Muster ist aus rechentechnischer Sicht einfach zu handhaben. In klinischen Studien werden bestimmte Variablen deren Erfassung mit hohen Kosten verbunden ist, oft nur bei einem repräsentativem Teil der Patienten erhoben.

$$M = \begin{pmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & & \vdots & 0 \\ \vdots & & \ddots & \vdots & 1 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Item Nonresponse: Item Nonresponse tritt auf, wenn ausgewählte Teilnehmer eine Frage oder einen Teil der Fragen nicht beantworten. Item Nonresponse ist umso häufiger, sobald die Intention der Frage nicht verstanden wird, eine unklare Information die Beantwortung nicht ermöglicht, oder auch der Wille fehlt trotz zugesicherter Anonymität bestimmte Daten preis zu geben.

$$M = \begin{pmatrix} 0 & \cdots & \cdots & 0 & 1 & 0 \\ \vdots & \ddots & & \vdots & \vdots & 1 \\ \vdots & & \ddots & \vdots & 1 & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{pmatrix}$$

Unit Nonresponse: Unit Nonresponse bedeutet, dass sich ein Teilnehmer einer Befragung komplett verweigert und keinerlei Auskunft gibt. Je nach Zielstellung kann man neue Probanden gewinnen, oder mit den vorhandenen Teilnehmern fortfahren. Es gibt in Deutschland seit den neunziger Jahren anstatt regelmäßiger Volkszählungen den sogenannten Mikrozensus. Hier werden vom Statistischen Bundesamt bestimmte Personen nach repräsentativen Kriterien ausgewählt und diese können sich dann in der Regel nicht der kompletten Teilnahme verweigern.

$$M = \begin{pmatrix} 1 & \cdots & \cdots & 1 \\ 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

Im Bereich der Meinungsforschung wird oft auf die Prozentzahl der Befragten hingewiesen die keine Meinung zum Thema hatten, oder eine Antwort verweigert haben. In den meisten Studien werden Totalausfälle aber ignoriert, da angenommen wird, dass sie keinen Effekt auf die Qualität der Umfrageergebnisse haben. Im folgenden Beispiel verweigert der erste Proband die Teilnahme komplett.

Zufälliges Datenmuster: Bei einem zufälligen Datenmuster lässt sich ohne weiteres keine spezielle Struktur erkennen. Es wäre jedoch verfrüht anzunehmen, dass es keine statistischen oder kausalen Beziehungen zwischen den vollständigen und fehlenden Daten gibt.

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Monotones Datenmuster: Mathematisch gilt ein Muster als monoton, falls die Variablen so geordnet werden können, dass für alle Beobachtungen in denen

eine Variable Y_j einen fehlenden Wert aufweist auch alle folgenden $Y_{j+1}.....Y_p$ $J = 1, ...P - 1$ Variablen fehlende Werte aufweisen. In der Praxis ist dieses Muster eher bei Panelstudien anzutreffen. An einem bestimmten Punkt der Studie fällt der Proband aus und kehrt nicht mehr zurück. Ab diesem Punkt fehlen alle Antworten zu den späteren Fragen.

$$M = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & & \iddots & 1 \\ \vdots & & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & & \vdots \\ 0 & 1 & \cdots & \cdots & 1 \end{pmatrix}$$

Laut Little Rubin (2002) ist ein Muster in der Praxis selten monoton aber oft nahezu monoton. Wenn ein Muster nahezu monoton ist, besteht auch die Möglichkeit erst wenige Daten zu imputieren, um ein monotones Muster zu erhalten. Viele Methoden zur Behandlung von fehlenden Daten kommen mit einem monotonen Muster besser zurecht als mit einem zufälligem Muster.

3.2 Mechanismen fehlender Daten

Das Fehlen von Daten wird laut Rubin (1976) als probalistisches Phänomen eingestuft. Der beobachtete Teil einer Datenmatix Y lässt sich als Y_{obs} darstellen und der nichtbeobachtete als Y_{mis}. $Y = (Y_{mis}, Y_{obs})$. Eine akkurate Beschreibung des kompletten Mechanismus der fehlenden Daten ist aber nahezu unmöglich. Rubin (1976a) hat die folgende Mechanismen benannt: MCAR (Missing-Completely-at-Random), MAR (Missing-at-Random) und MNAR (Missing-not-at-Random). Diese Mechanismen beziehen sich nicht auf kausale Beziehungen, sondern auf statistische Beziehungen zwischen den vollständigen und fehlenden Daten. Man definiert für die folgenden Abschnitte wieder M als Indikator Matrix für fehlende Daten. M wird im Zusammenhang mit den Mechanismen als Zusammenschluss von Zufallsvariablen betrachtet die durch eine gemeinsame Verteilung charakterisiert werden können. Diese Verteilung erfasst mögliche Beziehungen zwischen dem Auftreten von fehlenden Werten und Ausprägungen der Zufallsvariablen. Auch wenn man diese Verteilung nicht explizit spezifizieren muss, so muss man zustimmen, dass diese existiert, vgl. Schafer und Graham (2002). Es ist jedoch nahezu unmöglich alle Ursachen und Gründe für fehlende Daten in einem statistischen Modell zu erfassen. Ferner geht man davon aus, dass der Mechanismus der fehlenden Daten durch die bedingte Verteilung von M gegeben Y beschrieben wird $f(M \mid Y, \phi)$. Unbekannte Parameter der Verteilung werden durch ϕ charakterisiert.Vgl. Little und Rubin (2002).

3.2.1 MCAR

Missing-Completely-at-Random bedeutet, dass der Mechanismus der hinter den fehlenden Werten steht weder von den fehlenden Werten selbst noch von den kompletten Daten abhängt. Die mathematische Formulierung lautet wie folgt.

$$f(M \mid Y, \phi) = f(M \mid \phi) \quad \forall \quad Y_{mis}, \phi \tag{3.1}$$

Die Wahrscheinlichkeit ein Missing zu beobachten oder nicht ist in keiner Weise von den vorhandenen Daten Y abhängig. Das Auftreten von fehlenden Werten wird zum Beispiel verursacht durch Krankheit der Person zur Klausur, Fehler im Institut bei der Dateneingabe, Datenverlust des Rechners etc. In der Praxis tritt Missing-Completely-at-Random aber selten auf. Allison (2001) erwähnt, dass die MCAR Annahme verletzt sein würde, wenn zum Beispiel jüngere gegenüber älteren Personen die Angaben zum Einkommen verweigern würden. Er schlägt vor, hier einen einfachen t-Test, oder auch das nichparametrische Gegenstück den Kruskal-Wallis-Test zuverwenden. Dies geschieht indem man die Personen in Gruppen unterteilt die Ihr Einkommen berichtet haben und den Leuten die die Aussage verweigert haben. Man überprüft nun, ob sich diese Leute in der Altersstruktur entscheiden. Ein Verfahren für größere Datensätze wurde auch von Little (1988) entwickelt. Dieses Verfahren ist ähnlich dem Chi-Quadrat Unabhängigkeitstest. Die Methode besteht aus drei Schritten und soll an einem Beispiel erklärt werden. Für den Fall von drei Variablen mit drei Ausprägungen erhält man eine Kontingenztabelle mit 27 Zellen. Der Ablauf gliedert sich wie folgt:

- Erstelle eine Kontingenztabelle mit Zellwahrscheinlichkeiten die nur aus vollständigen Beobachtungen geschätzt werden.
- Erstelle eine Kontingenztabelle mit Zellwahrscheinlichkeiten die aus allen Beobachtungen geschätzt werden. Maximum-Likelihood Schätzungen dieser Zellwahrscheinlichkeiten erhält man mit Hilfe des EM Algorithmus auf den noch eingegangen wird.
- Multipliziere die Zellwahrscheinlichkeiten mit der gesamten Stichprobengröße. Führe einen Chi-Quadrat Test oder Likelihood-Ratio Test zum Vergleich der beiden Kontingenztabellen durch. Die Teststatistik wird durch Summation über alle Zellwahrscheinlichkeiten der beiden Tabellen berechnet. Für p-Werte größer als 0.05 kann die Nullhypothese nicht verworfen werden. Die fehlenden Werte sind dann mit gewisser Sicherheit MCAR.

Der MCAR Test ist zum Beispiel im SPSS MVA-Modul verfügbar. Falls die Annahme von MCAR haltbar ist, so können die kompletten Beobachtungen als einfache Stichprobe der Population betrachtet werden. Allison (2001) gibt als Beispiel das bewusste Fehlen bei Studien an. In klinischen Studien werden bestimmte Variablen deren Erfassung mit hohen Kosten verbunden ist, oft nur bei einem kleinen Teil der Patienten erfasst. MCAR stellt nur einen Spezialfall der MAR Annahme dar. In Abbildung 5 erkennt man fehlende Daten in den Klausurpunkten von Statistik 2 die zufällig erzeugt worden sind.

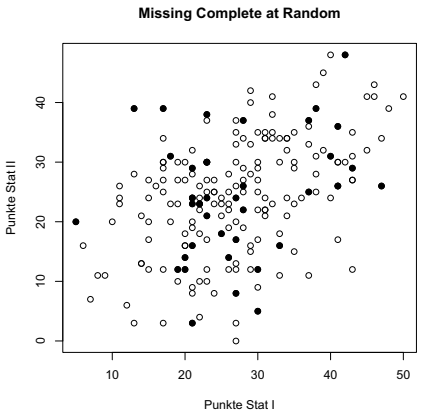

Abbildung 5: MCAR

3.2.2 MAR

Die MAR Annahme ist schwächer als die MCAR Annahme. Die Verteilung der fehlenden Daten hängt nicht von den fehlenden Daten Y_{mis} selbst ab, kann jedoch von anderen beobachteten Werten Y_{obs} abhängen.

$$f(M \mid Y, \phi) = f(M \mid Y_{obs}, \phi) \quad \forall \quad Y_{mis}, \phi \tag{3.2}$$

Betrachtet man den Fall von fehlenden Daten zur Einkommenshöhe und geht man davon aus, dass ältere Personen eher dazu neigen Angaben zum Einkommen zu verweigern als jüngere, so ist der Mechanisus hinter den fehlenden Einkommensdaten Missing-at-Random. Mittels einer Regression wäre es zum Beispiel nun möglich, Einkommensdaten mehr oder weniger zuverlässig vorauszusagen. Des Weiteren darf auch keinen Zusammenhang zwischen den fehlenden Werten der Variable und den Ausprägungen der Variablen selbst vorhanden sein. Dies wäre der Fall, wenn fehlende Werte von der Einkommenshöhe selber abhängen würden. Collins, Schafer und Kam (2001) haben gezeigt, dass in vielen realistischen Fällen eine fehlerhafte Annahme von MAR oft nur kleine Auswirkungen auf Schätzungen und Standardfehler haben kann. Es ist nicht testbar, ob fehlende Daten Missing-at-Random sind. Laut Allison (2001) sollten fehlende Werte im Falle von MAR streuen und sich nicht räumlich konzentrieren. Sollten man bei der graphischen Darstellung der Indikatormatrix also in einzelnen Variablenblöcken eine räumliche Konzentrierung feststellen, so kann man davon ausgehen, dass die Daten nicht mehr Missing-at-Random sind.

Abbildung 6: MAR

3.2.3 MNAR

Das Vorhandensein eines MNAR Mechanismus stellt den schwierigsten Fall dar.

$$f(M \mid Y, \phi) = f(M \mid Y_{obs}, Y_{mis}, \phi) \quad \forall \quad Y_{mis}, \phi \tag{3.3}$$

Die Verteilung von M selber, hängt auch von fehlenden Werten im Datensatz ab. Als plausibles Beispiel kann das Verweigern von Antworten zum Einkommen, von Leuten die ein hohes oder niedriges Einkommen haben betrachtet werden. Der Mittelwert erhöht bzw. verringert sich, die Standardabweichung sinkt. Um zuverlässige Schätzungen durchführen zu können, muss man den Prozess kennen der hinter den fehlenden Daten steht und ein wenigstens aproximativ korrektes Modell spezifizieren, vgl. Schafer und Graham (2002)

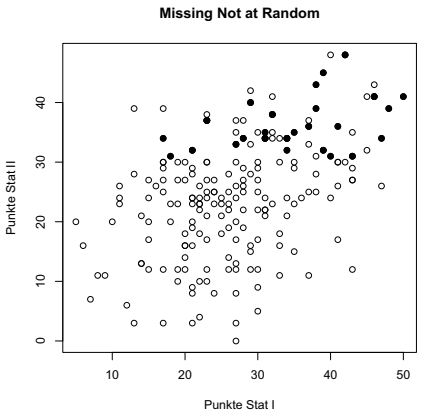

Abbildung 7: MNAR

3.3 Ignorierbarkeit des Missing Data Mechanismus

Laut Rubin (1976) kann der Missing Data Mechanismus als vernachlässigbar betrachtet werden, wenn die Daten MAR oder MCAR sind und die Parameter die den Misssing Data Prozess steuern unabhängig von den zu schätzenden Parametern sind. Allison (2001) gibt an, dass dann keine Notwendigkeit besteht den Prozess der hinter den fehlenden Daten steht zu modellieren. Seiner Meinung nach ist in der Praxis die MAR Annahme ausreichend, jedoch kann mit einer expliziten Modellierung des Mechanismus bessere Ergebnisse erzielt werden.

3.4 Nichtignorierbarkeit des Missing Data Prozesses

Wenn die Daten nicht MAR sind dann kann man den Missing Data Mechanismus nicht mehr vernachlässigen. Um gute Schätzungen der Parameter zu erhalten, muss der Mechanismus direkt modelliert werden. Allison (2001) gibt folgende plausible Beispiele. Personen mit hohem Einkommen neigen überdurchschnittlich oft dazu, Angaben zum Einkommen zu verweigern. Personen mit starkem Übergewicht machen weniger häufig Angaben zum Gewicht, Kriminelle machen mit geringer Wahrscheinlichkeit Angaben über ihre Vorstrafen, Studenten mit schlechten Noten geben diese ungern preis und Personen, die in klinischen Studien aufgrund einer verschlechterten Wirkung aussteigen sind häufiger als Personen, die aufgrund von positiver Wirkung die Studie verlassen. Aus den Daten kann man diesen unbekannten Mechanismus aber nicht direkt erkennen. Ein bekanntes Modell, dass sich dem Problem der Verzerrungen bei der Stichprobenauswahl annimmt, ist das Modell nach Heckman (1979). Dieses ökonometrische Modell hält vor allem bei mikroökonomischen Untersuchungen zur Arbeitsmarktthematik Einzug. Oft wird versucht, die sogenannte Lohngleichung, in Abhängigkeit von

bestimmten Variablen wie Dauer der Schulbildung, Alter, Geschlecht und Migrationshintergrund zu schätzen. Bestimmte Personengruppen mit einem geringen Ausbildungsgrad oder hohen Alters sind aber oft unterrepräsentiert. Das Heckman Modell liefert FIML (Full Information Maximum Likelihood) Schätzungen durch eine zweistufigen LIML Schätzung (Limited Maximum Likelihood). Laut Heckman (1979) erfolgt die Schätzung bei normalverteilten Daten wie folgt.

1. Schätze ein Probit Regressionsmodell für Y auf X, um das sogenannte inverse Mills Verhältnis zu erhalten.

2. Schätze nun die die Regressionskoeffizienten von Y auf X mittels der kleinste Quadrate Methoden und addiere das inverse Mills Verhältnis. Heckman (1979) beschreibt diesen Schätzer als konsistent, aber nicht effizient bei heteroskedastischen Fehlertermen. White (1982) beschreibt ein Verfahren das dieses Problem überwinden kann.

Durch die starken Annahmen der Normalverteilung und Homoskedastizität der Fehlerterme wird dieses Modell aber in der Literatur oft kritisiert. Andere Herangehensweisen für das MNAR Problem beschreibt Rubin (1987) mit den sogenannten Pattern-Mixture-Modellen. Allison (2001) schlägt vor, diese Modelle mit Multiple-Imputation zu kombinieren. Zuerst sollen eine Reihe von Imputationen unter einem ignorierbarem Modell durchgeführt werden und diese imputieren Werte werden dann zum Beispiel durch eine lineare Transformation verändert. Diese Transformationen können unter einer Reihe 'plausibler' Annahmen durchgeführt werden. Zum Beispiel können die imputierten Werte mit einer Konstante größer oder kleiner ein multipliziert werden.

4 Listenweiser und Paarweiser-Fallausschluss

In diesem Abschnitt sollen die beliebten Methoden des Fallweisen und Paarweisen Ausschlusses erläutert werden.

4.1 Fallweiser - (Listenweiser) Ausschluss

Das einfachste Verfahren bei Datensätzen mit fehlenden Daten ist der sogenannte Fallausschluss. In nahezu jeder statistischen Software ist dieser als Standardverfahren vorhanden. Bei dieser Methode werden nur die vollständigen Beobachtungen in die weitergehenden Analysen einbezogen. Jede Beobachtung die auch nur einen fehlenden Wert beinhaltet, wird von der Analyse ausgeschlossen. Man sollte jedoch die Wichtigkeit der einzelnen Beobachtungen für die weiteren Analysen prüfen. Problematisch ist dieses Verfahren insbesondere bei geringer Anzahl von Beobachtungen mit kompletten Daten. In der Literatur wird vorgeschlagen diese Herangehensweise nur bei Daten mit MCAR zu verwenden, vgl. Rubin (1987) und Allison (2001). Wenn die fehlenden Daten Missing-Completely-at-Random sind, so stellt die Auswahl nach Löschung der fehlenden Daten eine repräsentative Stichprobe dar. Bei einer repräsentativen Stichprobe gäbe es keine Probleme für weitere Analysen wie Strukturgleichungsmodelle, Regressionen etc. Die Parameter wie Mittelwerte für den reduzierten Datensatz wären dann unverzerrt. Bei Datensätzen aus der Praxis ist die MCAR Annahme aber selten erfüllt. Wenn die MAR Annahme gilt und man Beobachtungen löscht, die zum Beispiel einen fehlenden Wert bei der Einkommensangabe haben (Man weiss, dass dort Personen mit hohem Einkommen eher dazu neigen die Antwort zu verweigern.), so ist die verbleibende Stichprobe nicht mehr repräsentativ und die Parameter der Verteilung sind verzerrt. Eine qualifizierte Aussage aus den verbleibenden Daten ist nicht mehr möglich. Laut Allison (2001) ist aber insbesondere die Verwendung des Fallausschlusses bei Schätzungen von Regressionsanalysen selbst bei der Verletzung der MAR Annahme nicht so problematisch wie die Verwendung anderer Verfahren. In der Praxis ist Datenerfassung auch meist mit Kosten verbunden und ganze Beobachtungen zu löschen ist auch aus wirtschaftlichen Gründen nicht immer sinnvoll. Als Extrembeispiel für den Verlust an Beobachtungen führen Graham und Schafer (2002) bei nur 3 Prozent gleichverteilt fehlenden Daten in einem Datensatz mit 25 Variablen einen Verlust von 50 Prozent der Beobachtungen an. *(Dies entspricht* $1 - 0.97^{25}$.) Eine Variation der Methode besteht darin, auch ganze Variable die einen besonders hohen Anteil von fehlenden Werten haben, aus der Analyse herauszunehmen. Natürlich kann es schwerfallen, wichtige Attribute mit besonders hohem Anteil fehlender Werte aus der Analyse auszuschließen.

4.2 Effizienz bei Fallweisen Ausschluss

Für den Fall von bivariat normalverteilten Daten die ein monotones Muster aufweisen, kann die Verzerrung durch fehlenden Daten folgendermaßen dargestellt werden. Vgl. Little und Rubin (2002). Hierbei gibt es keine fehlenden Werten in Y_1. In Y_2 sind r der n Beobachtungen komplett und es fehlen n-r Beobachtungen. Für den Fall von MCAR spielt die Verzerrung der Daten eine untergeordnete Rolle. Durch Ausschluss der fehlenden Beobachtungen verringert sich ebenfalls die Effizienz bei der Schätzung des Mittelwertes in Y_1

$$\Delta_{cc}^{*} = \Delta_{cc} = \frac{n-r}{n} \tag{4.1}$$

Falls die Hälfte der Beobachtungen fehlt verdoppelt sich die Varianz. Für den Mittelwert von Y_2 hängt der Verlust der Effizienz jedoch nicht nur Anteil der fehlenden Daten, sondern auch von der Korrelation der Variablen ab.

$$\Delta_{cc} = \frac{(n-r)\rho^2}{n(1-\rho^2)+r\rho^2} \tag{4.2}$$

Δ_{cc} ist voll effizient für den Fall, dass fehlende Beobachtungen in Y_1 keine Informationen für die Regression von Y_2 auf Y_1 enthalten. Dies ist nur der Fall bei unkorrelierten Variablen.

4.3 Paarweiser Fallausschluss

Viele statistische Verfahren, wie zum Beispiel die Faktorenanalyse basieren auf der Korrelationsmatrix. Regressionsanalysen, Strukturgleichungsmodelle können auch mit Hilfe der Korrelationsmatrix geschätzt werden. Die Kovarianz zwischen drei Variablen X, Y, Z in denen jeweils fehlende Werte auftreten, kann aus den jeweils paarweise kompletten Beobachtungen geschätzt werden. Wenn man jetzt alle verfügbaren Informationen für die Berechnung der Standardabweichung benutzt, kann es unter Umständen zu Korrelationskoeffizienten grösser als 1 führen sowie nicht positiv definiten Korrelationsmatrizen (Anmerkung: Korrelationsmatrizen müssen positiv definit sein). Es ist jedoch nicht klar, ob die Stichprobengröße aus Variablen, in denen die Anzahl der vollständigen Beobachtungen am grössten oder am kleinsten ist, entnommen werden sollte. Unter MCAR bleiben aber deutlich weniger Informationen ungenutzt, als zum Beispiel beim Listenweisen Fallausschluss. Jedoch sollte Paarweiser Fallausschluss nur in Ausnahmefällen angewendet werden.

4.4 Vergleich Listenweiser - und Paarweiser Fallausschluss

An dieser Stelle sollen dazu drei Variablen mit je 10.000 normalverteilten Beobachtungen mit folgenden Parametern generiert werden.

$$\mu = (30,30,30)$$

$$\Sigma = \begin{pmatrix} 1 & 0.50 & 0.50 \\ 0.50 & 1 & 0.50 \\ 0.50 & 0.50 & 1 \end{pmatrix}$$

Es werden dann völlig zufällig 25 Prozent fehlende Werte in den 3 Variablen generiert. Hierbei sei angemerkt, dass beide Verfahren bei MCAR in der Lage waren die Mittelwerte und die Korrelationsmatrizen aus 2.4 genau zu schätzen.

4.4.1 Simulation MAR - Listenweiser-Fallausschluss

Im Folgenden werden die fehlenden Werte nach dem in der Praxis eher vorkommenden Mechanismus Missing-at-Random (MAR) generiert. Hierfür wurden in X und Y erneut zufällig 25 Prozent fehlende Werte generiert und in Z alle Beobachtungen gelöscht, die eine Ausprägung größer als 35 in Y aufweisen (entspricht ca. 25 Prozent der Daten in Y).

$$\mu = (30,30,27)$$

És lässt sich zeigen, dass bei Schätzung mit Listenweisen Fallausschluss sowohl die Korrelation zwischen Z und X sowie Z und Y unterschätzt wird, als auch die Korrelation zwischen X und Y.

$$\Sigma = \begin{pmatrix} 1 & 0.36 & 0.43 \\ 0.36 & 1 & 0.36 \\ 0.43 & 0.36 & 1 \end{pmatrix}$$

4.4.2 Simulation MAR - Paarweiser Fallausschluss

$$\mu = (30,30,27)$$

Dieses Problem ist nicht so gravierend bei der Schätzung mit den paarweise kompletten Beobachtungen. Der Paarweise Fallausschluss kommt näher an die ursprüngliche Korrelationsmatrix. Es kann sich also durchaus, lohnen beide Verfahren anzuwenden und die Ergebnisse mit der höheren Plausibilität näher zu untersuchen.

$$\Sigma = \begin{pmatrix} 1 & 0.49 & 0.46 \\ 0.49 & 1 & 0.37 \\ 0.46 & 0.37 & 1 \end{pmatrix}$$

Es sei jedoch auf die bekannten Probleme des Paarweisen Fallausschlusses verwiesen.

5 Single Imputations Techniken

Bei der Ersetzung durch Imputation ist die Grundidee, fehlende Daten durch plausible Werte aufzufüllen. Dabei geht es jedoch nicht vorrangig darum, die wahren Werte zu rekonstruieren. Das Hauptaugenmerk liegt vielmehr darin einen Datensatz zu erzeugen, der bei den weiteren Analysen für das Schätzen von zentralen Parametern wie Mittelwerten, Standardabweichungen, Korrelationen etc. gute Ergebnisse liefert. Man unterscheidet bei den Imputationstechniken zwischen Single-Imputation und Multiple-Imputation. Unter Umständen kann das Verwenden von Imputationen zu besseren Ergebnissen führen als Fallweiser Ausschluss, da mehr Beobachtungen und somit auch mehr Informationen zur Verfügung stehen. Durch die Imputation entsteht ein kompletter Datensatz, der sich nun mit herkömmlichen statistischen Verfahren analysieren lässt. Jedoch sollte nicht dem Trugschluss verfallen werden, dass man die Analysen nun unbesorgt wie mit einem Datensatz ohne fehlende Daten durchgeführt werden können. Die Probleme wurden von Little und Rubin (1987) und Dempster und Rubin (1983) ausführlich dokumentiert. Man unterscheidet des Weiteren zwischen explizierter Modellierung, welche auf einem formalen statistischen Modell mit expliziten Annahmen basiert und der implizierter Modellierung, bei dieser liegt der Fokus auf modellbasierten implizierten Annahmen. Methoden mit expliziten Annahmen sind z.B Mittelwertersetzung, Medianersetzung, Regressionsimputation oder der stochastischen Regressionsimputation. Methoden mit implizierter Modellierung sind z.B Hot-Deck oder Cold-Deck.

5.1 Einsatzmöglichkeiten von Software

Für die Schätzung von fehlenden Daten gibt es eine Reihe von kostenloser aber auch kommerzieller Software. Dabei sind die Voraussetzungen an die Vorkenntnisse der Nutzer sehr verschieden. Relativ einfach zu Handhaben sind Programme wie AMOS, MPLUS, Solas, BMDP oder das Missing-Value-Analysis Modul der SPSS Software. Die ersten beiden Programme werden vor allem zur Schätzung von Strukturgleichungsmodellen verwendet. Der Nachteil bei den erwähnten Programmen ist vor allem ein hoher Anschaffungspreis. In der kommerziellen Software SAS ist unter anderem eine Multiple-Imputation Prozedur verfügbar. Für SAS und Stata existieren Makros für zahlreiche Verfahren, die im Internet verfügbar sind. Allerdings verlangen diese Programme fortgeschrittene Programmierkenntnisse. Unter Windows sind einige kostenlose eigenständige Programme wie NORM oder MICE von Joseph Schafer verfügbar. Die Bedienung der Programme erweist sich als einfach. Auf die Verfahren wird später noch näher eingangen. Will man alle gängigen Verfahren in einer freien Software zur Verfügung haben, so kann man auf die Software R zurückgreifen. Hier gibt es ebenfalls Bibliotheken speziell für Maximum-Likelihood Schätzungen für fehlende kategorielle und gemischt kategorielle Daten. Soweit mir bekannt ist, sind diese Verfahren insbesondere für kategorielle Daten in keiner anderen Software implementiert. Oft ist es auch sinnvoll, vor der Benutzung der Software eine Transformationen der Variablen durchzuführen. Eine häufige Annahme der

Verfahren ist eine (multivariate) Normalverteilung in den Daten. Durch Logarithmieren, Ziehen der Quadratwurzel oder einer logistischen Transformation der Variablen kann die Verzerrung oft beseitigt werden, vgl. auch Allison (2001). In Tabelle 12 sind aktuelle Softwareprogramme abgebildet.

Software	Methode	Annahmen	Kommerziell	Bemerkungen
SPSS Base	MW[1]	MCAR	Ja	Standalone
SPSS MVA	EM	MAR	Ja	Zusatzmodul[2]
SAS Proc MI	MI[3]	MAR	Ja	STAT Modul benötigt
AMOS	FIML[4]	MAR	Ja	SEM
MPLUS	FIML[4]	MAR	Ja	SEM
Amelia I	MI[5]	MAR	Frei	Standalone
Amelia II	MI[5]	MAR	Frei	R-Modul
Norm	EM	MAR	Frei	Standalone
Norm	EM	MAR	Frei	R-Modul
MICE	MI[3]	MAR	Frei	Standalone
MICE	MI[3]	MAR	Frei	R-Modul
BMDP	EM, ML[4]	MCAR,MAR	Ja	Standalone
Stata	REG[6]	MAR	Ja	Standalone
MX	ML[4]	MAR	Frei	Standalone
Solas	MI	MAR	Ja	Standalone
CAT[7]	EM[8]	MAR	Frei	R-Modul
SeqKnn	HD[9]	MCAR	Frei	R-Modul
MIX[10]	EM[11]	MAR	Frei	R-Modul

Tabelle 12: Software für Missings Data

In diesem Kapitel erfolgt die Erläuterung der wichtigsten Verfahren. Zusätzlich wird mit Hilfe einer Simulationstudie die Leistungsfähigkeit von einigen Verfahren überprüft. Anhand dieser Ergebnisse sollen dem Leser dann auch Anwendungsempfehlungen gegeben werden.

[1] Mittelwertersetzung
[2] Zusatzmodul für SPSS, Standardfehler und Teststatistiken sind stark verzerrt, vgl v. Hippel (2004).
[3] Multiple-Imputation EM-Algorithmus und Data-Augmentation und MCMC, Multivariate Normalverteilung, verschiedene SAS Makros sind im Internet verfügbar
[4] Full-Information-Maximum-Likelihood, Multivariate Normalverteilung wird vorausgesetzt
[5] Multiple-Imputation mit einer Technik die sich 'importance sampling' nennt, Multivariate Normalverteilung wird auch hier vorausgesetzt
[6] Standard in Stata ist ein Verfahren das Werte mittels einer Regressionsersetzung imputiert, zusätzliche Makros sind auf der Homepage verfügbar
[7] Handhabung von kategoriellen Daten
[8] EM-Algorithmus für kategoriellen Daten
[9] Sequentielle Hot-Deck Verfahren
[10] Handhabung von gemischt kategoriellen Daten
[11] EM-Algorithmus für gemischt kategorielle Daten

5.2 Mittelwert Imputation

Die Grundidee bei der Mittelwert Imputation ist das Ersetzen von fehlenden Daten in einer Variablen Y_j durch den Mittelwert der vorhandenen Beobachtungen in dieser Variablen. Dabei wird offensichtlich, dass insbesondere bei nicht normalverteilten Daten die neue Verteilung eine inkorrekte Repräsentation der Originaldaten darstellt. Durch die Ersetzung mit nur einem Wert in der Mitte der Verteilung werden Parameter wie die Varianz oder die Schiefe einer Verteilung unterschätzt. Die Varianz wird laut Little und Rubin (2002) um den Faktor $\frac{(n^{(j)}-1)}{(n-1)}$ unterschätzt. Wobei $n^{(j)}$ die Anzahl der kompletten Beobachtungen darstellt. Die Kovarianz der aufgefüllten Daten $\tilde{s}_{jk}^{(jk)}$ zwischen den Variablen Y_j und Y_k wird um den Faktor $\frac{(n^{(jk)}-1)}{(n-1)}$ unterschätzt. $n^{(jk)}$ stellt die Anzahl der Beobachtungen dar, die in beiden Variablen komplett sind. Durch eine Rückgewichtung kann diesem Problem entgegengewirkt werden. Jedoch können diese Vorgehen ähnlich wie beim Paarweisen Fallausschluss zu nicht mehr positiv definiten Korrelationsmatrizen führen. Für die folgenden Grafiken wurden in der Variablen Y_2 (Punkte Statistik 2) zufällig zirka fünfzig Prozent fehlende Werte erzeugt. Die fehlenden Werte wurden durch den Mittelwert der verbleibenden Daten in Y_2 ersetzt.

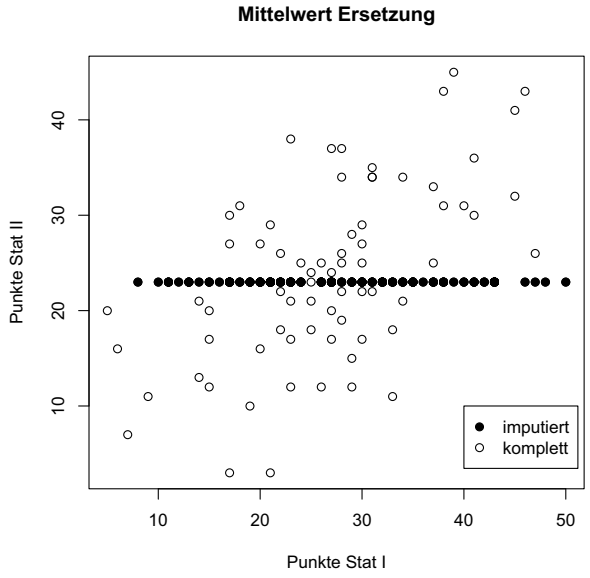

Abbildung 8: Mittelwert Imputation

Bei der Mittelwertersetzung lässt sich die ‚Zerstörung ‚der Verteilung durch die Imputation eines gleichen Wertes ersehen.

5.3 Median Imputation

Da Mittelwerte sehr stark durch Ausreißer beeinflusst werden, bietet es sich an den Median aufgrund der Robustheit zu verwenden. Bei kategoriellen Daten wird auf den Modus zurückgegriffen. Die Hauptprobleme der Mittelwertersetzung bleiben jedoch bestehen und daher ist dieses Verfahren keine echte Alternative.

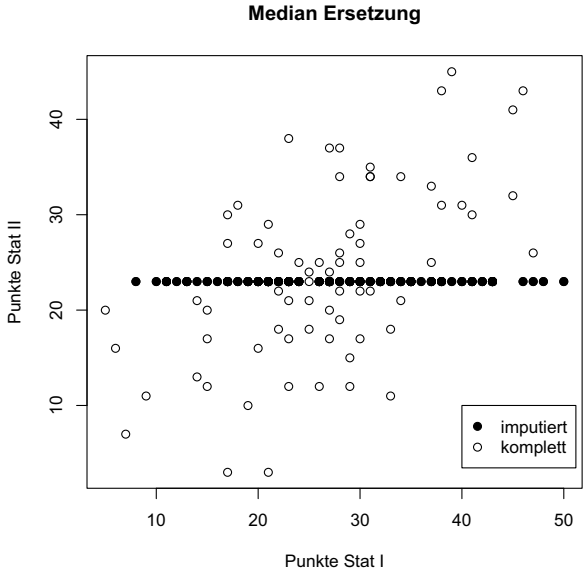

Abbildung 9: Median Imputation

Bei der Medianersetzung erkennt man dasselbe Problem wie bei der Mittelwertersetzung.

5.4 Regressionsersetzung

Die Imputation durch Regression kann besonders sinnvoll sein, wenn es einen starken linearen Zusammenhang zwischen den Variablen gibt. Ist jedoch kein Zusammenhang vorhanden, so reduziert sich das Modell zur Mittelwertersetzung. Im Falle von k Variablen mit n-r fehlenden Werten in der k-ten Variablen wird eine lineare Regressionsanalyse mit den r kompletten Beobachtungen geschätzt. Anhand der Koeffizienten der geschätzten Regressionsfunktion

$$\tilde{y}_{ik} = \tilde{\beta}_0 + \sum_{j=1}^{k-1} \tilde{\beta}_j y_{ij} \quad \forall \;\; i \in [r, n] \tag{5.1}$$

werden die fehlenden Werte nun in der k-ten Variablen imputiert. Für kategorielle oder gemischt kategorielle Daten kann an dieser Stelle auch an eine logistische Regression durchführt werden. Ein größeres Problem ist jedoch, dass die Regression nach einem univariaten aber zumindestens einem monotonem Muster fehlender Daten verlangt. Viele Statistikprogramme benutzen bei der Schätzung der Regressionskoeffizienten den Fallweisen Ausschluss, so dass die Stichprobengröße stark sinken kann.

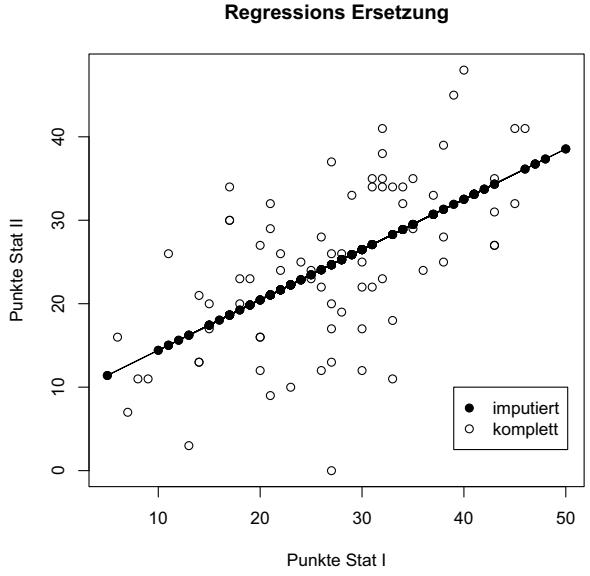

Abbildung 10: Regressionsersetzung

In der Praxis kann es somit bei einem zufälligen Muster fehlender Daten problematisch sein, mehrere Variablen in ein Regressionsmodell einfließen zu lassen. Für den einfachen Fall mit zwei Variablen Y_1 und Y_2 vereinfacht sich das Modell wie folgt

$$\tilde{y}_{i2} = \tilde{\beta}_0 + \tilde{\beta}_1 y_{i1} \quad \forall \ \ i \in [r, n]. \tag{5.2}$$

Die fehlenden Werte in Y_2 werden nun durch die Schätzungen aus dem linearen Modell ersetzt. In Abbildung 10 lassen sich ähnliche Probleme erkennen, wie auch bei der Mittelwertersetzung. Jetzt liegen die imputierten Werte auf der Regressionsgeraden, anstatt auf einer horizontalen Linie. Problematisch bleibt, dass die Regressionsparameter nur aus den Beobachtungen geschätzt werden können, die in beiden Variablen komplett vorhanden sind. Buck (1960) hat vorgeschlagen dieses Problem zu lösen, indem man die Regressionskoeffizienten nicht mit den Daten, sondern basierend auf der Korrelationsmatrix schätzt. Für den Fall von zwei Variablen X und Y lassen sich die Regressionskoeffizienten auch wie folgt berechnen.

$$\tilde{\beta}_0 = r_{x,y} \cdot \frac{S_y}{S_x} \qquad \tilde{\beta}_1 = r_{x,y} \cdot \frac{S_x}{S_y} \tag{5.3}$$

Die Berechnung von Standardabweichungen und Korrelation erfolgt z.B anhand des Fallweisen Ausschluss. Mit Hilfe der geschätzten Regressionskoeffizienten wird nun für jeden fehlenden Wert in einer Beobachtung über die vorhandenen Variablen der bedingte Mittelwert berechnet. Technisch umgesetzt wird diese Methode anhand des so genannten Sweep Operator, auf den hier nicht weiter eingegangen werden soll. Vgl. auch Schafer (1997). Diese Methode nach Buck (1960) ist ein Vorläufer des so genannten EM-Algorithmus auf den später ausführlicher eingegangen wird.

5.5 Stochastische Regression

Stochastische Regression ist die bereits bekannte Regressionsersetzung, nur mit einem zusätzlichen Fehlerterm.

$$\tilde{y}_{ik} = \tilde{\beta}_0 + \sum_{j=1}^{k-1} \tilde{\beta}_j y_{ij} + z \quad \forall \ \ i \in [r, n] \tag{5.4}$$

Dieses Verfahren wird auch als stochastische Regression bezeichnet. z stellt hierbei einen Zufallszug aus den normalverteilten Residuen der Regression von Y_k auf $Y_1....Y_{k-1}$ dar. Die praktische Umsetzung erfolgt durch Generierung einer normalverteilten Zufallsvariablen mit Mittelwert 0 und der Varianz der Residuen. Die Varianz der Residuen basiert auf der Regression der vollständigen Beobachtungen von Y_k auf $Y_1....Y_{k-1}$. Sollte ein zu vorhergesagter Wert außerhalb eines eventuell vorher definierten Bereichs liegen, so wird dieser Zug abgelehnt und ein neuer Zug wird durchgeführt.

An dieser Stelle mag sich der Leser fragen, ob diese Idee sinnvoll ist, da die einfache lineare Regression der bessere Punktschätzer ist. Vgl Little und Rubin (2002). Die stochastische Regression wird aber der wirklichen Variabilität der Daten eher gerecht. Schafer und Graham (2002) beschreiben, dass diese Methode nahezu unverzerrte Ergebnisse unter der MAR Annahme erreicht.

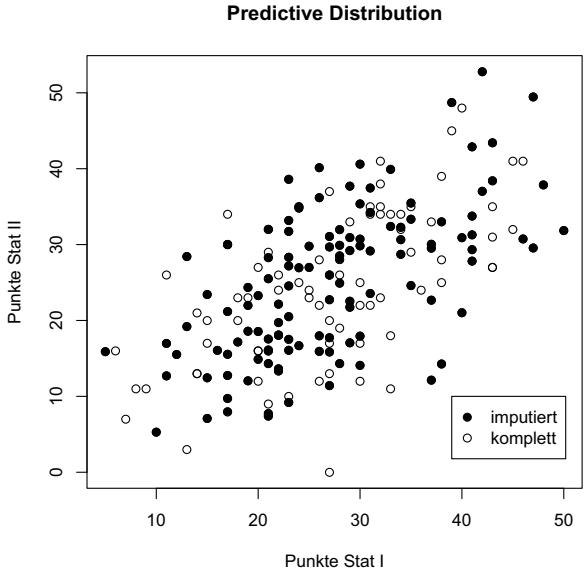

Abbildung 11: Stochastische Regression

Wie man in Abbildung 11 erkennen kann, scheint die Schätzung mit Hilfe der stochastischen Regression der Variabilität der Daten besser Rechnung zu tragen.

5.6 Verfahren basierend auf der Likelihood-Funktion

Diese Verfahren gehören zu den aktuellsten auf dem Gebiet zur Thematik. Die sogenannte Likelihood-Funktion kann man als Funktion von unbekannten Parametern θ auffassen. Dieses können Parameter wie Mittelwerte, Varianzen, die Schiefe oder die Kurtosis sein. Für den einfachen Fall einer unterstellten Normalverteilung interessieren uns vorallem $\theta = (\mu, \Sigma)$, da der Mittelwert und Varianz für Daten mit dieser Form eine hinreichend genaue Beschreibung liefert. Diese Likelihood-Funktion wird dann logarithmiert und wir erhalten die sogenannte loglikelihood. Da der Logarithmus eine monotone Transformation ist, ändert sich auch die Lage der Extremstelle nicht. Der bekannteste ML Schätzer ist der kleinste Quadrate Schätzer (KQ) der linearen Regressionsanalyse.

5.6.1 Parameterschätzungen bei vollständigen Daten

Für den Fall vollständiger Daten können wir die Maximum-Likelihood-Schätzungen für wichtige Parameter wie folgt erhalten. Y sei i.i.d multivariat normalverteilt mit p Variablen, die likelihood Funktion kann wie folgt faktorisiert werden.

$$L\left(Y, \mu, \Sigma\right) = |2\pi\Sigma|^{-\frac{np}{2}} exp - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^{\top}\Sigma^{-1}\left(x_i - \mu\right) \tag{5.5}$$

Logarithmieren von Gleichung 5.5 ergibt

$$l\left(Y, \mu, \Sigma\right) = -\frac{n}{2} log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^{\top}\Sigma^{-1}\left(x_i - \mu\right). \tag{5.6}$$

Die Werte für μ und Σ die diese loglikelihood Funktion aus Gleichung 5.6 maximieren sind

$$\hat{\mu} = \bar{y} \text{ und } \hat{\Sigma} = S.$$

$\bar{y}$ stellt einen Vektor bestehend aus den p Mittelwerte und S die empirische Kovarianzmatrix dar.

5.6.2 Parameterschätzungen bei unvollständigen Daten

Bei einem Datensatz mit fehlenden Werten stellen $\hat{\mu} = \bar{y}$ und $\hat{\Sigma} = S$ jedoch nicht mehr unbedingt eine Maximum-Likelihood-Schätzung dar. Die Schätzungen der Parameter aus den vollständigen Beobachtungen sind verzerrt. ML behandelt fehlende Daten als Zufallsvariablen, die aus der Likelihood-Funktion entfernt werden müssen (z.B durch Herausintegrieren), so als ob sie niemals vorhanden

wären vgl. Graham und Schafer (2002) und Allison (2002). Falls die fehlenden Daten MAR sind, kann die likelihood wie folgt faktorisiert werden.

$$L\left(\theta|Y_{obs}\right) = \left(Y_{obs}, Y_{mis}|\theta\right)dY_{mis} \tag{5.7}$$

Diese Likelihood-Funktion kann oft nur mit großem Rechenaufwand und numerischen Verfahren maximiert werden. Es sind Verfahren wie FIML oder der EM Algorithmus entwickelt worden, die unter bestimmten Bedingungen nahezu Maximum-Likelihood-Schätzungen liefern.

5.6.3 Full Information Maximum Likelihood Schätzer

In Softwarepaketen wie AMOS und MPLUS, die zur Berechnung von Strukturgleichungsmodellen entwickelt worden sind, ist der sogenannte FIML Schätzer verfügbar. Als einzige Annahmen werden hier MAR und die multivariate Normalverteilung vorausgesetzt. Enders (2001) beschreibt den Ablauf wie folgt. Der FIML Schätzer maximiert eine Likelihood-Funktion, die aus der Summe von n (fallweisen) Likelihood-Funktionen besteht. So wird für jede Beobachtung eine separate likelihood berechnet, die eine Diskrepanz zwischen den beobachteten Daten für die i-te Beobachtung und den aktuellen Parameterschätzungen misst. Für den Fall von multivariater Normalverteilung wird die folgende Funktion maximiert.

$$l_i\left(\mu_i, \Sigma_i\right) = K_i - \frac{1}{2}log|\Sigma_i| - \frac{1}{2}(x_i - \mu_i)^\top \Sigma_i^{-1}(x_i - \mu_i) \tag{5.8}$$

Dies geschieht mit fehlenden Daten folgendermaßen: x_i ist ein Vektor der vollständigen Daten von Beobachtung i und μ_i ist der Vektor der Mittelwertschätzungen der jeweiligen Variablen die in i beobachtetet wird, K_i ist eine Konstante deren Wert von der Anzahl der kompletten Datenpunkte in Beobachtung i abhängt. Die Berechnung der Determinante von Σ_i erfolgt ebenfalls nur aus den beobachteten Variablen aus Fall i. Summiert über alle n Beobachtungen erhält man die Gleichung 5.9.

$$l(Y, \mu, \Sigma) = \sum_{i=1}^{N} l_i \tag{5.9}$$

Für den Prozess der FIML Schätzung wird angenommen, dass Variablen die fehlende Werte aufweisen von anderen beobachteten Variablen abhängen. Konzeptionell ist der FIML Schätzer somit ähnlich zu einem Algorithmus der mittels Regression fehlenden Daten imputiert. Es ist jedoch zu bemerken, dass der Algorithmus keinerlei Daten imputiert, sondern nur alle verfügbaren Rohdaten ausnutzt um die zentralen Parameter zu schätzen. Enders (2001) beschreibt, dass der FIML Schätzer in AMOS und MPLUS dieselben Punktschätzungen für die Modellparameter liefern, wie der im nächsten Abschnitt beschriebene EM Algorithmus. Er verweist jedoch auch explizit auf Graham et al. (1997) die zeigen, dass die Verwendung einer mit Hilfe des EM Algorithmus geschätzten Kovarianzmatrix in AMOS oder MPLUS für weitere Analysen problematisch

ist. Sie führen dies auf verzerrte Standardfehler zurück, die aufgrund falscher Annahmen über die Stichprobengrösse beim EM Algorithmus zustandekommen können. Enders (2001) empfiehlt bei der Analyse von Strukturgleichungsmodellen in MPLUS und AMOS die dort implementierten FIML Schätzer zu benutzen.

5.6.4 EM Algorithmus

Der Expectation-Maximization-Algorithmus ist eine der aktuelle Methode zur Schätzung von fehlenden Werten. Der EM Algorithmus in seiner heutigen Form, wurde von Dempster et al. (1977) vorgestellt. Er ist besonders seit den 90er Jahren ein fester Bestandteil bei der effizienten Schätzung von fehlenden Daten. Populär wurde der EM durch die Implementierung in die Software R durch Schafer (1997) und SAS durch Y.C Yuan (2000). Der EM Algorithmus ist ein iterativer Algorithmus zur Maximum-Likelihood Schätzung bei Datensätzen mit fehlenden Daten. Er besteht aus 2 Schritten die iterativ wiederholt werden. Diese Schritte sind der sogenannten E-Schritt (Erwartung) und der M-Schritt (Maximierung). Beim EM werden im Regelfall eine ganze Reihe von Iterationen benötigt. Der Ablauf kann laut Little und Rubin (2002) wie folgt beschrieben werden:

1. Ersetze fehlende Werte durch plausible Startwerte.
2. Schätze die Parameter $\theta^{(t)}$
3. Ersetze die fehlenden Werte unter der Annahme, dass die geschätzen Parameter $\theta^{(t)}$ korrekt sind.
4. Schätze die Parameter $\theta^{(t+1)}$ und gehe zum vorigen Punkt zurück.
5. Iteriere bis zur Konvergenz d.h $\theta^{(t+1)} - \theta^{(t)} \leq \epsilon$.

Im E-Schritt werden die fehlenden Daten ersetzt und im M-Schritt die Parameter aktualisiert. Little und Rubin (2002) zeigen, dass unter bestimmten Bedingungen wie einer linearen loglikelihood der EM Algorithmus zuverlässig konvergiert. Das heisst jede Iteration erhöht die loglikelihood $l(\theta_{,obs})$. Wenn die Sequenz von $\theta^{(t)}$ konvergiert, so konvergiert diese auch zu einem lokalem Maximum. Der Beweis ist unter anderem in McLachlan und Krishnan (1997) zu finden. Ein Nachteil stellt jedoch die unter Umständen langsame Rate der Konvergenz dar. In der Literatur finden sich verschiedene Ansätze zur Erhöhung der Geschwindigkeit des Algorithmus durch Verbindung mit anderen Algorithmen wie Newton-Raphson oder dem Scoring-Algorithmus, vgl. Little und Rubin (2002, S. 186). Eine Implementierung in einer statistischen Software ist mir jedoch nicht bekannt. Eine praktische Umsetzung kann für einen Datensatz mit p Variablen und zufälligen Muster fehlender Daten mittels einfacher linearer Regression erfolgen. Eine Voraussetzung ist die multivariate Normalverteilung. Im ersten Schritt wird mittels Fallweisen oder Paarweisen Ausschluss die Kovarianzmatrix geschätzt. Für jedes Muster fehlender Daten werden fehlende

Werte unter Ausnutzung aller zur Verfügung stehenden Informationen (Variablen) mittels linearer Regression imputiert. Die Startwerte für Mittelwerte und Kovarianzen erhält man mit Hilfe von Fallweisen oder Paarweisen Ausschluss. Auf die jeweilige Schätzung wird ein Zufallszug aus der Residualverteilung der Regression addiert. Durch den Zufallszug ändern sich bei jeder Iteration die Schätzungen leicht. Der Algorithmus konvergiert dann wie oben beschrieben in endlichen Schritten oder bricht ab. Der Standard EM Algorithmus, wie er zum Beispiel in SPSS implementiert ist, wurde in den letzten Jahren durch Methoden wie Data Augmentation verbessert. Auf diese Methode soll nun im folgenden Abschnitt eingegangen werden. In der Abbildung 12 erkennt man einen Standard EM Algorithmus in SPSS ohne Zufallszug aus den Residuen.

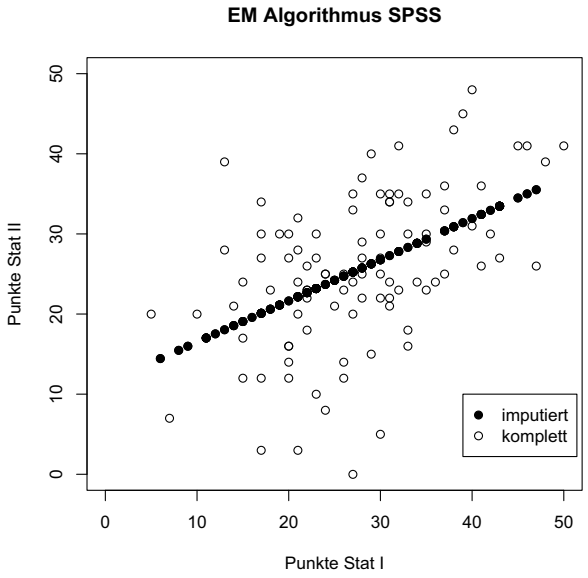

Abbildung 12: EM Algorithmus in SPSS

5.6.5 Data-Augmentation

Data-Augmentation (engl. to augment) bedeutet übersetzt in etwa „Datenvermehrung“. Data-Augmentation wurden von Tanner und Wong (1987) als iteratives Verfahren zur Simulation der a-posteriori Verteilungen von θ entwickelt. Es verbindet Merkmale von EM und von Multiple-Imputation. Allison (2002) beschreibt den Algorithmus wie folgt. Wähle zuerst eine Reihe von Variablen für den Imputationsprozess aus. Diese können sowohl aus Variablen mit fehlenden Daten, als auch aus den vollständigen Variablen stammen. Falls man Variablen zur Verfügung hat die man für die weiteren Analysen nicht benötigt, die aber eine hohe Korrelation zu den Variablen mit fehlenden Werten aufweisen, so sollte man diese ebenfalls verwenden. Da die Anzahl der verschiedenen Muster für große Datensätze leicht einige hundert erreichen kann, sollte man um den Rechenaufwand gering zu halten jeweils nur einige wenige korrelierte Variablen an den Algorithmus übergeben. Sind die Variablen ausgewählt, gliedert sich der Ablauf nach Allison (2001) wie folgt:

1. Wähle plausible Startwerte für die Parameter aus. Für multivariate Normalverteilung sind μ und Σ ausreichend. Die Parameter können vorher durch Listenweisen, Paarweisen oder am besten dem EM Algorithmus geschätzt worden sein.

2. Anhand dieser Parameter schätzt man die Regressionskoeffizienten, in dem jede Variable mit fehlenden Werten auf die vollständigen Variablen regressiert wird. So werden möglichst viele Informationen einbezogen. Dies wird für alle verschiedenen Muster fehlender Daten abgearbeitet.

3. Anhand dieser Regressionskoeffizienten werden Imputationen für alle fehlenden Daten durchgeführt. Zu jedem vorhergesagten Wert wird ein Zufallszug aus den normalverteilen Residuen hinzugefügt (I-Schritt).

4. Benutze den nun kompletten Datensatz um die Mittelwerte und Kovarianzen erneut zu schätzen.

5. Basierend auf den im letzten Schritt geschätzten Mittelwerten und Kovarianzen wird nun ein Zufallszug aus der sogenannten a-posteriori Verteilung der Mittelwerte und Kovarianzen durchgeführt.(P-Schritt)

6. Mit den aus Schritt 5 erhaltenen Mittelwerten und Kovarianzen geht man nun wieder zu Schritt 2 über. Diese Prozedur wird solange wiederholt, bis der Algorithmus konvergiert. Der Datensatz aus dem letzten Schritt wird verwendet.

In der Praxis findet man sich oft in einem gewissen Stadium der Unsicherheit. Die unter Punkt 5 beschriebende Ziehung aus der a-posteriori Verteilung der Parameter θ gehört in den Bereich der baysianischen Statistik. In der Statistik gibt es zwei Gruppen von Forschern, die sogenannten Frequentisten und die Baysianer. Nach Interpretation der Frequentisten ist Wahrscheinlichkeit der Grenzwert

der relativen Häufigkeit, in der baysianischen Herangehensweise wird Wahrscheinlichkeit als Grad von persönlicher Überzeugung definiert. Gegeben eine Menge von Messungen, Informationen oder Datenpunken können sich Wahrscheinlichkeiten ändern. Es werden alle unbekannten Parameter als Zufallsvariablen behandelt und den Parametern wird eine gemeinsame Verteilungsfunktion zugeordnet. Dabei wird den Parametern zuerst eine sogenannte a-priori Verteilung zugeordnet. Eine a-priori Verteilung ist eine Verteilung, die die Unsicherheit von Parametern darstellt, bevor Daten vorliegen. Bei der Anwendung von Bayes' Theorem kann man die a-posteriori Verteilung durch Multiplizieren der a-priori Verteilung mit der likelihood berechnen.

$$p(\theta|Y) = \frac{p(\theta)L(Y|\theta)}{p(Y)} \tag{5.10}$$

p(Y) ist hierbei eine Normalisierungskonstante. Die a-posteriori Verteilung spiegelt nun die Wahrscheinlichkeitsverteilung von θ basierend auf den beobachteten Daten wieder. Die a-posteriori Verteilung für ein Modell mit ignorierbarem Mechanismus fehlender Daten wird wie folgt definiert:

$$p(\theta|Y_{obs}) = Konstante \cdot p(\theta) \cdot f(Y_{obs}|\theta) \tag{5.11}$$

wobei $p(\theta)$ die a-priori Verteilung und $f(Y_{obs}|\theta)$ die Dichte der beobachteten Daten ist. Diese a-priori Verteilung ist oft jedoch eine rein subjektive Annahme. Deshalb werden auch oft spezielle a-priori Verteilung zur leichteren Berechnung der a-posteriori Verteilung verwendet. Eine weit verbreitete Strategie ist es, eine geeignete a-priori Verteilung auszuwählen, so dass die a-posteriori Verteilung zur selben Familie wie die a-priori Verteilung gehört. Die Auswahl der Familie hängt jedoch von der likelihood ab. Für mehr Hintergrundinformationen zu dieser Thematik kann D'Agostini (2003) empfohlen werden. In Tabelle 13 sind einige Beispiele dargestellt.

Likelihood $L(Y\|\theta)$	prior $p(\theta)$	posterior $p(\theta\|Y))$
multivariat normalverteilt	Inverse-Wishart	multivariat normalverteilt
normalverteilt	Inverse-Gamma	normalverteilt
multinomial	Derichlet	Derichlet
binomial	beta	beta
uniform	pareto	pareto

Tabelle 13: Verteilungsauswahl

In der Abbildung 13 ist Data-Augmentation aus der SPLUS/R Bibliothek NORM und der Software SAS 9.1 dargestellt .

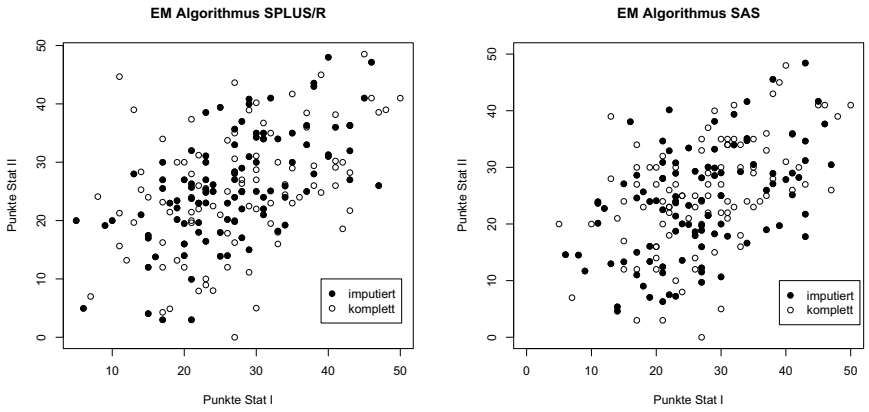

Abbildung 13: EM Algorithmus in SPLUS/R und SAS

5.6.6 Markov-Chain-Monte-Carlo Methode

Ein ist oft ein Problem, eine Folge von Ziehungen von zwei oder mehreren Zufallsvariablen aus der gemeinsamen Wahrscheinlichkeitsverteilung der beobachteten Daten $f(Y_{obs}|\theta)$ zu erhalten. Ein spezieller Ansatz, die Markov-Chain-Monte-Carlo Methode generiert (zufällige) Ziehungen von den a-posteriori Verteilungen der Parameter. Eine Markov Kette als Sequenz von Zufallsvariablen hat die Eigenschaft, dass die Ziehung eines Parameters nur von der Ziehung des vorigen abhängt. Beim Durchlaufen dieser Sequenz, unter der Annahme, dass bestimmte Stabilitätskriterien erfüllt sind, konvergiert die Verteilung der Elemete in eine stationäre Verteilung. Bei der Markov-Chain-Monte-Carlo Methode kann durch das Ziehen der Parameter auf diese Weise die a-posteriori Verteilung der wahren Parameter simuliert werden. Gibbs Sampling ist ein Algorithmus, der versucht diese unbekannte gemeinsame Verteilung zu aproximieren. Gibbs-Sampling eignet sich besonders dann, wenn die gemeinsame Verteilung eines Zufallsvektors unbekannt, die bedingte Verteilung einer jeden Zufallsvariable jedoch bekannt ist. Das Grundprinzip besteht darin, iterativ eine Variable auszuwählen und gemäß ihrer bedingten Verteilung einen Wert in Abhängigkeit von den Werten der anderen Variablen zu erzeugen. Die Werte der anderen Variablen bleiben in diesem Iterationsschritt unverändert. Aus der entstehenden Folge von Stichprobenvektoren lässt sich eine Markow-Kette herleiten. Es kann gezeigt werden, dass die stationäre Verteilung dieser Markow-Kette die gesuchte gemeinsame Verteilung des Zufallsvektors ist.

5.7 Hot-Deck Methoden

Die folgenden Verfahren gehören zu den nichtparametrischen Verfahren. Man unterscheidet zwischen Hot-Deck und Cold-Deck. Der Name Hot-Deck geht auf das Lesen von Daten von sogenannten Lochkarten für IBM Computer zurück. Diese Lochkarten wurden sehr heiß, sobald Daten angefragt wurden. Hot-Deck steht somit für das Ersetzen von fehlenden Werten aus einem aktuellen Datensatz. Bei der Cold-Deck Imputation hingegen werden fehlenden Werte durch Werte aus einer externe Quelle ersetzt. Für mehr Informationen sei auf Kalton and Kasprzyk (1986) verwiesen.

5.7.1 Einfaches Hot-Deck

Die Grundidee ist es, fehlende Werte durch Ziehung aus der Verteilung der beobachteten Daten zu ersetzen. Man spricht hier auch von der Verteilungsimputation. Die einfachste Herangehensweise ist es, für jede Variable Y_i mit $n - r$ Beobachtungen die fehlenden Werte aufweisen eine Zufallsziehung (mit Zurücklegen) aus den vorhandenen r Beobachtungen durchzuführen. Diese Methode wird auch Bootstrapping genannt. So hat jeder Wert der in den beobachteten Daten vorhanden ist, eine Wahrscheinlichkeit in Abhängigkeit von seiner Häufigkeit imputiert zu werden. Diese Methode kann somit problematisch werden, falls nicht alle Merkmalsausprägungen in den beobachteten Daten zur Verfügung stehen, oder eine Merkmalsausprägungen besonders häufig ist.

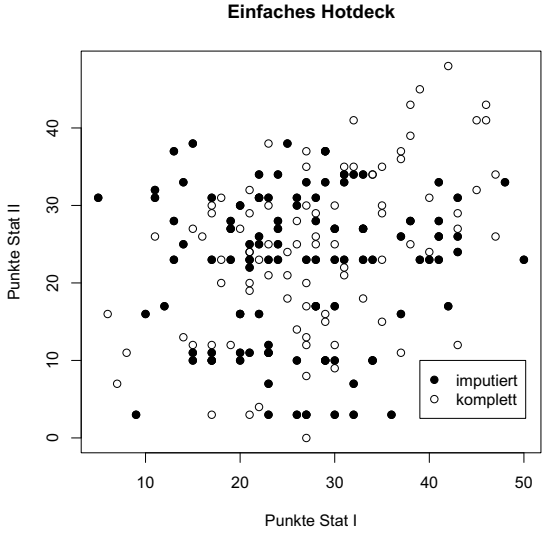

Abbildung 14: Verteilungsimputation

Eine etwas geringere Variabilität der Schätzungen gegenüber der stochastischen Regression ist in Abbildung 14 zu erkennen.

5.7.2 Nächste-Nachbar-Verfahren

Mit Hilfe des k-NN Hot-Deck Verfahren werden fehlende Werte durch Werte aus besonders „ähnlichen“ Beobachtungen imputiert. Es gibt Verfahren die neben kontinuierliche auch für kategorielle Daten geeignet sind. Auch ist es nicht nötig, explizite Modelle aufzustellen. Ist ein fehlender Wert in einer Beobachtung vorhanden, so durchsucht der Algorithmus den gesamten Datensatz nach den k „ähnlichsten“ und Beobachtungen und imputiert dann den Mittelwert aus diesen k Werten. Diese Suche kann bei grossen Datensätzen jedoch sehr zeitaufwendig sein. Es gibt verschiedene Verfahren. An dieser Stelle soll der Algorithmus für kontinuierliche Daten von Kim et al. (2004) erläutert werden. Dieser Algorithmus ist in der R-Bibliothek „SeqKnn“ verfügbar.

1. Unterteile den Datensatz Y in einen Datensatz Y_{com} der nur aus vollständige Beobachtungen besteht und einen Datensatz Y_{mis} der Beobachtungen mit mindestens einem fehlenden Wert aufweist

2. Die unvollständigen Beobachtungen werden nach der Anzahl der fehlenden Werte sortiert

3. Ein fehlender Wert der Beoachtung i in der Spalte j wird durch den Mittelwert der ähnlichsten k Beobachtungen aus den kompletten Daten ersetzt. Sobald die Beobachtung aufgefüllt ist wird sie nach Y_{com} verschoben.

Um die Rechenzeit zu minimieren werden alle fehlenden Werte in einer Beobachtung i simultan ersetzt. Dadurch ist dieses Verfahren auch für sehr große Datensätze geeignet. Zur Berechnung der Ähnlichkeit zweier Beobachtungen $x \in Y_{mis}$ und $y \in Y_{com}$ wird in der R-Bibliothek „SeqKnn“die Euklidische Distanz vewendet.

$$L_2(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5.12}$$

Die Distanz wird aus allen Variablen ermittelt. Aus Theorie und Praxis kennt man noch viele andere Distanzmaße wie Mahalanobis Distanz, Minkowski Metrik, Manhattan-Distanz u.v.a. Ein weiteres Problem stellt die Anzahl der einzubeziehenden k nächsten Beobachtungen dar. Eine zu kleine Anzahl von k kann zu einem übermäßigen Einfluss einzelner Beobachtungen führen. Eine zu grosse Anzahl von k würde unter Umständen zu einer Einbeziehung von zu „unähnlichen“ Beobachtungen führen. Arbeiten zu diesem Thema wurden in Troyanskaya et al. (2001) durchgeführt. Für diese Arbeit wird die Anzahl von $k = 3$ verwendet. Díeses Verfahren ist in der Lage eine gute Variabilität zu liefern. Allerdings konzentrieren sich, wie man in Abbildung 15 erkennen kann viele Punke innerhalb einer Punktwolke.

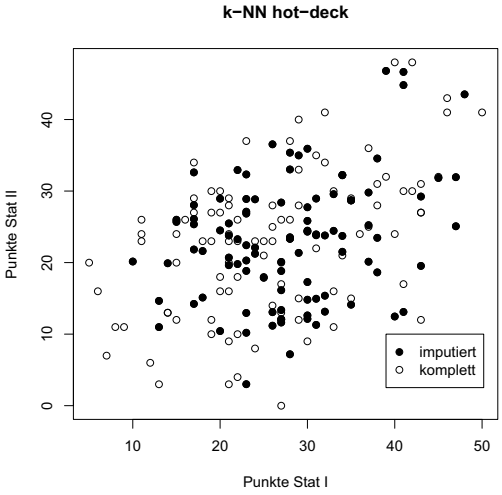

Abbildung 15: k-NN Ersetzung

5.8 Cold-Deck

Diese Methode wird vor allem bei der Auswertung von Umfragen benutzt. Bei der Cold-Deck Imputation werden fehlende Werte durch Werte aus einer externen Quelle ersetzt. Diese Werte können aus früheren oder externen Forschungen entnommen sein. Es können auch Werte die man für plausibel hält imputiert werden, z.B unterschiedliche Einkommenshöhen für besondere Altersklassen. Auch ist es oft möglich, Personen die an den früheren Umfragen teilgenommen haben, anhand bestimmter Angaben zu identifizieren. So wird man oft in Umfragen aufgefordert die ersten Buchstaben von Vor - und Nachnamen anzugeben. Die Anonymität bleibt so gewahrt und man ist unter Umständen in der Lage, fehlende Werte durch frühere Antworten zu ersetzen. Eine generelle Rechtfertigung zur Ersetzung von fehlenden Daten aus externen Quellen bleibt diese Methode jedoch schuldig. Bei der Verwendung der Methode sollte man eine Rechtfertigung haben, warum es besser ist externe Daten zu verwenden.

6 Multiple Imputation

Der Begriff Multiple-Imputation wird seit der erstmaligen Erwähnung von Rubin (1978b) verwendet. Multiple-Imputation basiert auf der einer Anzahl von $D \geq 2$ Zufallsimputationen. Dies führt zu D verschiedenen Datensätzen die sich leicht unterscheiden. Die Single Imputationstechniken besitzen den Vorteil, dass mit nur einem Datensatz gearbeitet wird. Allerdings lässt sich als Nachteil aufzählen, dass der Unsicherheit, die in den imputierten Daten steckt, unzureichend Rechnung getragen wird. Multiple-Imputation ist jedoch kein eigenständiges Verfahren, sondern nutzt die Unsicherheiten und Variabilität anderer Imputationsverfahren wie zum Beispiel des EM Algorithmus aus. Darüber hinaus besteht die Möglichkeit, mehrere Verfahren wie Hot-Deck und EM kombiniert zu verwenden. Die so erzeugten Datensätze können nun mit Standardmethoden analysiert werden. Der Ablauf kann wie folgt zusammengefasst werden.

1. Wähle ein geeignetes Imputationsverfahren aus.
2. Berechne D plausible Schätzungen für jeden fehlenden Wert. Man erhält D vollständige Datensätze.
3. Analysiere diese D Datensätze mit Hilfe von Standardmethoden.
4. Um Aussagen über die Variabilität der Ergebnisse zu treffen, werden die Ergebnisse aus 2 nach Formel 6.1 - 6.4 kombiniert.

Die Analyseschritte haben einen unterschiedlichen Schwierigkeitsgrad und können auch ohne weiteres unabhängig voneinander durchgeführt werden. Um die Ergebnisse der Methoden zu kombinieren wurden von Rubin (1987) folgende Formeln veröffentlicht. Die Angaben sind aus der aktuellsten Ausgabe von Little und Rubin (2002) entnommen.

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \tilde{\theta}_d \tag{6.1}$$

$\tilde{\theta}_d$ stellt hierbei eine Punktschätzung wichtiger Parameter dar. Dies können zum Beispiel Mittelwerte oder Varianzen sein.

Die Variabilität dieser Schätzungen hat zwei Komponenten, die durchschnittliche Varianz innerhalb der Imputationen

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^{D} W_d, \tag{6.2}$$

sowie die Komponente zwischen den Imputationen

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\tilde{\theta}_d - \bar{\theta}_D). \tag{6.3}$$

Die totale Variabilität der Parameter kann folgendermaßen berechnet werden,

$$T_D = \bar{W}_D + \frac{D+1}{D}B_D \tag{6.4}$$

wobei $\frac{D+1}{D}$ ein Anpassungsfaktor für eine endliche Anzahl von D ist.

Für große Stichproben sind Tests und Konfidenzintervalle approximativ t-verteilt

$$(\theta - \bar{\theta}_D)T_d^{-\frac{1}{2}} \approx t_v, \tag{6.5}$$

mit der Anzahl Freiheitsgrade

$$v = (d-1)\left(1 + \frac{1}{D+1}\frac{\bar{W}_D}{B_D}\right)^2. \tag{6.6}$$

Mit Hilfe der folgenden Kenngrößen können Aussagen getroffen werden, inwiefern die fehlenden Daten zur Unsicherheit der Schätzungen des Parameters θ beitragen. Aus der Verhältnisgröße,

$$r = \frac{\left(1 + D^{-1}\right)B}{\bar{W}_D} \tag{6.7}$$

kann man den relativen Anstieg der Varianz aufgrund der Anzahl der fehlenden Werte ableiten vgl. Rubin (1987).

Der Anteil der fehlenden Informationen über den Parameter θ wird wie folgt berechnet.

$$\lambda = \frac{r + 2/(v+3)}{r+1} \tag{6.8}$$

Rubin (1987) gibt an, dass für $D \to \infty$ Multiple-Imputation voll effiziente Schätzungen liefert.

Die relative Effizient (RE) eines Schätzer der auf D Imputationen basiert, anstatt auf einer unendlich Anzahl kann wie folgt dargestellt werden

$$RE \approx \left(1 + \frac{\lambda}{m}\right)^{-1}. \tag{6.9}$$

Aus Tabelle 14 erkennt man die relative Effizienz für ausgewählte Kombinationen von λ und D.

	λ				
D	10 %	20 %	30 %	50 %	70 %
3	0.97	0.94	0.91	0.86	0.81
5	0.98	0.96	0.94	0.91	0.88
10	0.99	0.98	0.97	0.95	0.93
20	1	0.99	0.99	0.98	0.97

Tabelle 14: Relative Effizienz der Multiplen-Imputation

Aus Tabelle 14 erkennt man ebenfalls, dass bei 20% fehlenden Informationen bereits $D = 5$ Imputationen ausreichen, um einen zu 96% effizienten Schätzer zu erhalten. Bei einem geringen Anteil fehlender Daten reicht bereits eine Anzahl von drei bis fünf Imputationen aus.

Die Abbildung 16 wurde mit Hilfe der Bibliothek MICE der Software R erstellt. Als Imputationsalgorithmus wurde die Standardeinstellung baysianische stochastische lineare Regression verwendet.

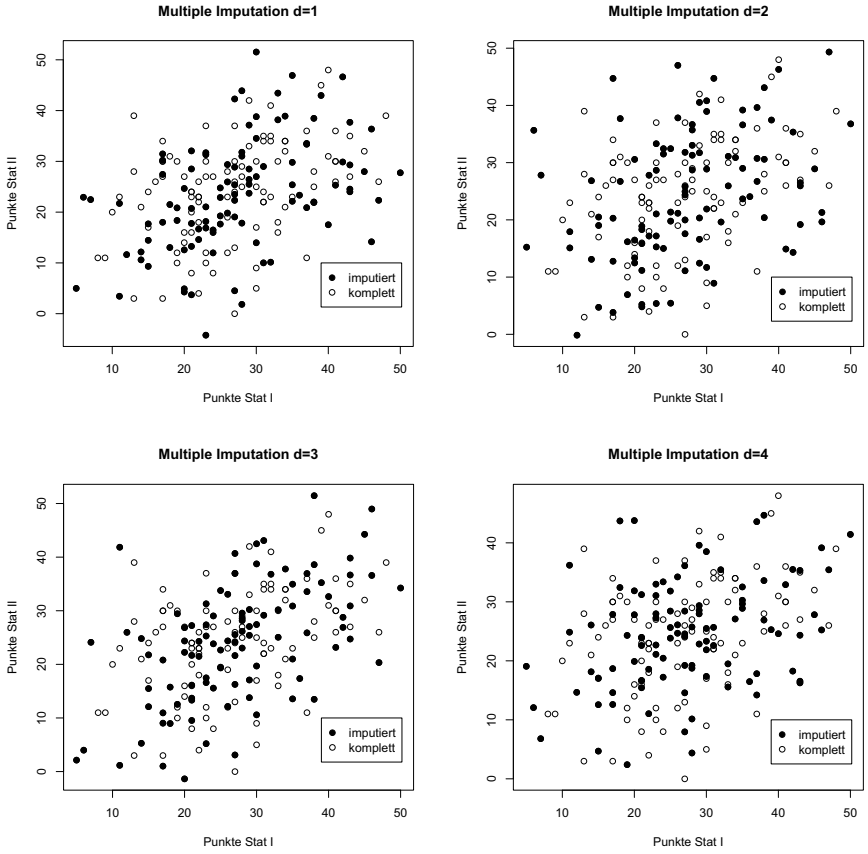

Abbildung 16: Multiple Imputation

7 Simulationsstudie

Im folgenden Teil der Arbeit soll mit Hilfe einer Monte-Carlo (MC) Simulation die Leistungsfähigkeit der beschriebenen Verfahren getestet werden. Eine Monte Carlo Studie ist, einfach formuliert, die wiederholte Durchführung von Zufallsexperimenten zur Lösung verschiedener Fragestellungen. Die Durchführung von MC-Simulationen ist vor allem dann von Vorteil, wenn die Eigenschaften von verschiedenen Schätzern theoretisch nur schwer abgeleitet werden können. Eine MC-Studie gliedert sich im Allgemeinen in vier Schritte: Die Modellierung des Datengenerierungsprozess (DGP), das Erzeugen von B-Daten-Samples, die Analyse der B Daten-Samples mit Hilfe eines geeigneten Verfahrens und die Analyse der Ergebnisse, vgl. Lehmann (2005) mit Bezug auf Kennedy (2004). Im Datengenerierungsprozess wird der unterstellte Zusammenhang zwischen den zu untersuchenden Variablen festgelegt. Bei der Auswertung werden dann die Schätzungen aus der Analyse miteinander verglichen. Man kann zum Beispiel Erwartungswert, Bias oder Standardfehler der Ergebnisse berechnen, vgl. Kennedy (2004). Für die Missing-Data Thematik wurde unter anderem von Little und Rubin (2002), die sich auf Efron (1994) beziehen, ein Ablauf beschrieben. Nach dem Datengenerierungsprozess, in dem Datensätze $S^{(b)}$ mit fehlenden Werten erzeugt wurden, erfolgt das Auffüllen der fehlenden Daten in $S^{(b)}$ durch eine bestimmte Imputationsmethode Imp. Man erhält den imputierten Datensatz $\widehat{S} = Imp(S)$, wo die gewünschten Parameter $\widehat{\theta}$ geschätzt werden. Eine konsistente Schätzung $\widehat{\theta}$ der Parameter θ erhält man durch die B-fache Wiederholung dieser Prozedur. Der Schätzer von θ ist dann der Mittelwert der B Schätzungen der Parameter.

$$\tilde{\theta} = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^b \qquad b = 1, ...B \tag{7.1}$$

7.1 Ablauf der Simulationsstudie

Für die Simulationsstudie werden die Punkte aus der Klausur Statistik 1 (X) und der Klausur Statistik 2 (Y) verwendet. In der Modellierung des Datengenerierungsprozess wird festgelegt, die gewünschten B Datensätze aus den Originaldaten durch Stichprobenverfahren zu gewinnen. Diese sollen mit fehlenden Variablen nach bestimmten Mechanismen (MCAR, MAR, MNAR) ersetzt werden. Jeder der B Datensätze unterscheidet sich von seinem Vorgänger durch fehlende Werte an unterschiedlichen Positionen. Fehlende Werte sollen nur in der Variable Punkte Statistik 2 auftreten. Es sollen für jeden Mechanismus verschiedene Anteile von fehlenden Daten erzeugt werden, dies sind 5%,10%,20%,50%,80%. Für Mechanismus fehlender Daten und jeden Anteil werden 100 Simulationen gestartet. Es werden so für jedes Verfahren insgesamt 1500 Simulation durchgeführt (Anm.: Die Datensätze werden auf der Festplatte gespeichert, jedes Verfahren benutzt so die gleichen 1500 Datensätze). Die Parameter, mit denen die Qualität der Verfahren geprüft wird, sind der Mittelwert und die Standardabweichung

der Punkte Statistik 2, sowie der Korrelationskoeffizient zwischen den Variablen Punkte Statistik 1 und Punkte Statistik 2 und die Regressionskoeffizienten $\beta_{X|Y}$ sowie $\beta_{Y|X}$, vgl. Schafer und Graham (2003).

Es werden folgende Verfahren Imp untersucht (Anm.: Listenweise Fallausschluss ist kein Imputationsverfahren):

- Listenweise Fallausschluss
- Mittelwertersetzung
- Regressionsimputation
- Stochastische Regressionsimputation
- Einfaches Hot-Deck
- Sequentielles (k-NN) Hot-Deck
- EM-Algorithmus in SPSS
- EM-Algorithmus in SPLUS/R
- PROC MI in SAS 9.1

7.1.1 Erzeugung der Datensätze unter MCAR

Aus den Daten werden durch Zufallsziehungen Beobachtungen ausgewählt. Diese Beobachtungen werden in der Variable Punkte Statistik 2 als fehlend kodiert.

7.1.2 Erzeugung der Datensätze unter MAR

Studierende, die eine geringe Punktzahl bei der Statistik 1 Klausur erreicht haben, werden in der Variable Punkte Statistik 2 als fehlend kodiert. Dies geschieht durch Zufallsziehungen aus den unteren Quantilen der Variable Punkte Statistik 1 (Anm.: Es macht inhaltlich keinen Unterschied, in diese Prozedur anstatt der unteren die oberen Quantile zu verwenden). Die Stichprobe für den 5% Anteil fehlender Daten setzt sich nur aus Beobachtungen zusammen, die bei den Statistik 1 Punkten aus dem 10% Quantil entnommen wurden. Für den 20% Anteil wurden nur Beobachtungen aus dem des 25% Quantil verwendet usw. Auf diese Weise kann auch ein hoher Anteil von 80% fehlenden Daten erreicht werden, der immer noch dem MAR Mechanismus folgt. Ein Problem dieses Ansatzes ist jedoch, dass sich die Datensätze mit steigendem Anteil an fehlenden Daten immer mehr ähneln. Eine hohe stochastische Komponente ist aber insbesondere für die Datensätze mit geringem Anteil an fehlenden Daten vorhanden.

7.1.3 Erzeugung der Datensätze unter MNAR

Der Ablauf unter MNAR ist ähnlich des Ablauf unter MAR. Bei MNAR werden jedoch die fehlenden Werte aus den Quantilen der Variable Punkte Statistik 2 selbst gezogen.

7.2 Listenweiser Fallausschluss

Bei Listenweisen Fallausschluss erkennt man in Tabelle 15 und Abbildung 44 die gute Qualität der Schätzungen im Falle von MCAR. Die Probleme der Methode werden jedoch unter MAR sichtbar (Anm.: Die Parameter des Originaldatensatzes befinden sich in der linken Spalte). Hier wird vor allem der Mittelwert überschätzt. Dies liegt daran, dass vermehrt Beobachtungen von Studenten mit schwachen Leistungen in Statistik 1 in Statistik 2 als fehlend kodiert wurden. Dadurch verbleiben überdurchschnittlich viele gute Studenten in der Stichprobe und diese ist nicht mehr repräsentativ. Die Korrelation und der Regressionskoeffizient $\beta_{X|Y}$ werden unterschätzt. Für den Fall von MNAR kommt es bereits bei einem sehr geringen Anteil von fehlenden Daten zu Misspezifikationen.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,36	24,37	24,36	24,42	24,34
$\sigma_Y = 9.85$	9,85	9,83	9,88	9,81	9,82
$\rho = 0.45$	0,44	0,44	0,44	0,43	0,45
$\beta_{X\|Y} = 0.42$	0,42	0,42	0,42	0,41	0,43
$\beta_{Y\|X} = 0.46$	0,46	0,46	0,47	0,45	0,47
MAR					
$\mu_Y = 24.35$	24,73	25,02	25,28	27,99	30,84
$\sigma_Y = 9.85$	9,75	9,77	9,88	9,49	9,59
$\rho = 0.45$	0,42	0,41	0,45	0,39	0,44
$\beta_{X\|Y} = 0.42$	0,38	0,35	0,36	0,3	0,39
$\beta_{Y\|X} = 0.46$	0,47	0,49	0,56	0,52	0,5
MNAR					
$\mu_Y = 24.35$	25,32	26,32	27,63	31,63	37,5
$\sigma_Y = 9.85$	9,12	8,3	7,9	6,77	4,77
$\rho = 0.45$	0,43	0,42	0,44	0,37	0,35
$\beta_{X\|Y} = 0.42$	0,44	0,47	0,52	0,53	0,67
$\beta_{Y\|X} = 0.46$	0,41	0,38	0,37	0,26	0,19

Tabelle 15: Listenweiser Fallausschluss

Die Statistikpunkte folgen, wie bereits in Kapitel 2 beschrieben, einer Normalverteilung. Sollten die Daten jedoch eine schiefe Verteilung haben und bestimmte Randgruppe die Antworten verweigert, so wird das Verfahren wahrscheinlich noch größere Probleme bekommen. Die Ergebnisse stimmen sehr gut mit der Literatur wie Allison (2002), Graham und Schafer (2003) und Schafer (1997) überein, die den Einsatz des Fallweisen Ausschluss außer bei MCAR ablehnen.

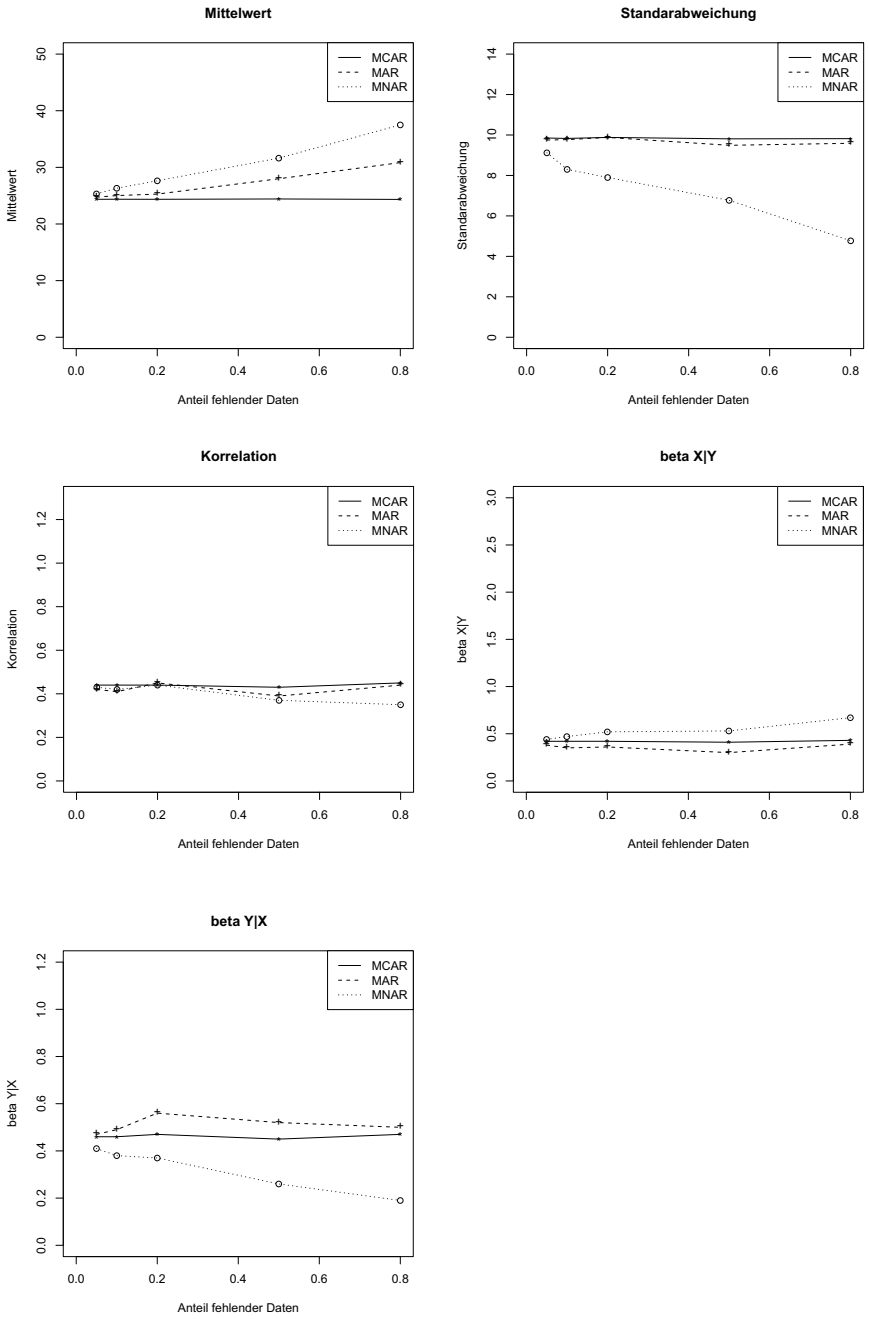

Abbildung 17: Listenweiser Fallausschluss

7.3 Mittelwertersetzung

Bei der Mittelwertersetzung erkennt man die Schwierigkeiten des Verfahrens, dargestellt in Tabelle 16 und Abbildung 18. Bereits bei einem geringeren Anteil an fehlenden Werten sind Parameter wie Standardabweichung, Korrelation und der Koeffizient der Regression von Y auf X $\beta_{Y|X}$ verzerrt. Die Ursache ist die wiederholende Imputation eines und des selben Wertes in die Verteilung von Y. Bei einem sehr geringen Anteil fehlender Werte sind die Ergebnisse unter MCAR noch annehmbar. Im Fall von MNAR kommt es, wie beim Fallweisen Ausschluss, bereits bei einem sehr geringen Anteil von fehlenden Daten zu Misspezifikationen.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,36	24,37	24,36	24,42	24,34
$\sigma_Y = 9.85$	9,59	9,31	8,83	6,9	4,34
$\rho = 0.45$	0,43	0,42	0,39	0,3	0,2
$\beta_{X\|Y} = 0.42$	0,42	0,42	0,42	0,41	0,43
$\beta_{Y\|X} = 0.46$	0,44	0,41	0,37	0,22	0,09
MAR					
$\mu_Y = 24.35$	24,73	25,02	25,28	27,99	30,84
$\sigma_Y = 9.85$	9,5	9,26	8,83	6,68	4,24
$\rho = 0.45$	0,38	0,34	0,34	0,21	0,18
$\beta_{X\|Y} = 0.42$	0,38	0,35	0,36	0,3	0,39
$\beta_{Y\|X} = 0.46$	0,39	0,34	0,32	0,15	0,08
MNAR					
$\mu_Y = 24.35$	25,32	26,32	27,63	31,63	37,5
$\sigma_Y = 9.85$	8,88	7,86	7,06	4,76	2,11
$\rho = 0.45$	0,42	0,4	0,39	0,27	0,15
$\beta_{X\|Y} = 0.42$	0,44	0,47	0,52	0,53	0,67
$\beta_{Y\|X} = 0.46$	0,39	0,33	0,3	0,14	0,03

Tabelle 16: Mittelwertersetzung

Aufgrund dieser Ergebnisse sollte man auf Mittelwertersetzung verzichten. Auch in der Literatur, wie Allison (2002), Graham und Schafer (2003) und Schafer (1997), wird das Verfahren abgelehnt. In der Praxis wird es jedoch noch sehr häufig verwendet.

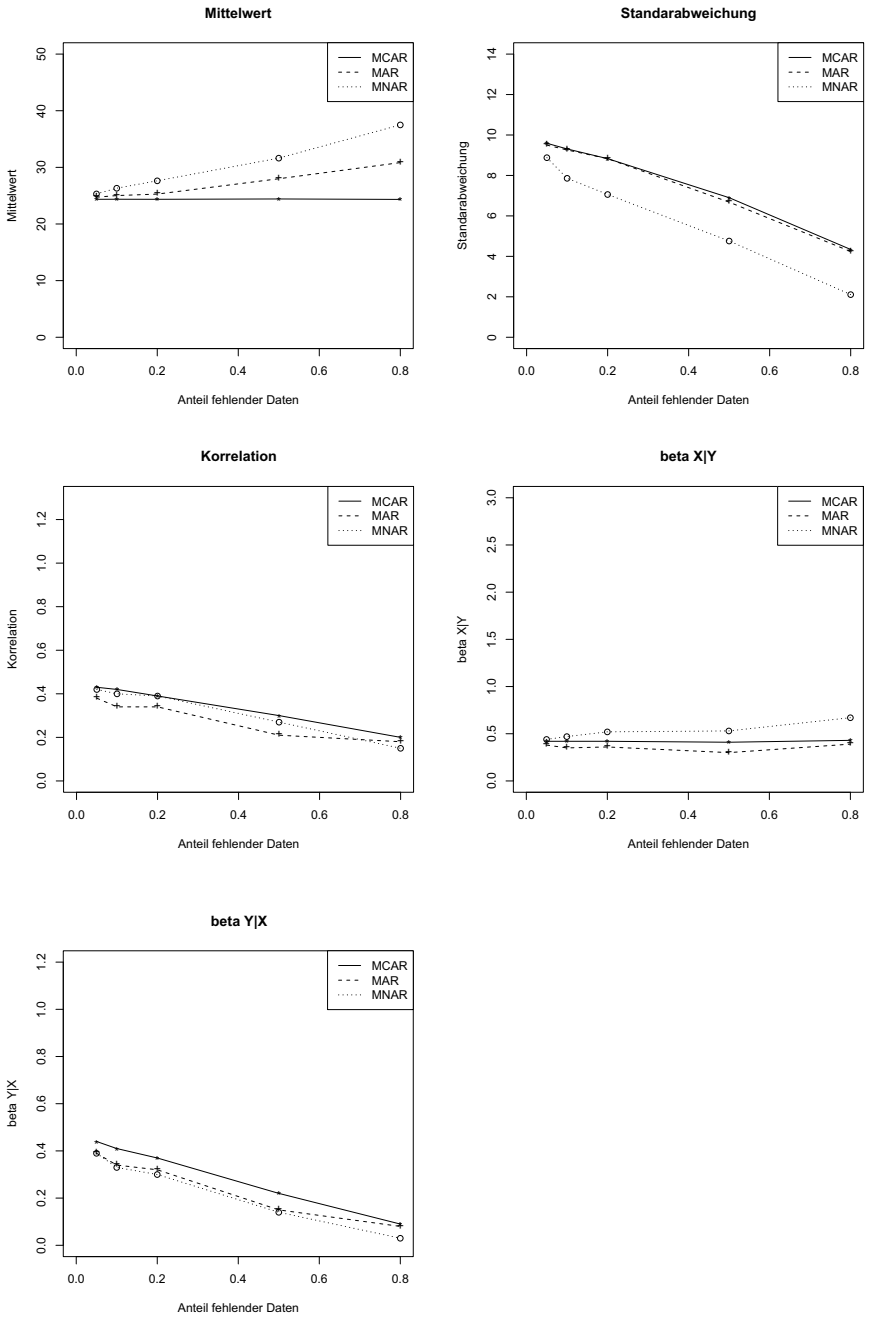

Abbildung 18: Mittelwertersetzung

7.4 Regressionsimputation

In Tabelle 17 und Abbildung 19 sind die Ergebnisse der Regressionsimputation aufgeführt. Die Mittelwerte werden unter MCAR und MAR zuverlässig geschätzt. Die anderen Parameter, insbesondere der Korrelationskoeffzient und der Koeffizient der Regression von X auf Y $\beta_{X|Y}$ sind ab ca. 20% fehlender Daten verzerrt.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,36	24,37	24,34	24,43	24,4
$\sigma_Y = 9.85$	9,64	9,41	9,05	7,53	5,91
$\rho = 0.45$	0,45	0,46	0,48	0,56	0,73
$\beta_{X\|Y} = 0.42$	0,44	0,46	0,5	0,7	1,17
$\beta_{Y\|X} = 0.46$	0,46	0,46	0,47	0,45	0,47
MAR					
$\mu_Y = 24.35$	24,32	24,18	23,63	24,39	25,17
$\sigma_Y = 9.85$	9,67	9,6	9,48	7,84	6,08
$\rho = 0.45$	0,45	0,48	0,55	0,62	0,75
$\beta_{X\|Y} = 0.42$	0,44	0,47	0,55	0,74	1,16
$\beta_{Y\|X} = 0.46$	0,47	0,49	0,56	0,52	0,5
MNAR					
$\mu_Y = 24.35$	25,19	26,07	27,26	30,78	36,3
$\sigma_Y = 9.85$	8,92	7,95	7,23	5,04	2,66
$\rho = 0.45$	0,43	0,44	0,48	0,48	0,65
$\beta_{X\|Y} = 0.42$	0,46	0,52	0,63	0,9	2,35
$\beta_{Y\|X} = 0.46$	0,41	0,38	0,37	0,26	0,19

Tabelle 17: Regressionsimputation

Die Regressionsimputation führt zu umso besseren Ergebnissen, falls eine Variable mit fehlenden Werten stark mit einer Variable mit wenigen fehlenden Werten korreliert ist, vgl. Allison (2001).

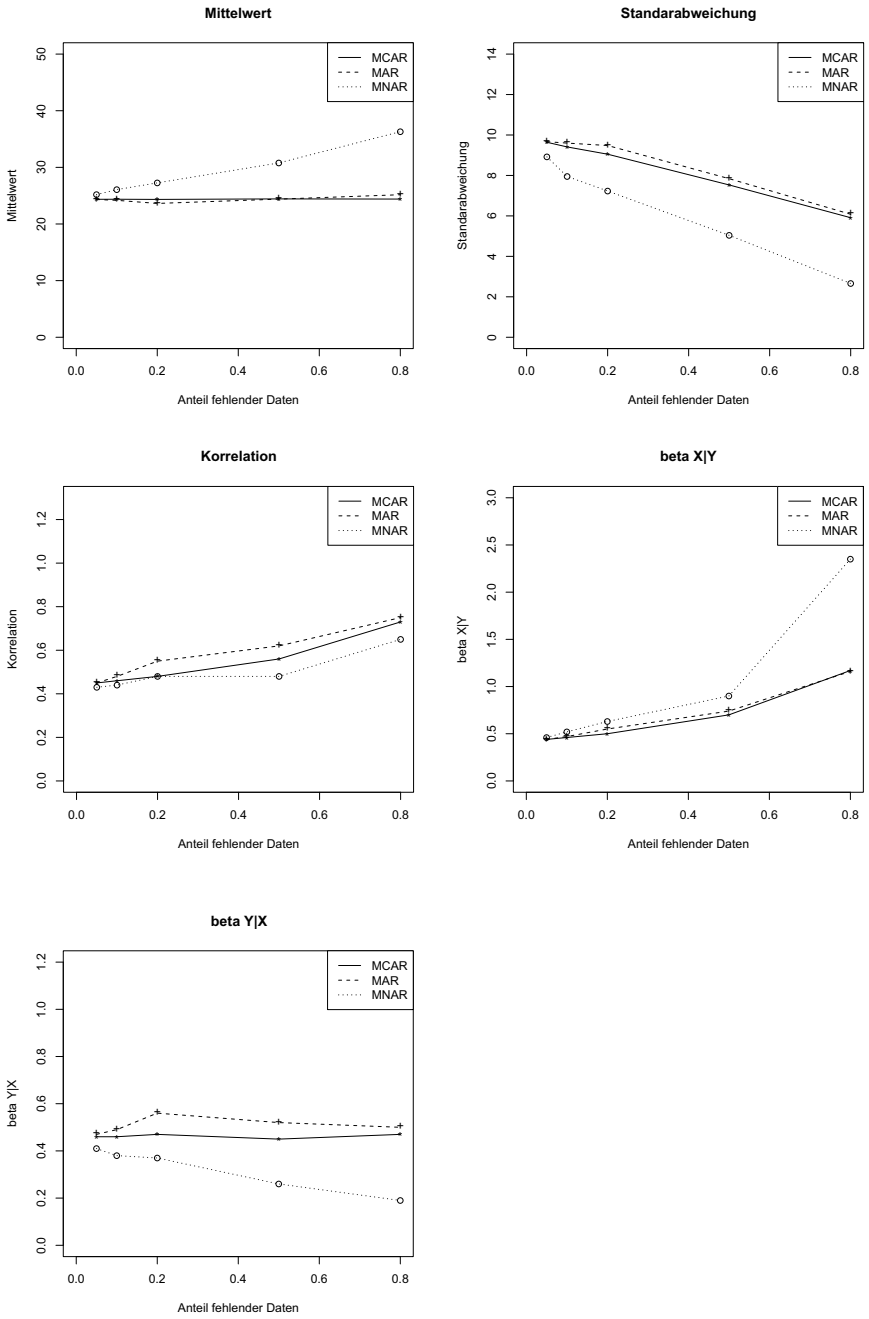

Abbildung 19: Regressionsimputation

7.5 Stochastische Regressionsimputation

Für die Simulation wurde ein R-Programm nach den Vorgaben von Allison (2001) geschrieben. Es wurde insbesondere darauf geachtet, die Fehlerterme aus einer neu generierten Zufallsvariablen z mit $\mu = 0$ und $\sigma_z = \sigma_{res}$ zu ziehen. Die Ergebnisse sind in Tabelle 18 und Abbildung 20 zusammengefasst. Im Fall von MCAR und MAR liegen die geschätzten Parameter sehr nah an den ursprünglichen Parametern.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,38	24,36	24,31	24,35	24,33
$\sigma_Y = 9.85$	9,87	9,83	9,91	9,8	9,74
$\rho = 0.45$	0,44	0,44	0,45	0,44	0,45
$\beta_{X\|Y} = 0.42$	0,42	0,42	0,42	0,42	0,43
$\beta_{Y\|X} = 0.46$	0,47	0,46	0,47	0,45	0,46
MAR					
$\mu_Y = 24.35$	24,31	24,19	23,63	24,33	25,23
$\sigma_Y = 9.85$	9,86	9,96	10,28	10,03	9,77
$\rho = 0.45$	0,44	0,46	0,51	0,49	0,47
$\beta_{X\|Y} = 0.42$	0,42	0,43	0,47	0,45	0,45
$\beta_{Y\|X} = 0.46$	0,47	0,48	0,56	0,52	0,5
MNAR					
$\mu_Y = 24.35$	25,2	26,09	27,26	30,78	36,32
$\sigma_Y = 9.85$	9,12	8,31	7,86	6,71	4,78
$\rho = 0.45$	0,42	0,42	0,44	0,35	0,37
$\beta_{X\|Y} = 0.42$	0,43	0,48	0,53	0,5	0,73
$\beta_{Y\|X} = 0.46$	0,41	0,37	0,37	0,25	0,19

Tabelle 18: Stochastische Regressionsimputation

Auffällig sind insbesondere die sehr effizienten Schätzungen auch bei sehr hohem Anteil fehlender Daten. Das Verfahren wird unter anderem von Allison (2001) empfohlen.

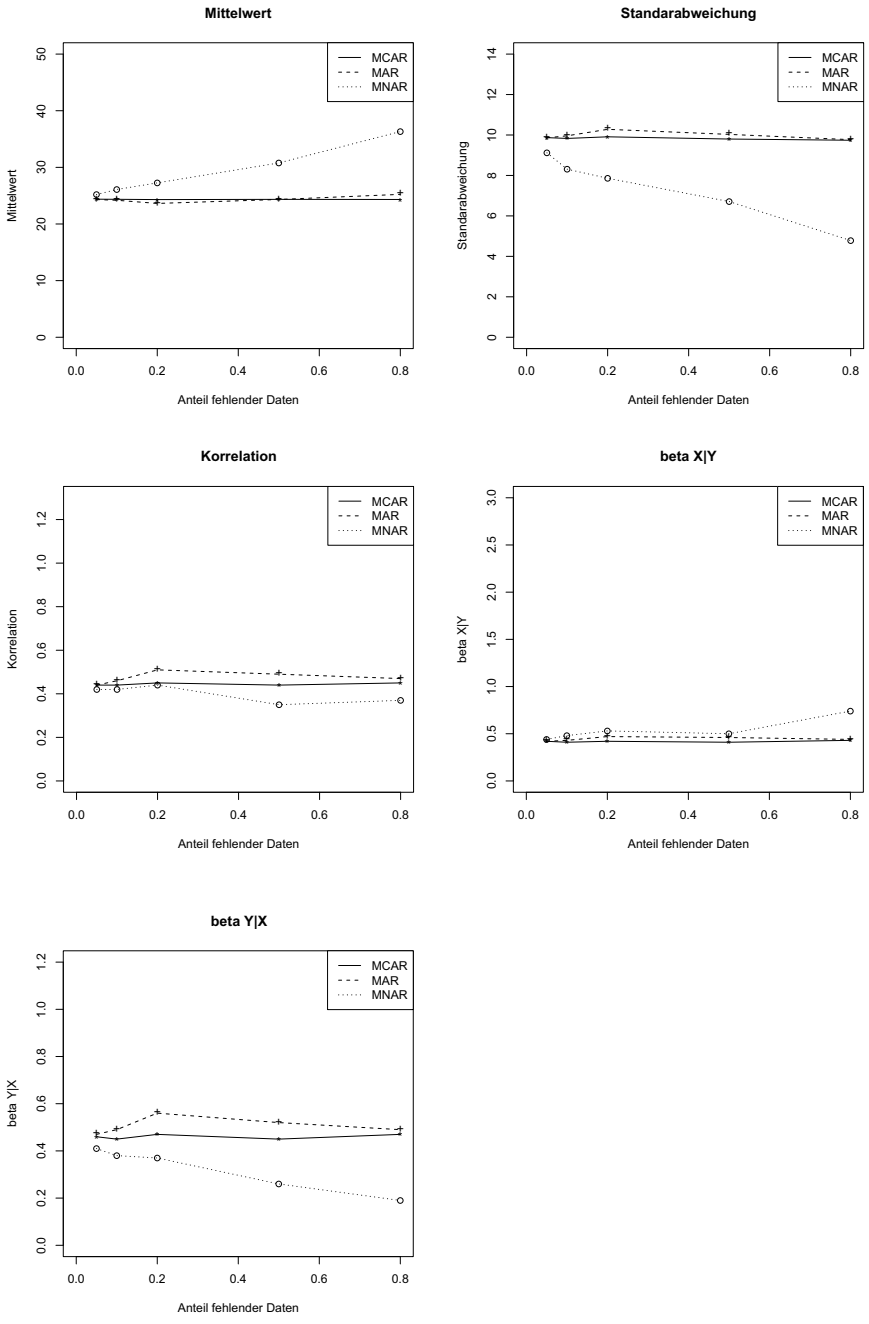

Abbildung 20: Stochastische Regressionsimputation

7.6 Einfaches Hot-Deck

Die Ergebnisse der Hot-Deck Imputation sind in Tabelle 19 und Abbildung 21 zu erkennen. Parameter wie Mittelwert und Standardabweichung werden bei geringem Anteil fehlender Daten zwar richtig spezifiziert, jeoch erkennt man eine deutliche Verzerrung von Korrelationkoeffizienten und den Regressionsparametern. Zufallsziehungen aus der Marginalverteilung von Y sind problematisch, da Informationen die in anderen Variablen vorhanden sind nicht ausgenutzt werden.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,35	24,38	24,39	24,37	24,38
$\sigma_Y = 9.85$	9,85	9,8	9,85	9,72	9,73
$\rho = 0.45$	0,42	0,4	0,35	0,22	0,11
$\beta_{X\|Y} = 0.42$	0,4	0,38	0,34	0,21	0,1
$\beta_{Y\|X} = 0.46$	0,44	0,41	0,37	0,22	0,11
MAR					
$\mu_Y = 24.35$	24,76	25,03	25,25	28,01	30,78
$\sigma_Y = 9.85$	9,74	9,77	9,88	9,52	9,47
$\rho = 0.45$	0,37	0,32	0,3	0,15	0,08
$\beta_{X\|Y} = 0.42$	0,35	0,31	0,29	0,15	0,08
$\beta_{Y\|X} = 0.46$	0,38	0,34	0,32	0,15	0,08
MNAR					
$\mu_Y = 24.35$	25,34	26,29	27,64	31,62	37,51
$\sigma_Y = 9.85$	9,11	8,29	7,91	6,72	4,69
$\rho = 0.45$	0,4	0,37	0,35	0,18	0,08
$\beta_{X\|Y} = 0.42$	0,42	0,42	0,42	0,26	0,15
$\beta_{Y\|X} = 0.46$	0,39	0,33	0,3	0,13	0,04

Tabelle 19: Hot-Deck

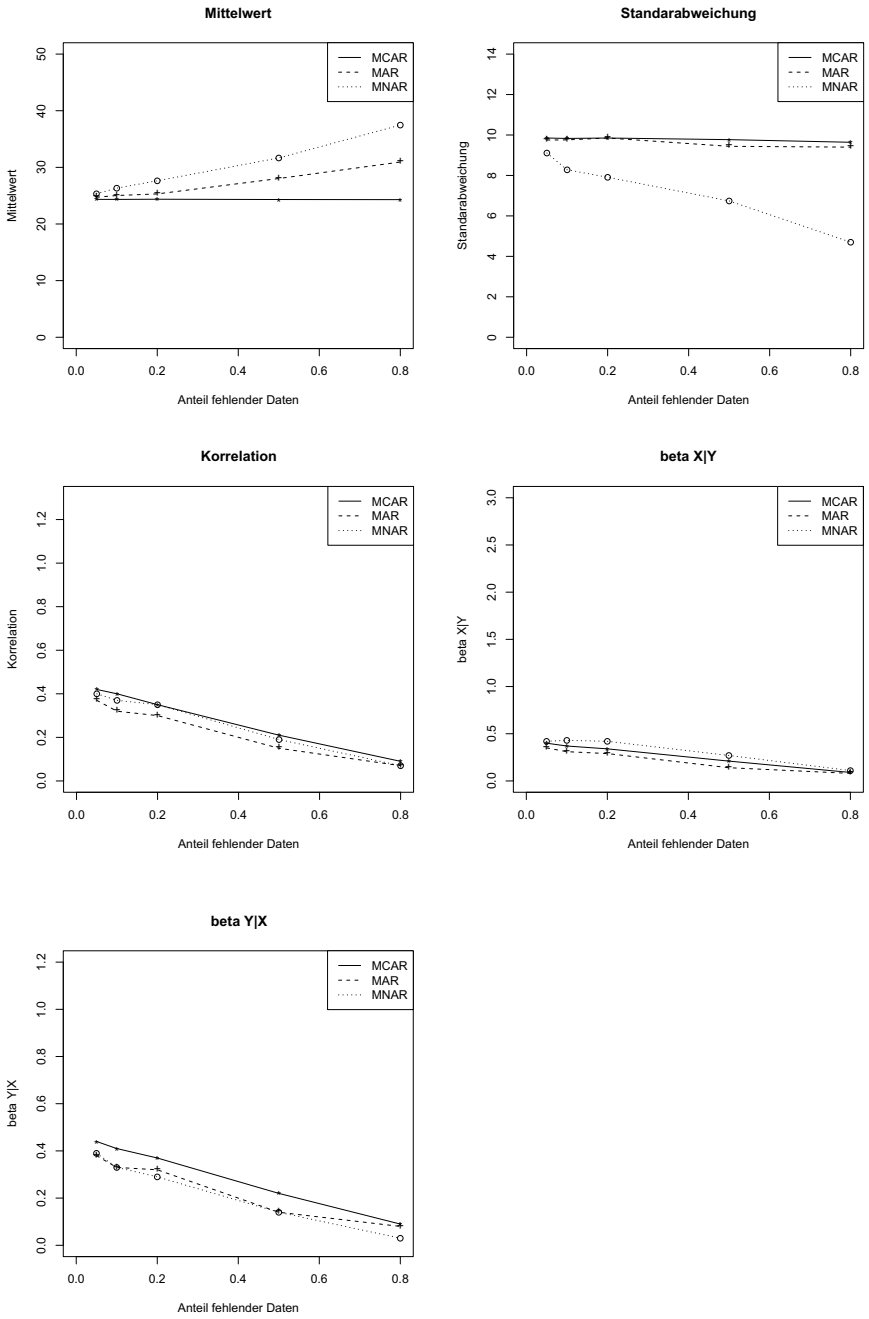

Abbildung 21: Hot-Deck

7.7 Sequentielles Hot-Deck

Die Ergebnisse des Sequentiellen Hot-Deck Verfahrens sind in Tabelle 20 und Abbildung 22 zusammengefasst. Das Verfahren aus der R-Bibliothek SeqKnn erzielt dabei sehr gute Simulationsergebnisse. Die Rechenzeit, die das Verfahren durch die sequentielle Abarbeitung der Beobachtungen benötigte, war dabei äußerst gering. Bei bis zu 20% fehlenden Werten werden alle Parameter zuverlässig geschätzt. Bei höherem Anteil wird insbesondere die Standardabweichung unterschätzt.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,31	24,32	24,22	24,07	23,52
$\sigma_Y = 9.85$	9,77	9,65	9,5	8,8	7,76
$\rho = 0.45$	0,44	0,44	0,44	0,41	0,34
$\beta_{X\|Y} = 0.42$	0,42	0,43	0,43	0,44	0,41
$\beta_{Y\|X} = 0.46$	0,46	0,45	0,44	0,39	0,28
MAR					
$\mu_Y = 24.35$	24,44	24,52	24,39	25,45	25,74
$\sigma_Y = 9.85$	9,75	9,69	9,39	8,43	7,37
$\rho = 0.45$	0,42	0,41	0,45	0,41	0,43
$\beta_{X\|Y} = 0.42$	0,41	0,4	0,45	0,46	0,57
$\beta_{Y\|X} = 0.46$	0,44	0,42	0,45	0,38	0,34
MNAR					
$\mu_Y = 24.35$	25,17	26,01	27,16	30,77	36,28
$\sigma_Y = 9.85$	8,97	8,06	7,44	5,98	3,71
$\rho = 0.45$	0,42	0,41	0,42	0,34	0,19
$\beta_{X\|Y} = 0.42$	0,44	0,48	0,52	0,53	0,5
$\beta_{Y\|X} = 0.46$	0,4	0,35	0,33	0,21	0,08

Tabelle 20: Sequentielles Hot-Deck

Durch die Einbeziehung von „ähnlichen“ Beobachtungen bleibt die Struktur des Datensatzes im Gegensatz zum einfachen Hot-Deck besser erhalten. Für einen geringen Anteil an fehlenden Daten sind die Ergebnisse somit sehr sinnvoll.

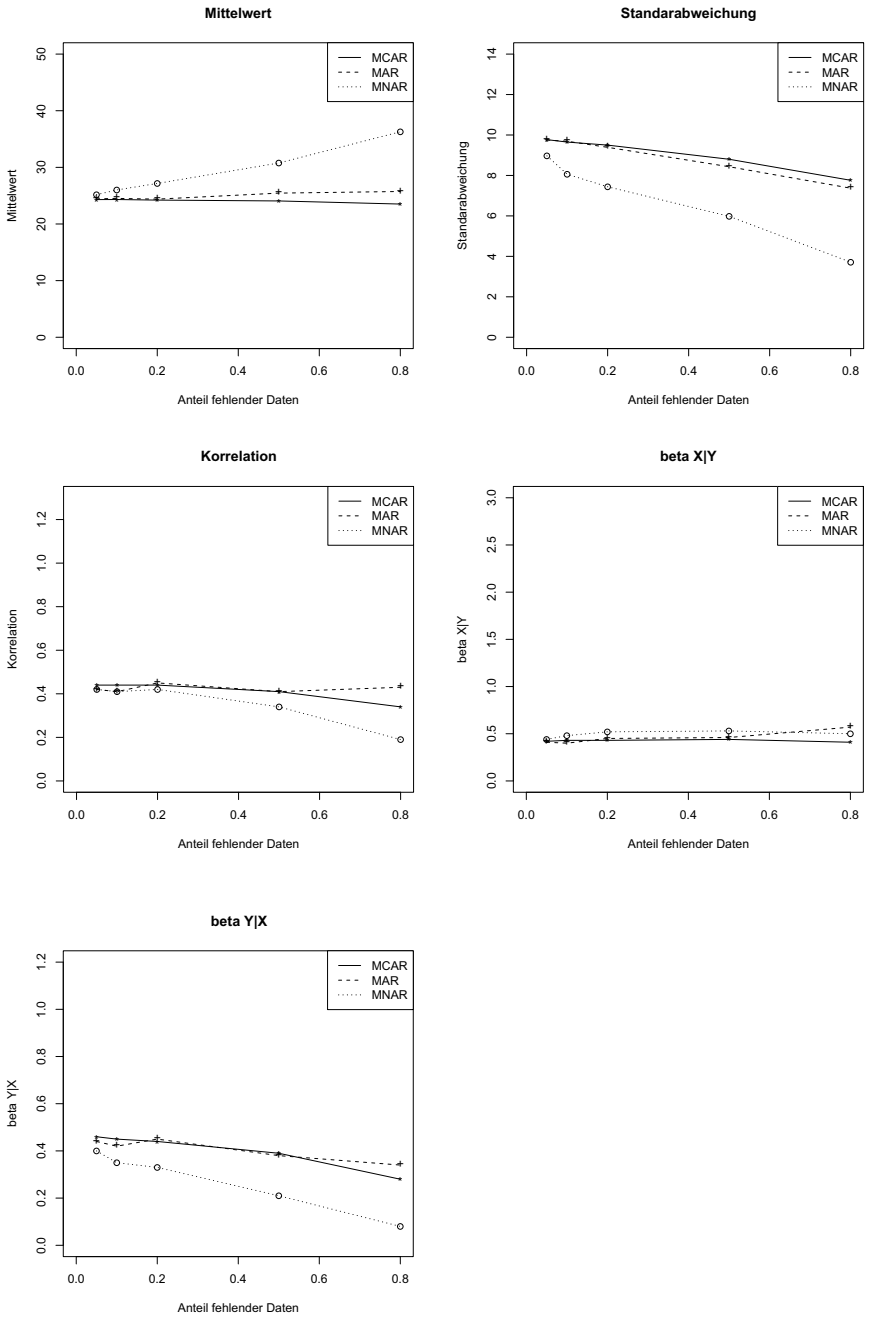

Abbildung 22: Sequentielles Hot-Deck

7.8 SPSS Modul Missing-Value-Analysis (MVA)

Für die Simulation wurde die SPSS Version 13 (Syntax) verwendet. Wenn man die Ergebnisse in Tabelle 21 und Abbildung 23 betrachtet fällt auf, dass die Resultate fast exakt mit denen der Regressionsimputation übereinstimmen. Die Standardfehler und Teststatistiken sind stark verzerrt, vgl. v. Hippel (2004).

In SPSS-MVA sind folgende Methoden verfügbar.

- Listenweise Fallausschluss
- Regressionsersetzung
- EM-Algorithmus

Man erkennt aber, dass der EM Algorithmus in SPSS-MVA die fehlenden Werte ohne Ziehung aus der Verteilung der Residuen imputiert. Verbesserte Verfahren, wie Data-Augmentation oder Multiple-Imputation, sind in SPSS noch nicht verfügbar.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,36	24,37	24,34	24,43	24,41
$\sigma_Y = 9.85$	9,64	9,41	9,05	7,53	5,9
$\rho = 0.45$	0,45	0,46	0,48	0,56	0,73
$\beta_{X\|Y} = 0.42$	0,44	0,46	0,5	0,7	1,17
$\beta_{Y\|X} = 0.46$	0,46	0,46	0,47	0,45	0,47
MAR					
$\mu_Y = 24.35$	24,32	24,18	23,63	24,43	26,34
$\sigma_Y = 9.85$	9,67	9,6	9,48	7,81	5,55
$\rho = 0.45$	0,45	0,48	0,55	0,61	0,69
$\beta_{X\|Y} = 0.42$	0,44	0,47	0,55	0,74	1,17
$\beta_{Y\|X} = 0.46$	0,47	0,49	0,56	0,51	0,42
MNAR					
$\mu_Y = 24.35$	25,19	26,07	27,26	30,79	36,4
$\sigma_Y = 9.85$	8,92	7,95	7,23	5,04	2,61
$\rho = 0.45$	0,43	0,44	0,48	0,48	0,64
$\beta_{X\|Y} = 0.42$	0,46	0,52	0,63	0,89	2,33
$\beta_{Y\|X} = 0.46$	0,41	0,38	0,37	0,26	0,18

Tabelle 21: SPSS MVA

In der Literatur, vgl. v. Hippel (2004), wird von der Benutzung von SPSS-MVA abgeraten.

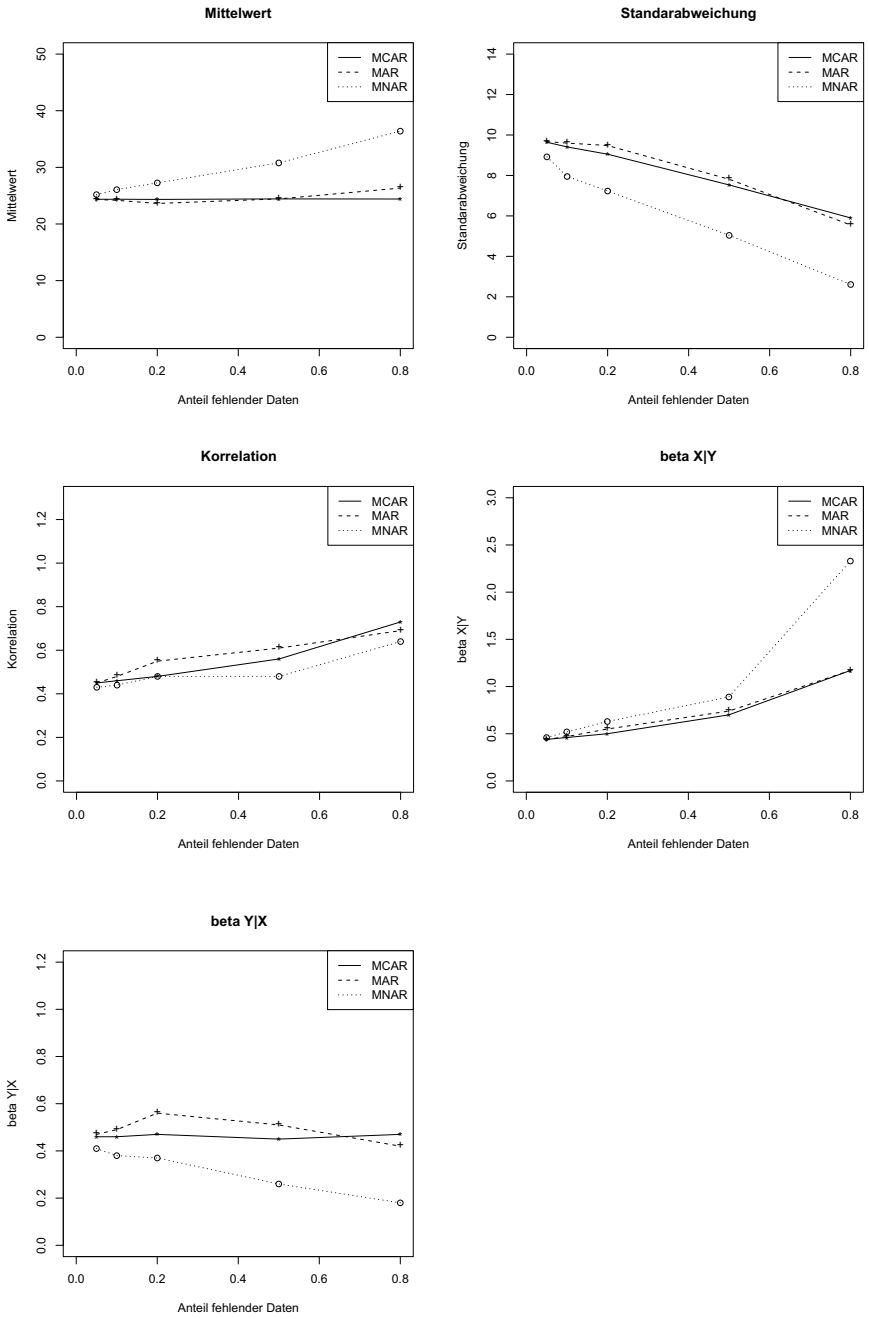

Abbildung 23: SPSS MVA

7.9 EM-Algorithmus in SPLUS 6.0 und R Version 2.5

Voraussetzungen für die Verwendung ist die multivarate Normalveteilung. Das Verfahren kann für alle Muster von fehlenden Daten benutzt werden. In Tabelle 22 und Abbildung 24 sind die Ergebnisse des EM-Algorithmus aus der R-Bibliothek Norm dargestellt. In der Bibliothek sind folgende Verfahren verfügbar.

- EM-Algorithmus (mit Zufallsziehung aus den Residuen)
- Data-Augmentation (mit Inverse-Wishart prior oder alternativ einer gleichverteilten a-priori Verteilung)
- Data-Augmentation (mit nicht-informativer a-priori Verteilungen und MCMC Methoden)

Bei der Data Augmentation mit nicht-informativer a-priori Verteilung werden nur soviel Werte imputiert, bis ein monotones Muster der fehlenden Daten entsteht, vgl. Schafer (1997). Auch bei einem hohen Anteil fehlender Daten sind Parameter wie Standardabweichung, Korrelation und der Regressionskoeffizienten unter einem ignorierbarem Fehlendmechanismus (MCAR, MAR) unverzerrt.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24,46	24,56	24,47	24,7	24,67
$\sigma_Y = 9.85$	9,99	9,91	10,03	9,76	9,69
$\rho = 0.45$	0,44	0,44	0,44	0,42	0,45
$\beta_{X\|Y} = 0.42$	0,41	0,41	0,41	0,4	0,44
$\beta_{Y\|X} = 0.46$	0,47	0,46	0,47	0,44	0,47
MAR					
$\mu_Y = 24.35$	24,41	24,37	23,76	24,67	25,45
$\sigma_Y = 9.85$	9,91	9,96	10,35	9,81	9,64
$\rho = 0.45$	0,42	0,43	0,5	0,45	0,46
$\beta_{X\|Y} = 0.42$	0,4	0,41	0,45	0,43	0,45
$\beta_{Y\|X} = 0.46$	0,44	0,46	0,55	0,48	0,49
MNAR					
$\mu_Y = 24.35$	25,28	26,23	27,36	30,97	36,44
$\sigma_Y = 9.85$	9,23	8,3	7,93	6,74	4,68
$\rho = 0.45$	0,42	0,41	0,42	0,36	0,33
$\beta_{X\|Y} = 0.42$	0,42	0,46	0,5	0,5	0,68
$\beta_{Y\|X} = 0.46$	0,41	0,36	0,35	0,26	0,17

Tabelle 22: EM Splus/R

Für einen nichtignorierbaren Mechanismus (MNAR) sind die Ergebnisse aber auch hier stark verzerrt. Für die Simulation wurde Data-Augmentation mit einer gleichverteilten prior Verteilung verwendet, vgl. Schafer (1997).

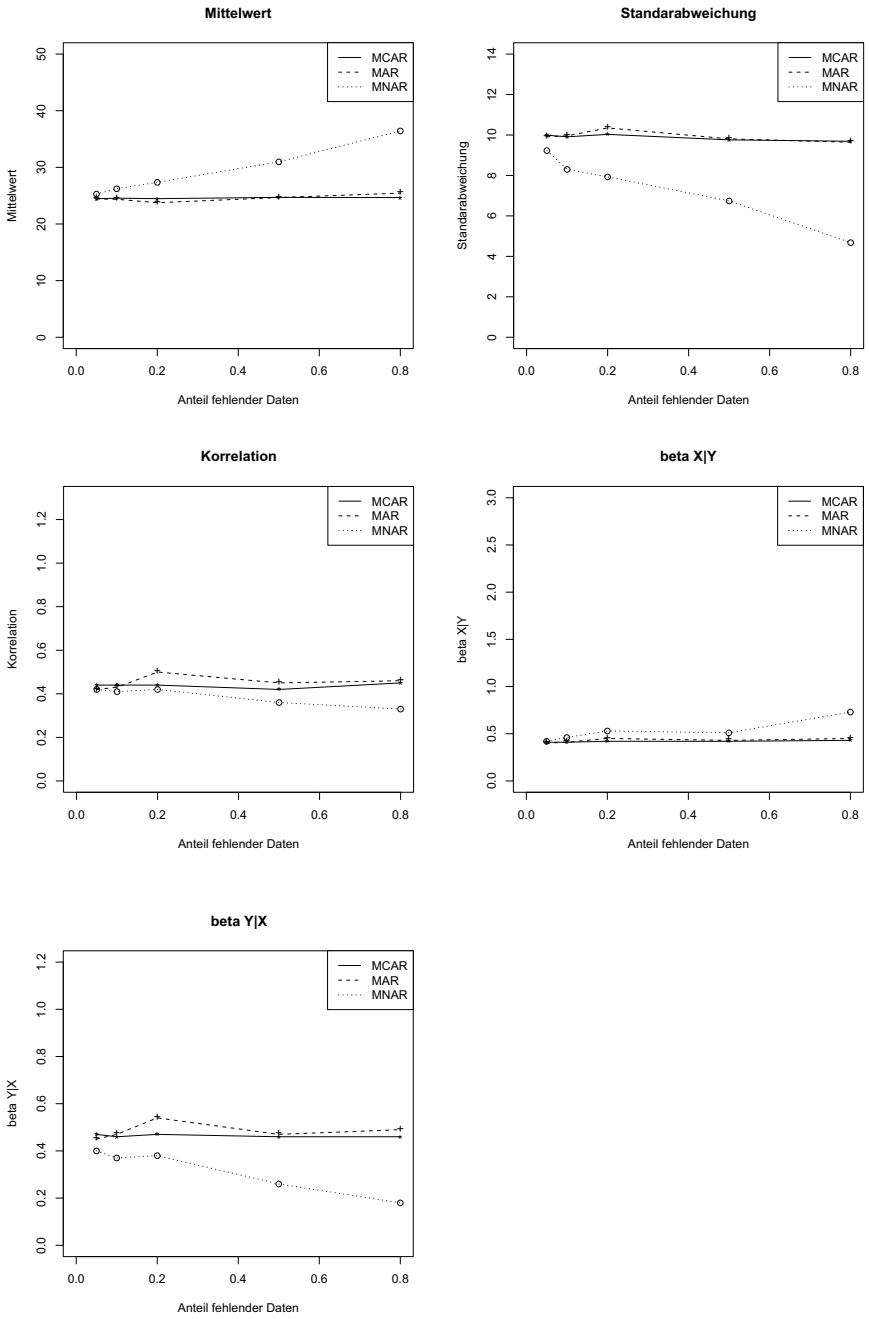

Abbildung 24: EM Splus/R

7.10 PROC MI in SAS

In Tabelle 23 und Abbildung 25 sind die Ergebnisse der SAS Proc MI dargestellt. Fehlende Werte sind in allen Variablen erlaubt. Es stehen 3 Verfahren zur Auswahl. Ein Regressionsverfahren, eine nichtparametrisches Propensity-Score Methode, vgl. Little und Rubin (2002), und eine Markov-Chain-Monte-Carlo Methode. Die Voraussetzung für die Anwendung der Regressionsmethode under der MCMC Methode ist eine multivariate Normalverteilung in den Daten. Bei der Data-Augmentation mit Hilfe des MCMC Algorithmus sollte der Anteil der fehlenden Daten laut Schafer (1997) nicht zu groß sein. Eine genaue Spezifikation findet sich jedoch nicht. Für einen geringen Anteil fehlender Daten füllt der MCMC Algorithmus Daten bis zu einem monotonen Muster auf. Es besteht auch die Möglichkeit, den ganzen Datensatz mit Hilfe von MCMC Methoden aufzufüllen. Ist ein monotones Muster erreicht, (I-Schritt) kann auch mit einem stochastischen Regressionsverfahren fortgefahren (P-Schritt) werden, vgl. Schafer(1997). In der Proc MIANALYZE besteht später die Möglichkeit, die imputierten Datensätze zu vergleichen. Für die Simulation wurde die Data-Augmentation mittels MCMC mit 200 Iterationen gewählt.

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
$\mu_Y = 24.35$	24.15	24.39	23.98	23.83	24.55
$\sigma_Y = 9.85$	9.78	9.73	9.93	9.79	9.9
$\rho = 0.45$	0.45	0.44	0.44	0.42	0.58
$\beta_{X\|Y} = 0.42$	0.43	0.43	0.42	0.40	0.55
$\beta_{Y\|X} = 0.46$	0.47	0.46	0.47	0.44	0.62
MAR					
$\mu_Y = 24.35$	24,11	24,2	23,44	23,84	25,21
$\sigma_Y = 9.85$	9,91	9,9	10,35	10,18	9,13
$\rho = 0.45$	0,48	0,46	0,53	0,51	0,51
$\beta_{X\|Y} = 0.42$	0,45	0,43	0,48	0,47	0,52
$\beta_{Y\|X} = 0.46$	0,5	0,48	0,58	0,55	0,5
MNAR					
$\mu_Y = 24.35$	24,97	26,08	26,92	30,3	36
$\sigma_Y = 9.85$	9,12	8,12	8,13	6,8	4,7
$\rho = 0.45$	0,44	0,41	0,48	0,37	0,48
$\beta_{X\|Y} = 0.42$	0,45	0,47	0,55	0,51	0,98
$\beta_{Y\|X} = 0.46$	0,43	0,35	0,41	0,27	0,24

Tabelle 23: SAS Proc MI

Unter MAR liefert das Verfahren bereits ab 10 % fehlenden Daten ungenaue Schätzungen. Wichtige Parameter wie der Korrelationskoeffizient und die Regressionsparameter sind verzerrt. Ersetzung von fehlenden Daten ausschließlich mit MCMC sollte daher vermieden werden.

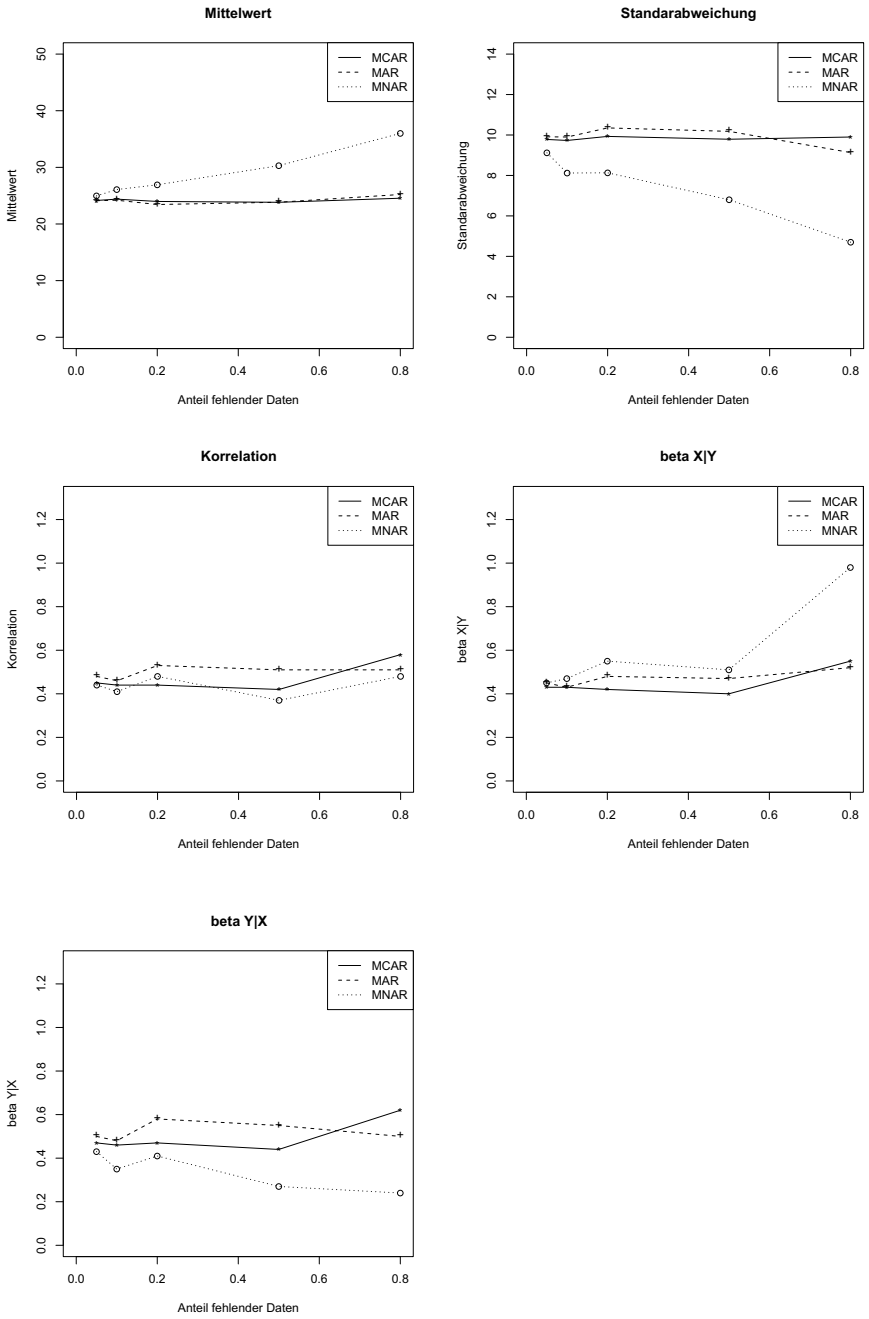

Abbildung 25: SAS Proc MI

7.11 Zusammenfassung Ergebnisse unter MCAR

Unter MCAR schätzen alle Verfahren die Parameter bei einem geringen Anteil an fehlenden Daten zuverlässig. Ab 50% fehlenden Daten haben vor allem Mittelwertimputation, Hot-Deck und SPSS Probleme. Listenweiser Fallausschluss, die stochastische Regression, SAS, SPLUS und das Sequentielle Hot-Deck Verfahren liefern auch bei hohem Anteil fehlender Daten zuverlässige Ergebnisse.

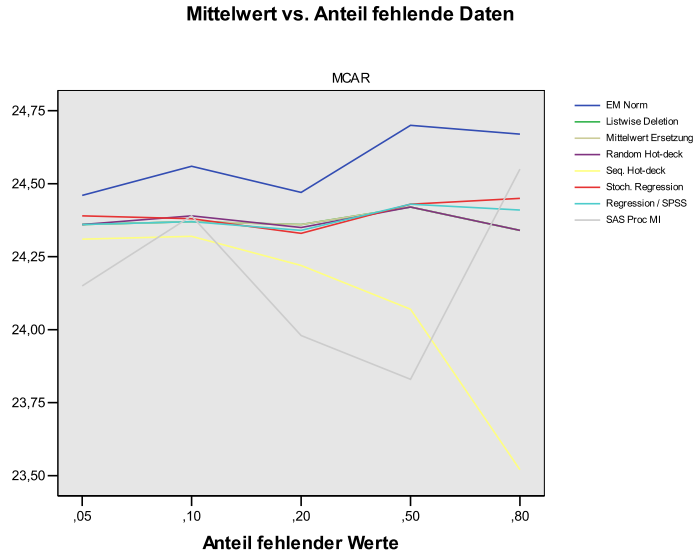

Abbildung 26: Mittelwert unter MCAR

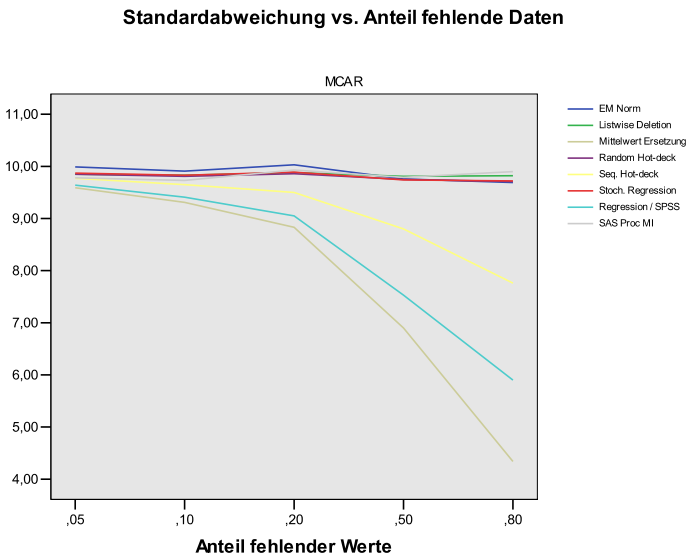

Abbildung 27: Varianz unter MCAR

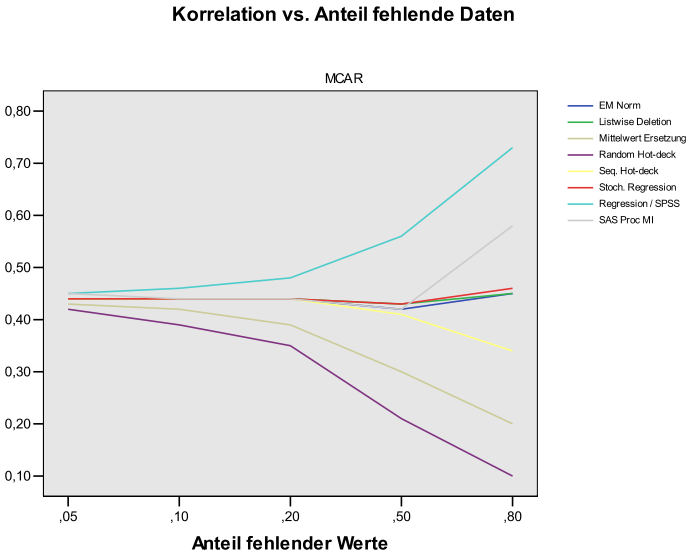

Abbildung 28: Korrelationskoeffizient unter MCAR

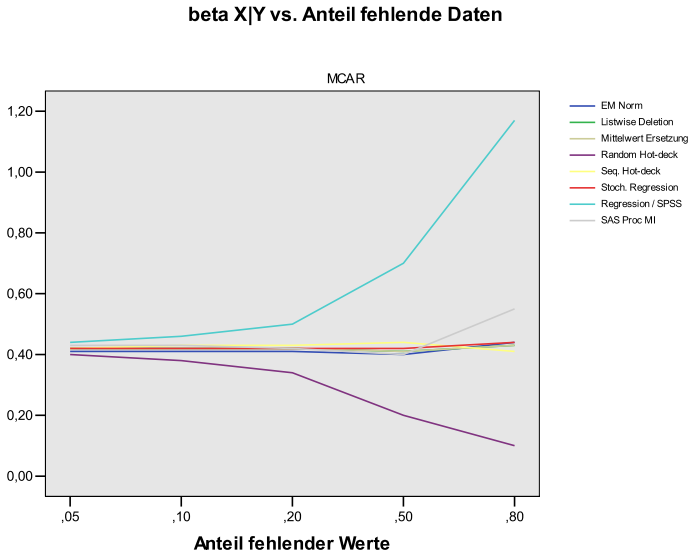

Abbildung 29: beta X—Y unter MCAR

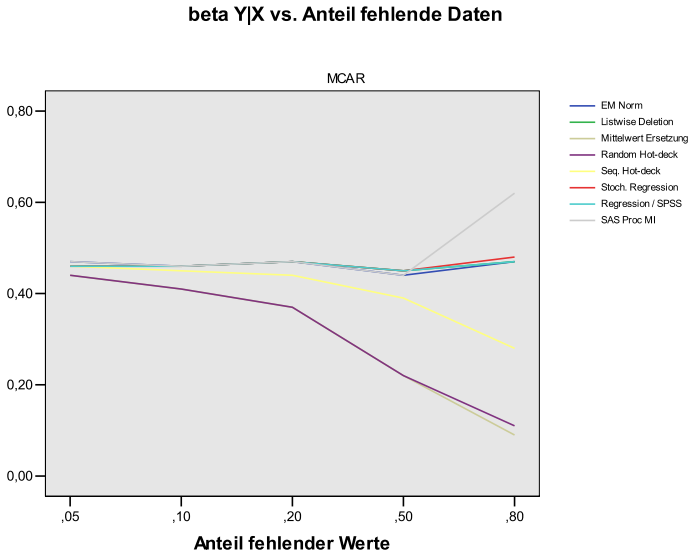

Abbildung 30: beta Y—X unter MCAR

7.12 Zusammenfassung Ergebnisse unter MAR

Die Probleme einiger Verfahren werden bei Missing-at-Random sichtbar. Hier erzielen das einfache Hot-Deck, der EM Algorithmus in SPSS und der Listenweise Fallausschluss keine sehr guten Ergebnisse. Auffallend ist die hohe Qualität der Data-Augmentation in R und der stochastischen Regressionsersetzung. Die MCMC Data-Augmentation in SAS hat bei steigendem Anteil fehlender Daten Probleme bei der Schätzung der Regressionsparameter.

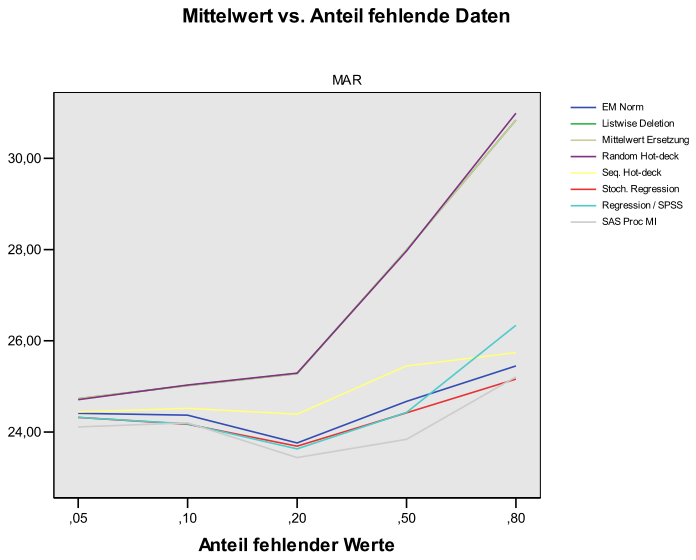

Abbildung 31: Mittelwert unter MAR

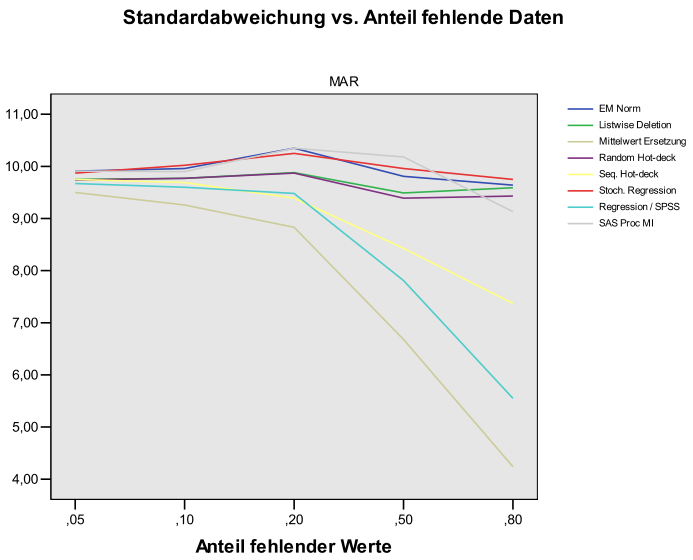

Abbildung 32: Varianz unter MAR

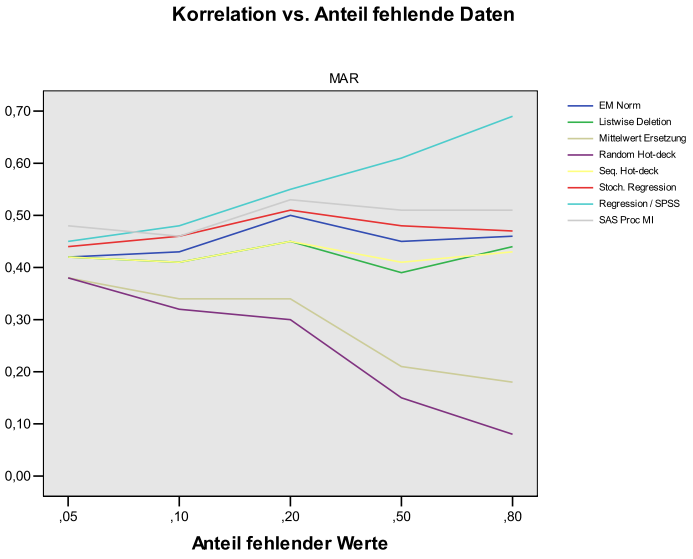

Abbildung 33: Korrelationskoeffizient unter MAR

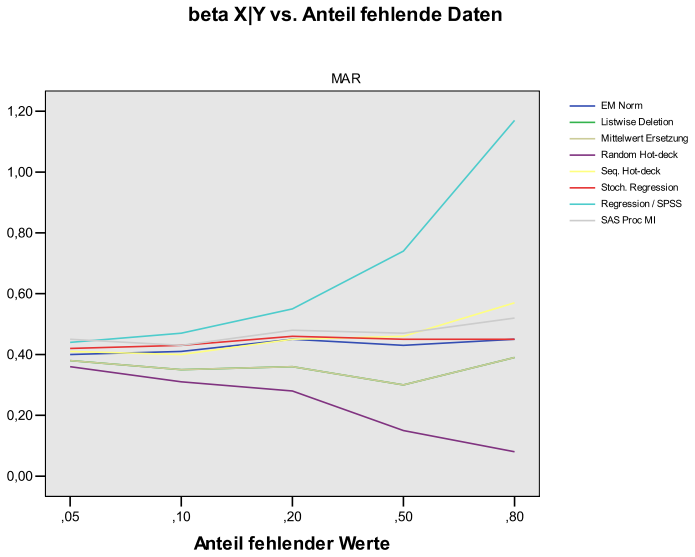

Abbildung 34: beta X—Y unter MAR

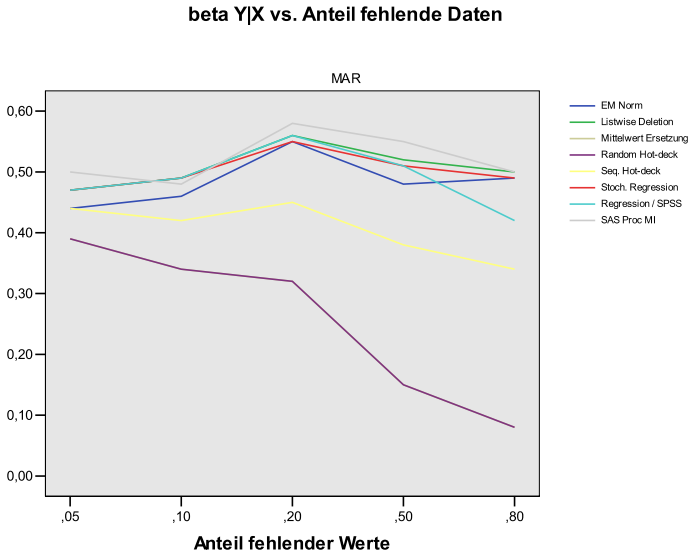

Abbildung 35: beta Y—X unter MAR

7.13 Zusammenfassung Ergebnisse unter MNAR

Für Missing-not-at-Random kommt es bereits bei einem sehr geringen Anteil von fehlenden Daten zu Fehlspezifikationen bei allen Verfahren. Kein Verfahren kann hier eine zuverlässige Schätzung liefern. Man erkennt besonders gravierend die Verzerrung bei Mittelwerten, der Standarabweichung und der Regression von Y auf X. Analysen, die mit diesen Datensätzen durchgeführt werden, führen zwangsläufig zu falschen Ergebnissen.

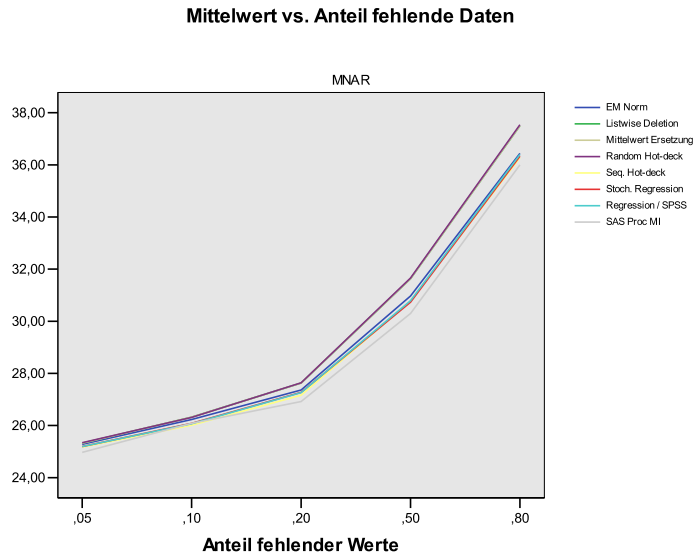

Abbildung 36: Mittelwert unter MNAR

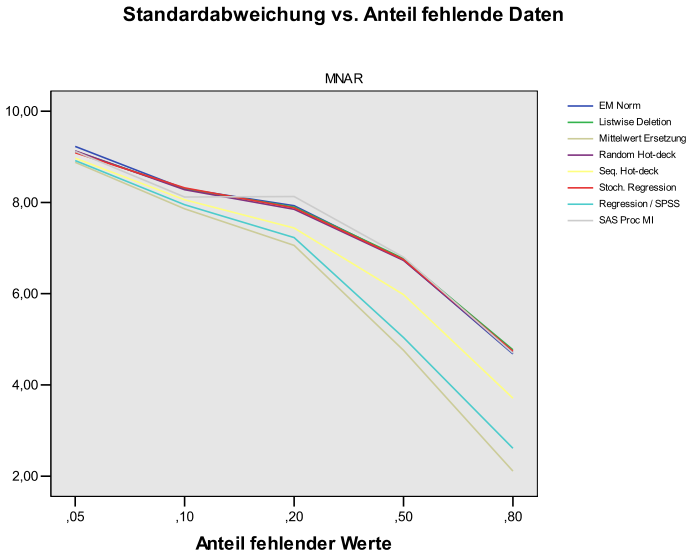

Abbildung 37: Varianz unter MNAR

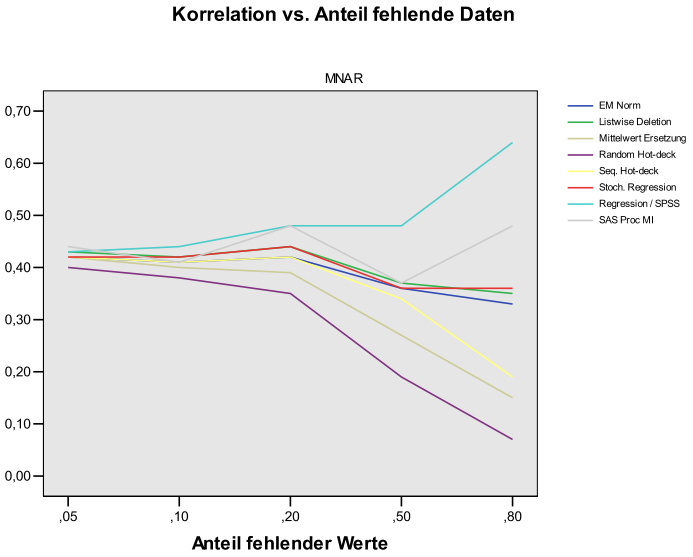

Abbildung 38: Korrelationskoeffizient unter MNAR

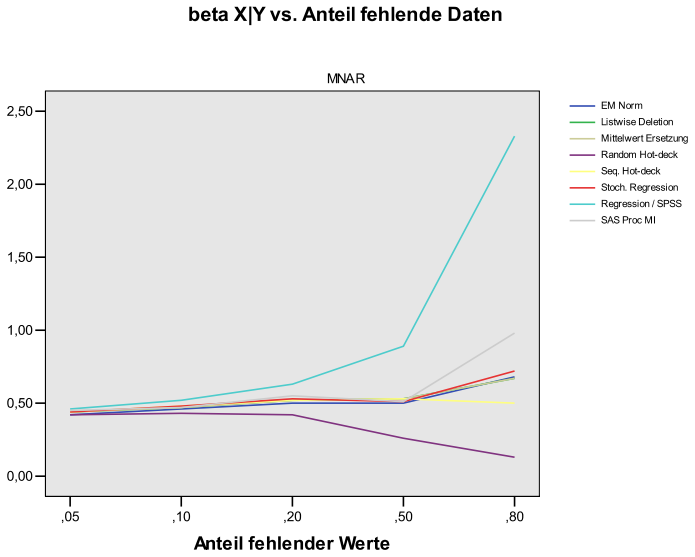

Abbildung 39: beta X—Y unter MNAR

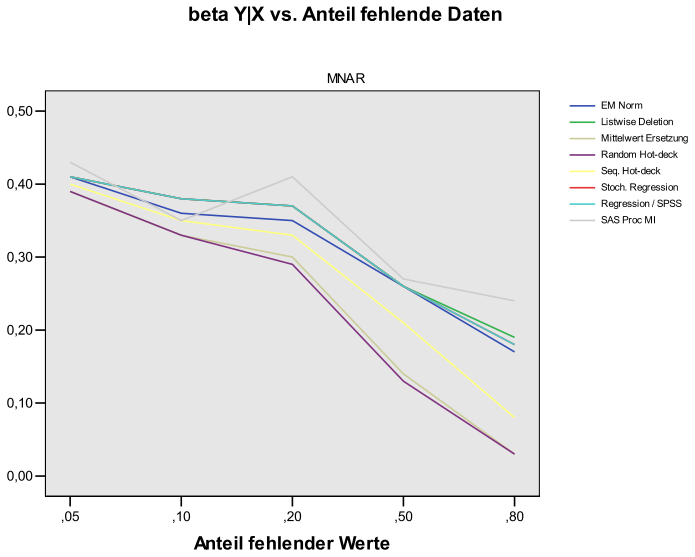

Abbildung 40: beta Y-X unter MNAR

8 Kategorielle Daten

In zahlreichen Umfragen werden häufig ausnahmslos kategorielle Variablen verwendet. Die meisten Verfahren zur Missing-Data Thematik wurden jedoch für stetige Daten entwickelt. In diesem Kapitel wird eine Maximum-Likelihood Methode für ordinale (geordnet kategoriell) Daten erläutert. In der Literatur wird oft vorgeschlagen für kategorielle fehlende Daten Methoden zu verwenden, die eigentlich für normalverteilte Daten entwickelt worden sind. In der Praxis gibt es jedoch kaum Datensätze die wirklich durch eine multivariate Normalverteilung aproximiert werden können. Trotzdem können diese Methoden zu brauchbaren Ergebnissen führen, vgl. Schafer (1997) und Alison (2002). Es wird ferner vorgeschlagen, die imputierten Werte in die nächste Kategorie zu runden.

8.1 Bespiel einer einfachen Maximum-Likelihood-Schätzung mit Kontingenztabellen

Für die Stichprobe von 195 Studenten werden die Statistik 1 und Statistik 2 Noten wie folgt umkodiert: (1.0...4.0)= 0 (Bestanden) 5.0=1 (Nicht Bestanden). Danach werden 20% der Beobachtungen in Statistik 2 mit einem ignorierbaren Mechanismus (MAR,MCAR) als fehlend kodiert. Es wird MCAR verwendet. Für die 156 vollständigen Beobachtungen erhält man folgende Kontingenztabelle.

		Y		
		0	1	Total
X	0	108	13	121
	1	27	8	35
Total		135	21	156

Tabelle 24: Kontingenztabelle Note 1 vs. Note 2

Für die anderen Beobachtungen wird nur die Statistik 1 Note beobachtet. Von den 39 fehlenden Beobachtungen nehmen die Statistik 1 Noten folgende Werte an: 35 Beobachtungen haben die Ausprägung X=0 und 4 Beobachtungen haben die Ausprägung X=1. Es sind folgende Wahrscheinlichkeiten zu schätzen.

		Y	
		0	1
X	0	p_{11}	p_{12}
	1	p_{21}	p_{22}

Tabelle 25: Kontingenztabelle Note 1 vs. Note 2

Stellen die 156 Beobachtungen die vollständigen Daten dar und es gäbe keine fehlenden Daten, dann sehe die Likelihood-Funktion wie folgt aus.

$$L = (p_{11})^{108}(p_{12})^{13}(p_{21})^{27}(p_{22})^{8} \tag{8.1}$$

unter der Nebenbedingung: $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

Die Wahrscheinlichkeiten $\tilde{p}_{ij}$ erhält man somit wie folgt.

$$\tilde{p}_{ij} = \frac{n_{ij}}{n} \tag{8.2}$$

Somit ergibt sich für

$$\tilde{p}_{11} = \left(\tfrac{108}{156}\right) = 0.692,$$

$$\tilde{p}_{12} = \left(\tfrac{13}{156}\right) = 0.083,$$

$$\tilde{p}_{21} = \left(\tfrac{27}{156}\right) = 0.173,$$

$$\tilde{p}_{22} = \left(\tfrac{8}{156}\right) = 0.05.$$

Jedoch stellen diese Schätzungen keine ML Schätzungen dar, da noch zusätzliche 39 Beobachtungen zur Verfügung stehen die in X beobachtet sind, jedoch für Y fehlende Werte aufweisen. Diese Beobachtungen müssen in der likelihood verwendet werden. Der Ablauf des Verfahrens wird wie folgt beschrieben, vgl. Allison (2001). Für einen ignorierbaren Mechanismus (MCAR,MAR) ist die likelihood für X=0 die marginale Wahrscheinlichkeit für das Auftreten von X=0 $(p_{11} + p_{12})$. Die likelihood für X=1, ist die marginale Wahrscheinlichkeit für das Auftreten von X=1 $(p_{21} + p_{22})$. Die likelihood sieht dann wie folgt aus.

$$L = (p_{11})^{108}(p_{12})^{13}(p_{21})^{27}(p_{22})^{8}(p_{11} + p_{12})^{35}(p_{21} + p_{22})^{4} \tag{8.3}$$

Für die meisten ML Probleme kann diese Likelihood-Funktion nur mit numerischen Methoden maximiert werden. Für den einfachen Fall, mit einem monotonen Datenmuster, kann man die bedingte Verteilung von Y gegeben X, sowie die marginale Verteilung von X jedoch relativ einfach berechnen. Für eine 2 x 2 Kontingenztabelle hat der ML Schätzer die generelle Form

$$\tilde{p}_{ij} = \tilde{p}(Y = j|X = i)\tilde{p}(X = i). \tag{8.4}$$

Die bedingten Wahrscheinlichkeiten werden aus den vollständigen Daten geschätzt. Dies geschieht durch Division der Zellhäufigkeiten durch die jeweiligen Randwahrscheinlichkeiten von X. Die marginalen Wahrscheinlichkeiten von X werden berechnet, indem man die Zeilenhäufigkeiten zu der Häufigkeit von X, die fehlende Daten in Y aufweisen, addiert und durch den gesamten Stichprobenumfang dividiert.

$$\tilde{p}_{11} = \left(\tfrac{108}{121}\right)\left(\tfrac{121+35}{195}\right) = 0.714$$

$$\tilde{p}_{21} = \left(\tfrac{13}{121}\right)\left(\tfrac{121+35}{195}\right) = 0.086$$

$$\tilde{p}_{12} = \left(\tfrac{27}{35}\right)\left(\tfrac{35+4}{195}\right) = 0.154$$

$$\tilde{p}_{22} = \left(\tfrac{8}{35}\right)\left(\tfrac{35+4}{195}\right) = 0.046$$

Wir haben nun Maximum-Likelihood Schätzungen der Zellwahrscheinlichkeiten erhalten. Eine allgemeine Methode soll im nächsten Abschnitt vorgestellt werden.

8.2 Generelle Maximum-Likelihood Schätzung für kategorielle Daten

Eine Anwendung eines iterativen EM-Algorithmus ist auch bei kategoriellen Daten möglich. Der Rechenaufwand ist jedoch viel größer, als beim EM-Algorithmus für kontinuierliche Daten, da für mehr als 2 Variablen mit k Ausprägungen bereits mit mehrdimensionalen Kontingenztabelle gearbeitet wird. Unter bestimmten Voraussetzungen folgen die Einträge in der Kontingenztabelle einer Multinomialverteilung. Die Multinomialverteilung wiederum gehört zur Familie der Exponentialverteilungen. Bei einem saturierten Nullmodell kann der EM folgendermaßen angewandt werden, vgl. Ming-Yi (1999). Der E-Schrit besteht darin, die Erwartungswerte der Zellhäufigkeiten zu berechnen.

$$\hat{n}^g_{ijk\ldots t} = E(n_{ijk\ldots t}|n, \theta_{ijk\ldots t}) = m_{ijk\ldots t} + \sum_{w,v} \gamma_{wv}\phi^g_{ijk\ldots t\in S_w} \quad (8.5)$$

Wobei

$$\phi^g_{ijk\ldots t\in S_w} = \frac{\hat{\theta}^g_{ijk\ldots t}\delta(ijk\ldots t\in S_w)}{\sum_{ijk\ldots t}\hat{\theta}^g_{ijk\ldots t}\delta(ijk\ldots t\in S_w)}$$

die aktuelle Schätzung der bedingten Wahrscheinlichkeit von Zelle (i,j,k...t) ist, gegeben dass eine Beobachtungen in eine Kombination von Kategorien S_w mit dem Muster w fällt. $\delta(\cdot)$ stellt die Indikatorfunktion dar. Der Exponent g steht für die g-te Iteration.

Der M-Schitt kalkuliert dann die Parameterschätzungen aufgrund der geschätzten Zellhäufigkeiten:

$$\hat{n}^g_{ijk\ldots t} = \frac{\hat{n}^g_{ijk\ldots t}}{N}$$

Als Startwerte für den Algorithmus können die beobachteten Zellhäufigkeiten verwendet werden. Ein ausführliche Beschreibung findet sich in Schafer (1997). Die Herangehensweise ist ähnlich wie bei der Data-Augmentation mit kontinuierlichen Daten. Wichtige Stichworte sind, Derichlet a-priori Verteilung, multidimensionale Kontingenztabellen, multinomiale Likelihood-Funktion, multinomiale a-posteriori Verteilung und Markov-Chain-Monte-Carlo Verfahren.

8.3 Simulationsstudie

Der Ablauf der Simulationsstudie ist ähnlich der Simulationsstudie mit kontinuierlichen Daten. In der Praxis sind insbesondere beim Erstellen von Fragebögen zwischen 5-7 Kategorien in den Items üblich. Daher sollen die Noten wie folgt umkodiert werden: 1.0=1, 1.3=1, 1.7=1, 2.0=2, 2.3=2, usw. Dadurch werden aus den 11 Kategorien insgesamt 5 Kategorien gebildet. Die untersuchten Parameter für die Qualität der Schätzungen sind Modus, Korrelation (Kendalls τ), die (25%,50%,75%) Quantile und die Entropie. Die Entropie,

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{8.7}$$

ist ein Maß zur Charakterisierung einer Verteilung.

Es werden folgende Verfahren Imp untersucht (Anm.: Listenweise Fallausschluss ist kein Imputationsverfahren):

- Listenweise Fallausschluss
- Verteilungsimputation
- EM-Algorithmus CAT in SPLUS/R
- PROC MI in SAS 9.1

8.4 Ergebnisse Simulationsstudie

Die Ergebnisse der Simulationsstudie sind leider nicht wie erwartet. Wichtige Parameter, wie die Quantile und die Entropie, ändern sich selbst unter MAR und MNAR kaum. Die Korrelation wird insbesondere bei der Verteilungsimputation (Hot-Deck) stark unterschätzt. Die Schätzungen des Listenweise Fallausschluss sind insbesondere unter MCAR sehr gut. Der EM-Algorithmus für kategorielle Daten kann hier nicht überzeugen, denn unter Missing-at-Random wird die Korrelation teils stark unterschätzt. Die besten Ergebnisse liefert die

Proc MI aus SAS. Für kategoriellen Daten kann man insgesamt, aufgrund der schlechten Simulationsergebnisse keine klare Aussage treffen. Es lässt sich festhalten, dass Verfahren die in der kommerziellen Software, wie SAS umgesetzt sind gut abschneiden. An dieser Stelle ist nicht klar, ob die Ergebnisse durch eine falsche Wahl der zu untersuchten Validierungsparameter, oder durch eine ungeeinete Datengrundlage entstanden sind. Alternativ hätte man die Unsicherheiten in den Schätzungen, auch mit Hilfe des Mean-Squared-Error (MSE) messen können, vgl. Härdle et al. (2004). Weitere Untersuchungen zum Thema sind nötig. Im folgenden sind die Simulationsergebnisse dargestellt. Die ursprünglichen Parameter stehen in der linken Spalte der Tabellen.

8.5 Listenweiser Fallaussschluss

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
Mod=3	3	3	3	2,97	2,81
$\rho = 0.29$	0,29	0,3	0,3	0,3	0,29
25%Quantil=2	2	2	2	2	2,06
Median=3	3	3	3	3	2,98
75 %Quantil=4	4	4	3,95	3,88	3,7
Entropie=0.29	0,29	0,29	0,3	0,33	0,4
MAR					
Mod=3	3	3	3	3	3
$\rho = 0.29$	0,27	0,25	0,23	0,23	0,22
25%Quantil=2	2	2	2	2,01	2,02
Median=3	3	3	3	3	3
75%Quantil=4	4	4	4	4	4
Entropie=0.29	0,29	0,29	0,29	0,31	0,31
MNAR					
Mod=3	3	3	3	3	4,66
$\rho = 0.29$	0,27	0,26	0,23	0,2	0,29
25% Quantil=2	2	2	2	3	2,04
Median=3	3	3	3	3	3,01
75% Quantil=4	4	4	4	4	5
Entropie=0.29	0,28	0,28	0,28	0,29	0,33

Tabelle 26: Listenweiser Fallausschluss

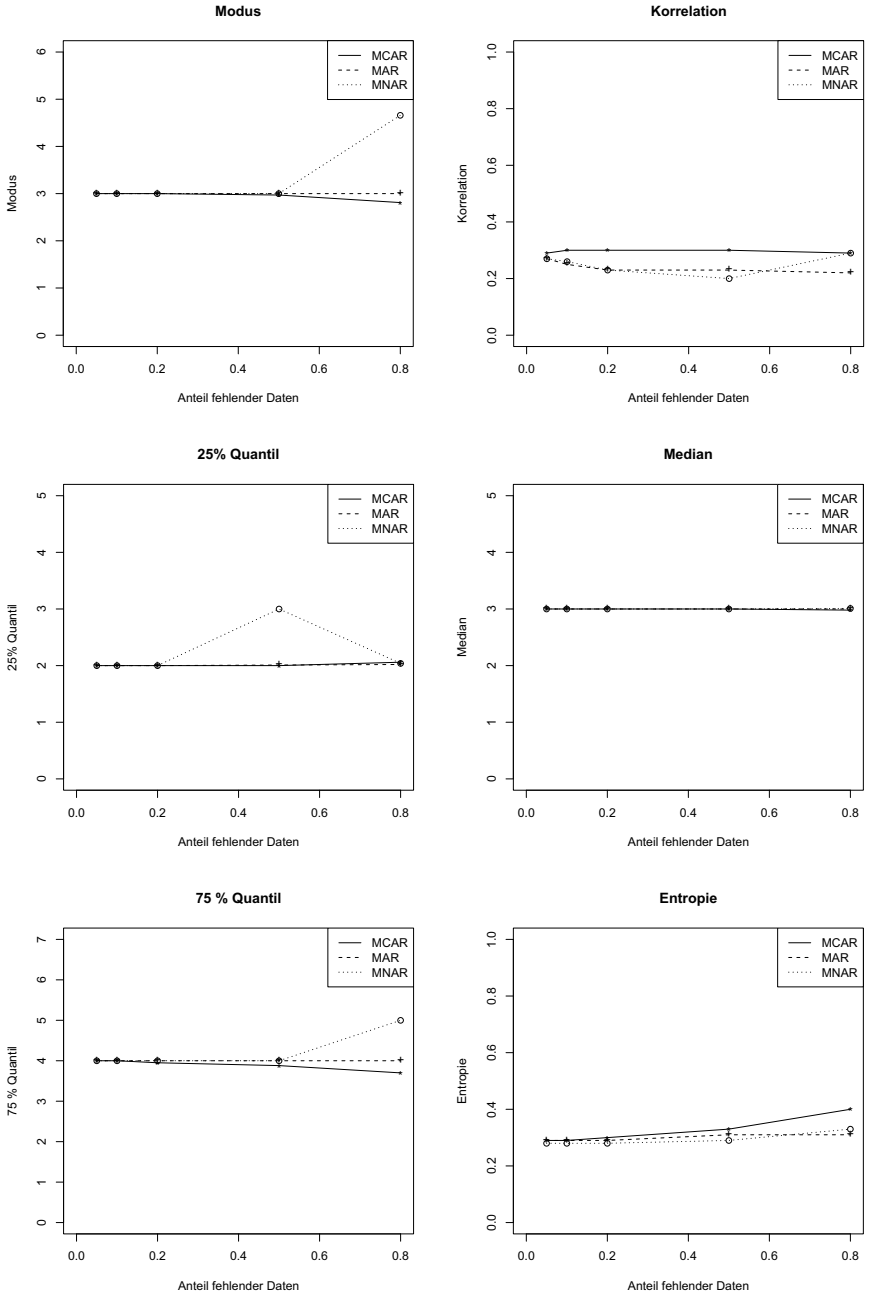

Abbildung 41: Listenweiser Fallausschluss

8.6 Verteilungsimpuation

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
Mod=3	3	3	3	2,89	2,76
$\rho = 0.29$	0,28	0,27	0,24	0,15	0,06
25% Quantil=2	2	2	2	2	2,08
Median=3	3	3	3	3	2,93
75 %Quantil=4	4	4	3,9	3,82	3,82
Entropie=0.29	0,29	0,29	0,29	0,29	0,28
MAR					
Mod=3	3	3	3	3	3
$\rho = 0.29$	0,25	0,22	0,19	0,16	0,15
25% Quantil=2	2	2	2	2,04	2,07
Median=3	3	3	3	3	3
75 %Quantil=4	4	4	4	4	3,99
Entropie=0.29	0,29	0,28	0,28	0,28	0,28
MNAR					
Mod=3	3	3	3	3	4,51
$\rho = 0.29$	0,26	0,24	0,2	0,13	0,14
25% Quantil=2	2	2	2,02	2,99	2,12
Median=3	3	3	3	3	3,01
75 %Quantil=4	4	4	4	4,06	5
Entropie=0.29	0,28	0,27	0,27	0,27	0,28

Tabelle 27: Verteilungsimputation

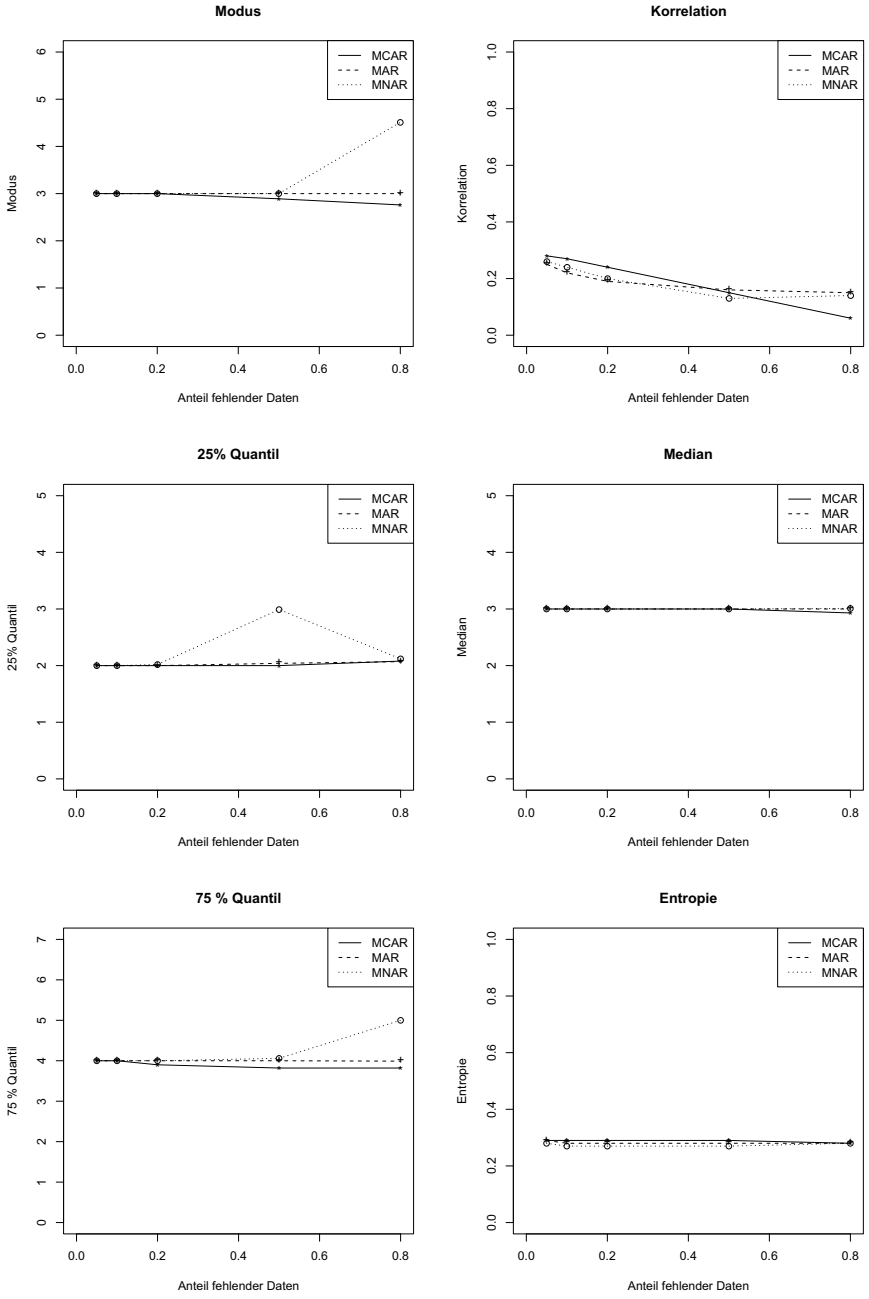

Abbildung 42: Verteilungsimputation

8.7 EM-CAT SPLUS/R

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
Mod=3	3	3	2,99	2,95	3,16
$\rho = 0.29$	0,29	0,26	0,24	0,14	0,13
25% Quantil=2	2	2	2	2	2,17
Median=3	3	3	3	3	3,12
75 %Quantil =4	4	4	4	4	4,13
Entropie=0.29	0,29	0,29	0,29	0,29	0,28
MAR					
Mod=3	3	3	3	3	3,01
$\rho = 0.29$	0,23	0,21	0,17	0,12	0,21
25% Quantil=2	2	2	2	2,04	2
Median=3	3	3	3	3	3,01
75 %Quantil=4	4	4	4	4	4,01
Entropie=0.29	0,29	0,29	0,29	0,29	0,29
MNAR					
Mod=3	3	3	3	3,48	4,8
$\rho = 0.29$	0,26	0,23	0,2	0,17	0,2
25% Quantil=2	2	2	2,35	2,72	2,34
Median=3	3	3	3	3,34	3,35
75 %Quantil=4	4	4	4	4,11	5
Entropie=0.29	0,28	0,28	0,28	0,29	0,29

Tabelle 28: EM-CAT SPLUS/R

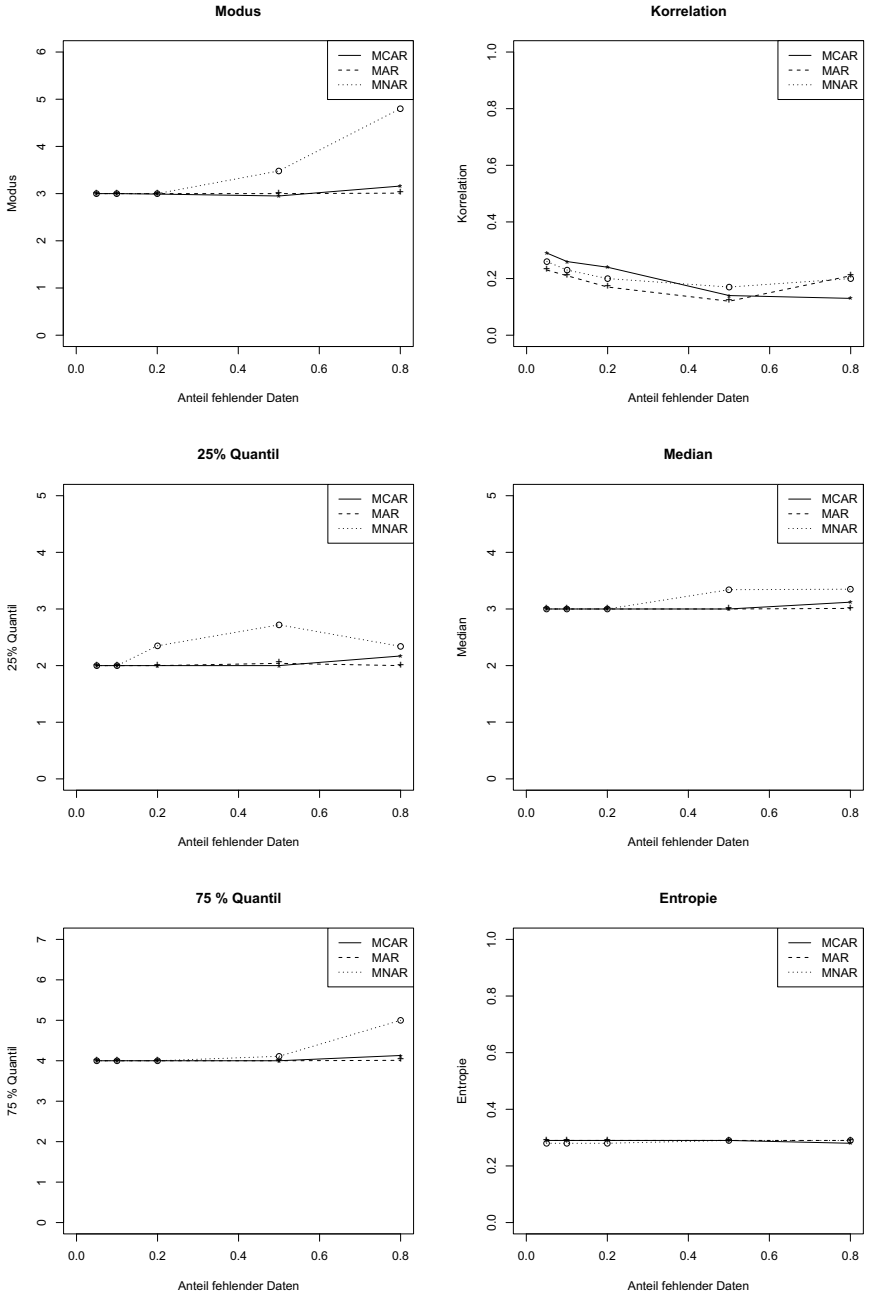

Abbildung 43: EM-CAT SPLUS/R

8.8 SAS Proc MI mit MCMC

Parameter	0.05	0.10	0.20	0.50	0.80
MCAR					
Mod=3	3	3	3	2,97	2,81
$\rho = 0.29$	0,29	0,29	0,27	0,26	0,27
25%Quantil=2	2	2	2	2	2,17
Median=3	3	3	3	2,98	3
75 %Quantil=4	3,99	3,99	3,9	3,78	3,77
Entropie=0.29	0,33	0,38	0,46	0,71	0,92
MAR					
Mod=3	3	3	3	3	3
$\rho = 0.29$	0,28	0,27	0,25	0,24	0,24
25% Quantil=2	2	2	2	2,02	2,03
Median=3	3	3	3	3	3
75 %Quantil=4	4	4	4	3,92	3,84
Entropie=0.29	0,32	0,36	0,41	0,57	0,6
MNAR					
Mod=3	3	3	3	3	4,76
$\rho = 0.29$	0,27	0,26	0,23	0,21	0,27
25% Quantil=2	2	2	2,02	2,97	2,19
Median=3	3	3	3	3,01	3,12
75 %Quantil=4	4	4	4	4,06	4,48
Entropie=0.29	0,32	0,34	0,39	0,54	0,72

Tabelle 29: SAS Proc MI mit MCMC

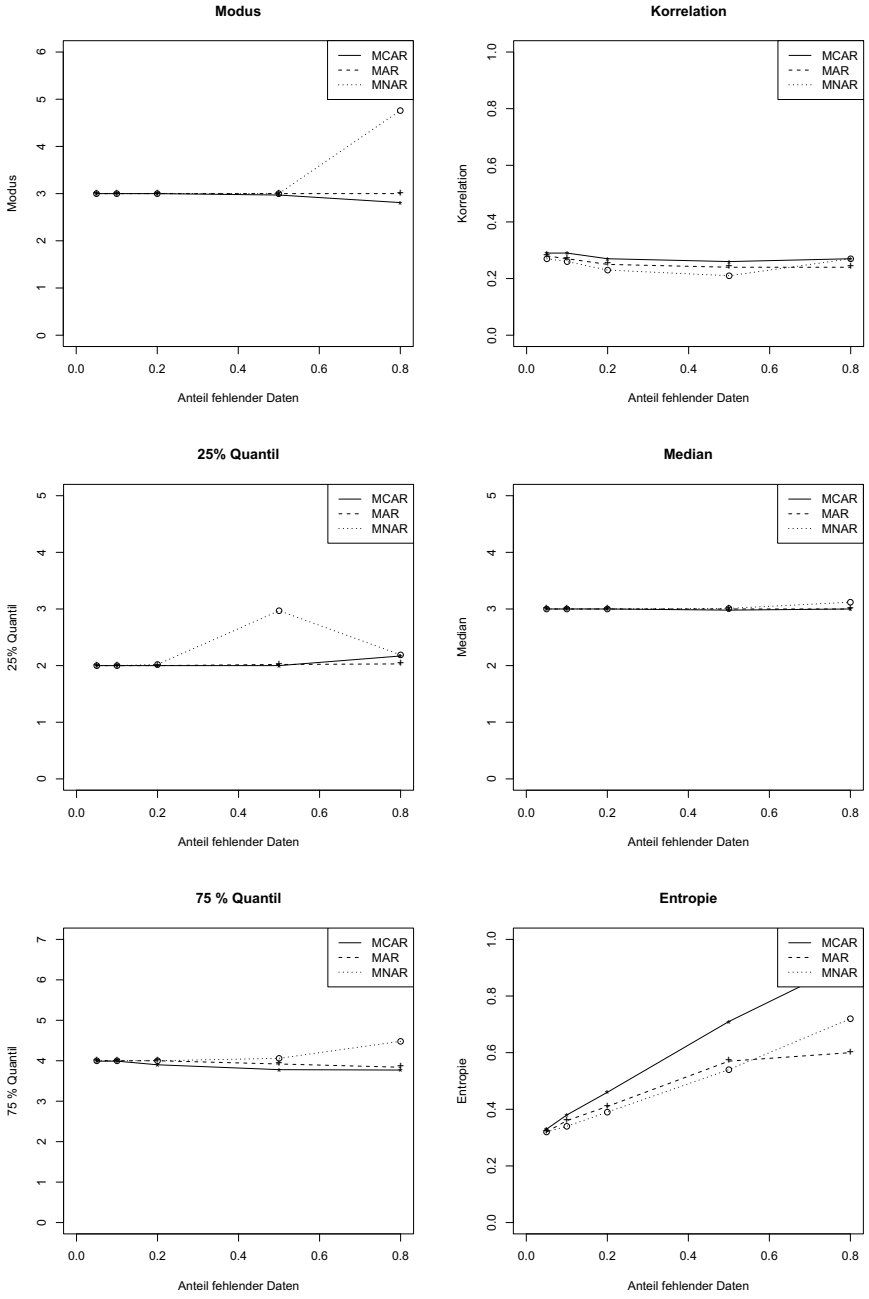

Abbildung 44: SAS Proc MI mit MCMC

9 Fazit

Der Umgang mit fehlenden Daten kann sehr schwierig sein. Falls man Einfluss auf die Erhebung von Daten hat, sollte darauf geachtet werden, dass Fragen klar formuliert sind und den Teilnehmern ein vertraulicher Umgang mit den Daten garantiert wird. Auch bietet sich die Möglichkeit, Fragen einzubauen, die mit Fragen mit einem erfahrungsgemäß hohem Anteil fehlender Daten korreliert sind. So korreliert das Einkommen in der Regel mit dem Alter oder der Wohngegend. Sollte man später jedoch mit fehlenden Daten arbeiten, so gibt es gravierende Unterschiede bei der Verwendung von Missing-Data Verfahren. Nach den Simulationen für stetige Daten hat sich gezeigt, dass Methoden wie der EM-Algorithmus und Data-Augmentation für einen ignorierbaren Datenmechanismus sehr gute Ergebnisse erzielen. Verfahren wie Mittelwertersetzung, Fallausschluss, Verteilungsimputationen und Regressionsmethoden ohne stochastische Komponente sind zu vermeiden. Der Fallweise Ausschluss erzielt nur unter MCAR gute Ergebnisse. Will man Analysen wie Strukturgleichungsmodelle durchführen, so sollte man unbedingt Verfahren wie den EM-Algorithmus oder Data-Augmentation verwenden, da nur dort die Korrelationsmatrix zuverlässig geschätzt wird. Eine große Auswahl, an Einstellung für Data-Augmentation, findet man in der SAS Proc MI. Auf den Einsatz von SPSS sollte verzichtet werden. Die Probleme der Verteilungsimputation (Hot-Deck) erkennt man sowohl bei stetigen, als auch bei kategoriellen Daten. Hier werden Zusammenhänge in den Daten zerstört. Im Falle eines nichtignorierbaren Mechanismus der fehlenden Daten (MNAR), erzielt kein Verfahren zufriedenstellende Ergebnisse.

10 Literatur

Allison, P.D. (2001). Missing Data, in Series: Quantitative Applications in the social sciences, *Sage University Paper*.
Brandes, U. (2005) Statistische Bewertung und Analyse der Klausurergebnisse Statistik (Grundstudium) , *Diplomarbeit*, Berlin.
Buck, S.F. (1960). A methodof estimation of missing values in multivariate data suitable for use with an electronic computer, *J. Roy. Statost. Soc.* B22, 302-306 Dempster, A.P., Laird, N.M and Rubin, D.B (1977), Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statisticl Society Series B*, 1-38.
Dempster, A.P. and Rubin, D.B (1983), Introduction in Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography, *New York: Academic Press. Journal of the Royal Statisticl Society Series B*, 1-38.
Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 463-479.
Enders, C.K (2001). The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. *Educational and Psychological Measurement 61.*
Härdle, W. und Simar, L. (2003). Applied Multivariate Statistical Analysis. *Springer*, Berlin.
W. Härdle, M. Müller, S. Sperlich and A. Werwatz, (2004). Nonparametric and Semiparametric Models. *Springer*, Berlin.
Heckman, J. J. 1979. Sample Selection Bias as a Specification Error. *Econometrica 47(1)*. 153-161.
Ming-Yi, H. (1999). Model Checking for Incomplete High Dimensional Categorical Data. *PhD. Thesis*, Los Angeles.
Kalton, G. und Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology 12*, 1-16.
Kennedy, P.E (2004). A guide in Econometrics. *Malden, Mass.: Blackwell.*
Ki-Yeol Kim, Byoung-Jin Kim, Gwan-Su Yi (2004). Reuse of imputed data in microarray analysis increases imputation efficiency, *BMC Bioinformatics*, 5-160.
Lehmann, P. (2005). Ein Vergleich von Logit und CART - Eine Monte Carlo Studie, *Diplomarbeit*, Berlin.
Little, R.J.A. and Rubin, D.B. (1987), Statistical Analysis with missing data, *Wiley and Sons*, New York.
Little, R.J.A., Rubin, D.B. (2002), Statistical Analysis with missing data, *John Wiley*, New Jersey.
McLachlan, G. and Krishnan, T. The EM Algorithm and Extensions. *John Wiley and Sons, New York, 1997*
Rönz, B. (2001). Skript - Computergestützte Statistik I, *Vorlesungsskript*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.
Rönz, B. (2000). Skript - Computergestützte Statistik II, *Vorlesungsskript*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.
Rubin, D.B. (1976a). Inferences and missing data, *Biometrika*, 581-592

Rubin, D.B. (1978b). Multiple imputation in sample surveys, *American Statistical Association*, 1978, 20-34
Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, *Wiley and Sons*, New York.
Schafer, Joseph L. (1997), Analysis of Incomplete Multivariate Data, *Chapman and Hall*, New York.
Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2): 147-177.
Silverman, B.W (1986). Density Estimation for Statistics and Data Analysis, Vol 26 of *Monograhs on Statistics and Applied Probability*, Chapman and Hall, London.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. Hastie, T., Tibshirani, R., Bostein, D., und Altman, R.B. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatic, 17(6)*, 520-525.
von Hippel, P.T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician 58(2)*,160-164.
Welch, B. L. (1947), The generalization of Students problem when several different population variances are involved, *Biometrica 34*, 28-35.
Yuan, Yang C. (2003). Multiple Imputation for Missing Data: Concepts and New Development. *Rockville, MD: SAS Institute Inc.* (Paper 267-25).