Universität Trier

Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK)

Trier, den 3. 7. 2013

Bernhard Baltes-Götz

Behandlung fehlender Werte in SPSS und Amos

Inhaltsverzeichnis

Herausgeber: Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK)
an der Universität Trier
Universitätsring 15
D-54286 Trier
WWW: http://www.uni-trier.de/index.php?id=518
E-Mail: zimk@uni-trier.de
Tel.: (0651) 201-3417, Fax.: (0651) 3921
Autor: Bernhard Baltes-Götz (E-Mail: baltes@uni-trier.de)

Vorwort

In diesem Manuskript geht es um das bei empirischen Studien fast allgegenwärtige Problem fehlender Werte. Für traditionelle Behandlungsmethoden (z.B. fallweiser Ausschluss, Ersetzung durch Mittelwerte) und moderne Alternativen (z.B. direkte Maximum Likelihood - Schätzung, multiple Imputation) werden …

- statistische Grundlagen erläutert,
- Anwendungsbeispiele mit SPSS Statistics 21 und Amos 21 vorgeführt.

Die aktuelle Version des Manuskripts ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

ZIMK (Rechenzentrum) > Infos für Studierende > EDV-Dokumentationen >

Statistik > Behandlung fehlender Werte in SPSS und Amos

Leider sind in diesem Manuskript einige Teile unter Zeitdruck entstanden, so dass Unzulänglichkeiten zu befürchten sind. Kritik und Verbesserungsvorschläge zum Manuskript werden dankbar entgegen genommen (z.B. unter der Mail-Adresse baltes@uni-trier.de).

Trier, im Juli 2013 Bernhard Baltes-Götz

1 Einleitung

Fehlende Werte sind bei empirischen Studien aus zahlreichen Gründen kaum zu vermeiden. So rechnet Acock (2005, S. 1014) z.B. bei der Frage nach dem Einkommen mit einer Ausfallrate von 30%. Klassische statistische Auswertungsverfahren (z.B. Regressions-, Faktoren- oder Diskriminanzanalyse) erfordern jedoch *komplette* Fälle. Je mehr Variablen beteiligt sind, desto kleiner wird die Schnittmenge mit den Fällen, die bei *allen* Variablen einen gültigen Wert abgeliefert haben. Unter der vereinfachenden Annahme, dass für k beteiligte Variablen die Ereignisse

$$\{\text{gültiger Wert bei Variable } j\},\ j = 1, .., k$$

unabhängig sind, erhält man bei einer für alle Variablen identischen Wahrscheinlichkeit p_s für einen gültigen Wert bei einer einzelnen Variablen in Abhängigkeit von der Anzahl k folgende Wahrscheinlichkeit p_c für einen vollständigen Datensatz:

$$p_c = p_s^k$$

Bei $p_s = 0{,}97$ fällt der Anteil kompletter Fälle in Abhängigkeit von der Variablenzahl schnell unter 50%:

k	p_c
5	0,86
10	0,74
20	0,54
30	0,40

Allerdings entscheidet in der Regel *nicht* der pure Zufall über das Fehlen eines Wertes. Oft hängt die Wahrscheinlichkeit für das Fehlen eines Wertes bei einer Variablen i von den Ausprägungen *anderer* Variablen ab. So könnte die Ausfallwahrscheinlichkeit bei der Frage nach dem Einkommen von der ebenfalls erfragten Einstellung zur Steuerehrlichkeit abhängen. Konstellationen dieses Typs können von modernen statistischen Verfahren gut behandelt werden.

Leider hängt nicht selten die Wahrscheinlichkeit für das Fehlen eines Wertes von dessen Ausprägung ab, was z.B. bei der Frage nach dem Einkommen zu befürchten ist. Auch moderne statistische Verfahren sind überfordert, wenn diese Abhängigkeit auch nach Berücksichtigung von beobachteten Ursachen für das Auftreten fehlender Werte besteht, wenn also im Einkommensbeispiel bei Personen mit gleicher Einstellung zur Steuerehrlichkeit die Wahrscheinlichkeit für einen fehlenden Einkommenswert von seiner Höhe abhängt.

Die in Statistikprogrammen meist voreingestellte fallweise Behandlung fehlender Werte (Beschränkung auf die vollständigen Fälle) gehört *nicht* zu den modernen statistischen Verfahren. Hier drohen:

- Verzerrte Schätzergebnisse
  Wenn nicht der pure Zufall über das Auftreten fehlender Werte entscheidet, resultieren verzerrte Parameterschätzer und entsprechend falsche Schlüsse.
- Verlust an Präzision
  Es ist bedauerlich, wenn die im Datensatz enthaltene Information zu einem erheblichen Teil verloren geht. Dies führt zu vergrößerten Standardfehlern und Konfidenzintervallen bei Parameterschätzungen sowie zu einer reduzierten Power bei Hypothesentests.

Um die Beeinträchtigung der Forschung durch fehlende Werte gering zu halten, sind folgende Maßnahmen erforderlich:

- Fehlende Werte vermeiden
  Bei der Datenerhebung sind fehlende Werte nach Möglichkeit zu vermeiden, was aber nur in seltenen Fällen perfekt gelingen wird. Relativ günstige Bedingungen bestehen z.B. bei der Online-Forschung mit Internet-Techniken, wo Auskunftspersonen nach dem Abschicken eines lückenhaft ausgefüllten Formulars um vollständige Antworten gebeten werden können.

- Bestmögliche statistische Behandlung fehlender Werte
  Bei der statistischen Auswertung sind Verfahren zu verwenden, die unter möglichst allgemeinen Bedingungen fehlende Werte kompensieren und verzerrte Forschungsergebnisse verhindern können. Anschließend wird beschrieben, welche Verfahren zur Behandlung fehlender Werte in den Produkten in der SPSS-Software-Familie verfügbar sind.

IBM SPSS Statistics bietet im Erweiterungsmodul **Missing Values** *zwei* Prozeduren zur Analyse und Behandlung fehlender Werte. In der folgenden Auflistung der verfügbaren Leistungen tauchen etliche im weiteren Kursverlauf noch zu erläuternde Begriffe auf:

- **MVA**
  Die ältere Prozedur MVA (*Missing Values Analysis*), deren Leistungen auch über den Menübefehl **Analysieren > Analyse fehlender Werte** abrufbar sind, bietet u.a.:
  - Univariate Analysen (z.B. Anteile fehlender Werte, Anzahl der Ausreißer)
  - Einfaches Ersetzen fehlender Werte (z.B. per multipler Regression)
  - Schätzung von Mittelwerten, Varianzen und Kovarianzen per EM-Algorithmus
  - Test nach Little zur Überprüfung der MCAR-Bedingung (rein zufälliges Auftreten fehlender Werte)
- **Multiple Imputation**
  Diese seit SPSS Statistics 17 verfügbare Prozedur, deren Leistungen auch über den Menübefehl **Analysieren > Multiple Imputation** abrufbar sind, unterstützt neben einer Analyse der aufgetretenen Muster fehlender Werte die *multiple Imputation*. Dabei entstehen *mehrere* (z.B. fünf) vervollständigte Datensätze, um die Unsicherheit bzgl. der beim Ersetzen fehlender Werte verwendeten Parameter zu berücksichtigen. Bei den eigentlich intendierten Auswertungen ist einiger Aufwand erforderlich, den SPSS Statistics zum Glück in vielen Fällen automatisiert:
  - Wiederholung mit jedem einzelnen Imputationsdatensatz
  - Zusammenfassung der Ergebnisse

Von den eben genannten Verfahren zur Behandlung fehlender Werte ist nur die multiple Imputation durchweg zu empfehlen. Mit der direkten **FIML-Methode** (*Full Information Maximum Likelihood*) steht eine weitere, im selben guten Ruf stehende Lösung für das Problem fehlender Werte zur Verfügung. Im Rahmen der IBM SPSS -Produktfamilie wird diese Methode vom Strukturgleichungsanalyseprogramm **IBM SPSS Amos** angeboten.

Im Manuskript können bei weiten nicht alle Detailprobleme im Zusammenhang mit fehlenden Werten behandelt werden:

- Wir konzentrieren uns auf Regressionsmodelle und ignorieren z.B. Probleme und Techniken bei der Schätzung von univariaten Verteilungsaspekten (z.B. Erwartungswert).
- Man kann zwischen komplett fehlenden Fällen und fehlenden Einzelwerten unterscheiden. Für das zuerst genannte Problem sind Gewichtungsverfahren vorgeschlagen worden, die im Manuskript nicht behandelt werden (siehe z.B. Little & Rubin 2002, Abschnitt 3.3).
- Ebenso werden die speziellen Probleme von Längsschnittstudien (Panelstudien) mit vorzeitig ausgestiegenen Fällen ignoriert.

2 Klassifikation fehlender Werte

Wir betrachten das Auftreten fehlender Werte als stochastisches Phänomen und definieren zu jeder bei einer statistischen Analyse beteiligten X_j eine Missing Data (MD) – Indikatorvariable M_j:

$$M_j = \begin{cases} 1, & \text{falls der Beobachtungswert zu } X_j \text{ fehlt} \\ 0, & \text{sonst} \end{cases}$$

Rubin (1976) hat über Beziehungen zwischen den MD-Indikatorvariablen und den eigentlichen Beobachtungsvariablen eine allgemein anerkannte Klassifikation fehlender Werte begründet, die anschließend vorgestellt werden soll. Später werden wir die verschiedenen MD-Behandlungsmethoden danach beurteilen, bei welchen Rubin-Typen sie anwendbar sind.

2.1 MCAR

Eine statistische Analyse mit den Variablen X_1, …, X_k erfüllt die MCAR-Bedingung (*Missing Completely At Random*), wenn für jede Variable X_j gilt: Die Wahrscheinlichkeit für einen fehlenden Wert bei X_j hängt weder von der X_j -Ausprägung noch von den Ausprägungen der restlichen Variablen ab:

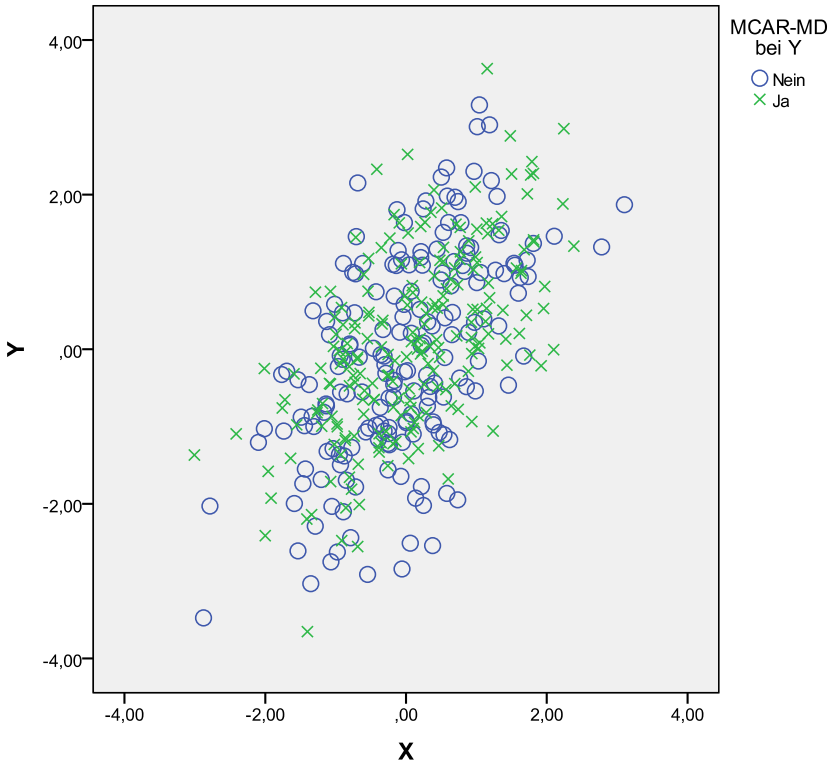
$$\mathrm{P}(\{M_j = 1\} \mid X_1,..., X_k) = c_j \quad (\in [0, 1])$$

Die für den Ausfall eines X_j –Wertes verantwortlichen Ursachen stehen in keiner Beziehung zu den Variablenausprägungen. Es darf durchaus eine Überlappung mit den Ursachen für den Ausfall von Werten bei anderen Variablen geben (z.B. aufgrund des Persönlichkeitsmerkmals *Faulheit* bei einer Befragung), so dass Fälle ohne X_j –Wert auch bei anderen Variablen ein erhöhtes Ausfallrisiko haben. Diese MD-Ursachen müssen aber von allen beobachteten Variablen in der Analyse unabhängig sein.

Ist die MCAR-Bedingung erfüllt, haben für jede Variable X_j die beiden Teilpopulationen mit $\{M_j = 1\}$ und $\{M_j = 0\}$ bei allen Variablen X_1, …, X_k dieselbe Verteilung. Diese Bedingung lässt sich für alle Variablen mit einem von j verschiedenen Index m überprüfen. Bei metrischen Variablen wird man sich in der Regel auf die Erwartungswerte der beiden Teilpopulationen beschränken und unter Verwendung der Fälle mit gültigem X_m - Wert die Nullhypothese gleicher Erwartungswerte über einen t-Test für unabhängige Stichproben prüfen. Signifikante Testergebnisse sprechen gegen die lokale MCAR-Bedingung und geben Hinweise auf Prädiktoren für die MD-Wahrscheinlichkeit bei X_j. Diese sollten bei einer Analyse mit MAR-pflichtiger MD-Behandlung (siehe unten) einbezogen werden, um die MAR-Bedingung plausibel zu machen.

Ein von **Little** entwickeltes Testverfahren, das SPSS Statistics bei vorhandenem Modul *Missing Values* beherrscht, erlaubt die *globale* Beurteilung der MCAR-Bedingung. Littles Test berechnet für jedes Muster fehlender Werte (z.B. für die Teilstichprobe mit gültigen Werten bei den Variablen X_1, X_2 und X_3 sowie fehlenden Werten bei den Variablem X_4, ..., X_k) einen streuungsnormierten Abstand des Vektors mit den Teilstichprobenmittelwerten vom Vektor mit den Mittelwerten der Gesamtstichprobe. Die Abstandmaße zu den einzelnen MD-Mustern werden mit der jeweiligen Teilstichprobengröße gewichtet und aufsummiert, wobei eine Prüfgröße entsteht, die bei erfüllter MCAR-Bedingung einer χ^2-Verteilung folgt. Bei akzeptierter Nullhypothese kann man sich berechtigt fühlen, eine MCAR-pflichtige MD-Behandlung einzusetzen (z.B. den simplen fallweisen Ausschluss).

Wie die lokale und globale MCAR-Testung mit SPSS Statistics durchgeführt wird, erfahren Sie in Abschnitt 3.2. Anschließend soll ein graphischer Eindruck von einer MCAR- Verteilung vermittelt werden. Das folgende Streudiagramm zeigt die gemeinsame empirische Verteilung einer Variablen *X* mit vollständig vorhandenen Werten und einer Variablen *Y* mit teilweise fehlenden Werten nach dem MCAR-Prinzip. Fälle mit vorhandenen Beobachtungswerten für *X und Y* sind durch einen blauen Kreis dargestellt. Datenpunkte mit fehlendem *Y*-Wert sind durch ein grünes Kreuz markiert:



Im Little-Test wird für diese Daten erwartungsgemäß die MCAR-Nullhypothese akzeptiert:

EM-Kovarianzen[a]

	X	Y
X	,99378	
Y	,74987	1,80602

a. MCAR-Test nach Little: Chi-Quadrat = 1,383, DF = 1, Sig. = ,240

Die von SPSS per EM-Algorithmus (siehe Abschnitt 5.1) ermittelte und in der Tabelle protokollierte Kovarianz ist erwartungsgemäß relativ präzise geschätzt (wahrer Wert: 0,7).

2.2 MAR

Wenn mit M_j die Indikatorvariable für das Fehlen des X_j-Wertes bezeichnet wird, dann verlangt die MCAR-Bedingung, dass M_j von allen Variablen X_1, …, X_k unabhängig sein. Demgegenüber fordert die MAR-Bedingung (*Missing At Random*), dass M_j nach Kontrolle der Abhängigkeiten von *beobachten* Variablen nicht mehr von X_j abhängen darf.

In der Einleitung wurde ein Beispiel mit den Variablen Einkommen und Einstellung zur Steuerehrlichkeit erwähnt. Es liegt *kein* Verstoß gegen die MAR-Bedingung vor, wenn die Ausfallwahrscheinlichkeit bei der Frage nach dem Einkommen von der Einstellung zu Steuerehrlichkeit abhängt, solange bei Personen mit derselben Einstellung zur Steuerehrlichkeit die Wahrscheinlichkeit für einen fehlenden Einkommenswert nicht von dessen Höhe abhängt.

Offenbar ist die MAR-Bedingung weniger streng als die MCAR-Bedingung und damit realistischer. Unter der MCAR-Bedingung ist auch die MAR-Bedingung erfüllt.

Die Bezeichnung *MAR* ist unglücklich gewählt, weil sie den Inhalt des Begriffs ziemlich im Unklaren lässt und zudem leicht mit *MCAR* verwechselt werden kann.

Um die MAR-Bedingung zu klären, beschränken wir uns auf *zwei* Variablen, die zur Vermeidung von Indexaufwand als *X* und *Y* bezeichnet werden sollen. Zunächst vereinfachen wir noch weiter und nehmen an, dass nur bei der Variablen *Y* fehlende Werte auftreten, dass also die Variable *X* einen kompletten Wertevektor besitzt. In dieser Situation besagt die MAR-Bedingung, dass die Wahrscheinlichkeit für einen fehlender *Y*-Wert zwar von der Variablen *X* abhängen darf, aber für einen festen *X*-Wert nicht von der Variablen *Y*. Die bedingte Irrelevanz der unbekannten *Y*-Ausprägung lässt sich mit bedingten Wahrscheinlichkeiten präziser formulieren:

$$\mathrm{P}(\{M_Y = 1\} \mid X, Y) = \mathrm{P}(\{M_Y = 1\} \mid X)$$

Unter dieser Voraussetzung ist für jede *X*-Ausprägung die bedingte Verteilung der fehlenden *Y*-Werte identisch mit der bedingten Verteilung der vorhandenen *Y*-Werte, so dass genügend Information über die fehlenden Werte vorliegt.

Im realistischeren Fall, dass beide Variablen fehlende Werte aufweisen, sind vier Muster fehlender Werte möglich, und die MAR-Bedingung verlangt für deren Wahrscheinlichkeiten (nach Little & Rubin 2002, S. 18):

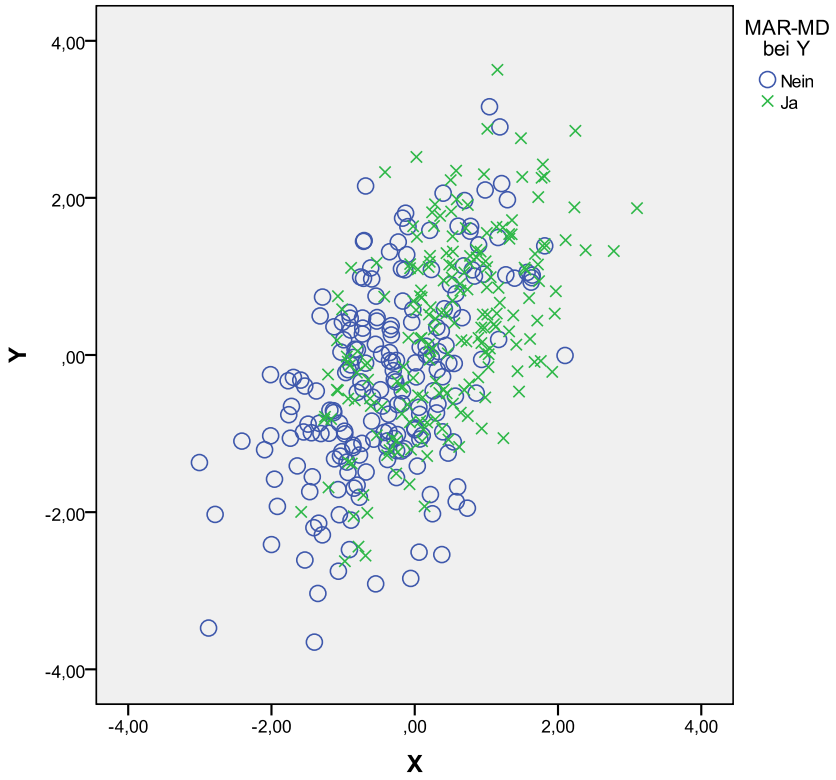
$$\mathrm{P}(\{M_X = 1, M_Y = 1\} \mid X, Y) = c \quad (\in [0, 1])$$
$$\mathrm{P}(\{M_X = 1, M_Y = 0\} \mid X, Y) = \mathrm{P}(\{M_X = 1, M_Y = 0\} \mid Y)$$
$$\mathrm{P}(\{M_X = 0, M_Y = 1\} \mid X, Y) = \mathrm{P}(\{M_X = 0, M_Y = 1\} \mid X)$$
$$\mathrm{P}(\{M_X = 0, M_Y = 0\} \mid X, Y) = 1 - c - \mathrm{P}(\{M_X = 1, M_Y = 0\} \mid Y) - \mathrm{P}(\{M_X = 0, M_Y = 1\} \mid X)$$

Es dürfte klar sein, wie *vollständige* Variablen in diese Gleichungen aufzunehmen sind. Je mehr Beobachtungen mit prognostischer Relevanz für die Wahrscheinlichkeiten fehlender Werte eingehen (im Idealfall über *vollständige* Variablen) eingehen, desto realistischer ist die MAR-Bedingung.

Im Wesentlichen verlangt die MAR-Bedingung für jedes MD-Muster (z.B. fehlende Werte bei den Variablen X_1 und X_2, vorhandene Werte bei den restlichen Variablen X_3 und X_4), dass bei jeder festen Kombination vorhandener Werte (im Beispiel: bei jedem (X_3, X_4)-Wertepaar) die Verteilung der MD-belasteten Variablen identisch ist bei den Teilpopulationen mit bzw. ohne Beobachtungswerte. Leider kann die MAR-Bedingung *nicht* überprüft werden, weil man dazu die fehlenden Werte kennen müsste.

Viele attraktive Techniken zur Lösungen von MD-Problemen setzen die MAR-Bedingung voraus (z.B. die ML-Techniken mit direkter Parameterschätzung oder EM-Schätzung von Verteilungsmomenten, die multiple Imputation). Diese Methoden liefern unverzerrte Parameterschätzungen, ohne dass ein Modell für das Zustandekommen fehlender Werte bekannt sein müsste. Man spricht daher auch von einem *ignorierbaren* Missing Data - Mechanismus, wenn zumindest die MAR-Bedingung erfüllt ist (siehe z.B. Allison 2002, S. 5).

Der folgende Plot zeigt eine vollständige Variable *X* und eine Variable *Y* mit fehlenden Werten (durch ein Kreuz markiert). Im MAR-Sinn hängt in der simulierten Population die Wahrscheinlichkeit für einen fehlenden *Y*-Wert von der *X*-Ausprägung ab, ist für feste *X*-Werte jedoch unabhängig von der *Y*-Ausprägung:



Die MCAR-Bedingung ist bei diesen Daten deutlich verletzt, was zu einem hoch signifikanten Little-Test führt:

EM-Kovarianzen[a]

	X	Y
X	,99378	
Y	,68346	1,61432

a. MCAR-Test nach Little: Chi-Quadrat = 67,823, DF = 1, Sig. = ,000

Die von SPSS per EM-Algorithmus (siehe Abschnitt 5.1) ermittelte und in der Tabelle protokollierte Kovarianz kann als sehr gute Schätzung für den wahren Wert (= 0,7) gelten.

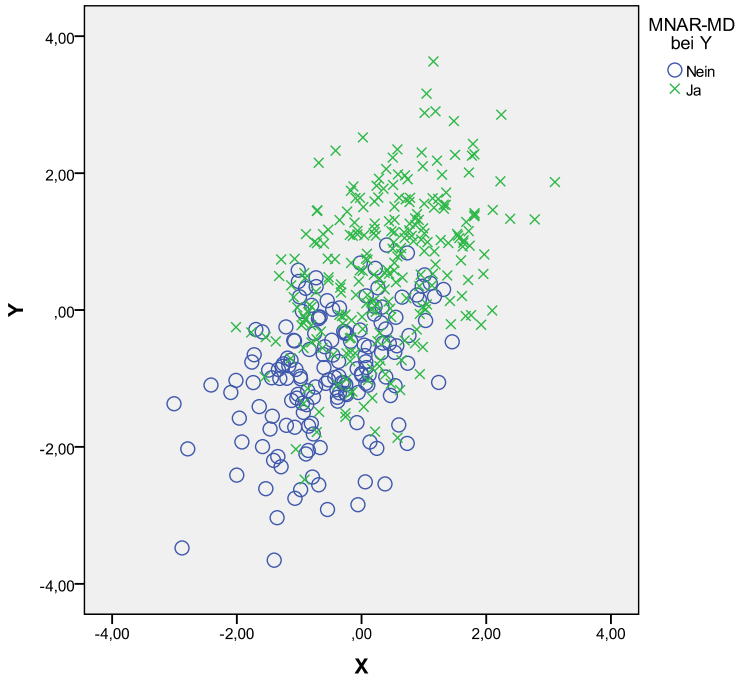
Um die MAR-Plausibilität zu steigern, sollten bei einer Studie möglichst viele Merkmale mit Einfluss auf die Wahrscheinlichkeit für fehlende Werte miterhoben werden. In sozialwissenschaftlichen Studien sind dabei z.B. folgende Merkmale von Interesse: Bildung, Alter, Geschlecht, Lebenszufriedenheit, Motivation zur Teilnahme an der Studie. Am Ende einer schriftlichen Befragung (z.B. via Internet) kann man sich mit einiger Aussicht auf eine ehrliche Antwort bei den Teilnehmern erkundigen, ob sie ernsthaft geantwortet haben.

Wenn fehlende Werte Bestandteil der Untersuchungsplanung sind, ist die MAR- oder auch die MCAR-Bedingung erfüllt (siehe Enders 2010, S. 21). Man kann z.B. aus Kostengründen bei manchen Fällen auf eine aufwändige Messung verzichten:

- Geschieht dies in Abhängigkeit von einer bestimmten Ausprägung bei einem erhobenen Merkmal, resultiert die MAR-Bedingung.
- Geschieht dies zufallsabhängig, ist die MCAR-Bedingung erfüllt.

2.3 MNAR

Ist bei einem Datensatz die MAR-Bedingung *nicht* erfüllt (und damit auch die MCAR-Bedingung nicht), spricht man von der *MNAR – Bedingung* (*Missing Not At Random*). Für mindestens eine Variable hängt die MD-Wahrscheinlichkeit auch nach Kontrolle aller Einflüsse von beobachteten Variablen immer noch vom fehlenden Wert ab. Bei der im folgenden Plot mit einer vollständigen Variablen *X* und einer unvollständigen Variablen *Y* dargestellten Konstellation hängt die MD-Wahrscheinlichkeit bei *Y* auch nach Berücksichtigung des vorhandenen *X*-Werts von der unbekannten *Y* - Ausprägung ab:



Im Little-Test für diese Daten wird die Verletzung der MCAR-Bedingung aufgedeckt:

EM-Kovarianzen[a]

	X	Y
X	,99378	
Y	,42813	,86053

a. MCAR-Test nach Little: Chi-Quadrat = 75,840, DF = 1, Sig. = ,000

Weil auch die (nicht testbare) MAR-Bedingung verletzt ist, gelingt per EM-Algorithmus (siehe Abschnitt 5.1) keine brauchbare Schätzung der Kovarianz (wahrer Populationswert: 0,7). Weil die Steigung der Regression von *Y* auf *X* durch den Quotienten aus der Kovarianz und der Varianz des Prädiktors zu bestimmen ist, resultiert aus den per EM-Algorithmus geschätzten Normalverteilungsparametern die erheblich verzerrte Schätzung von 0,43 (wahrer Populationswert: 0,7).

Durch die Einbeziehung von Hilfsvariablen, die als Ursachen bzw. Korrelate für das Fehlen von Werten in Frage kommen, lässt sich das MNAR-Risiko reduzieren. Über Hilfsvariablen, die mit MD-belasteten Variablen korreliert sind, lässt sich die Auswirkung der MNAR-Bedingung abmildern. Kann man z.B. in der eben simulierten Situation eine Hilfsvariable aus dem Hut zaubern, die mit *Y* zu 0,655 korreliert ist und einen kompletten Wertevektor besitzt, wird die Verzerrung bei der Schätzung der Kovarianz von *Y* und *X* gemildert:

EM-Kovarianzen[a]

	X	Y	YH
X	,99378		
Y	,47053	,91140	
YH	,32741	,52585	,88772

a. MCAR-Test nach Little: Chi-Quadrat = 104,311, DF = 2, Sig. = ,000

Gelegentlich wird statt MNAR auch die Abkürzung NMAR (*Not Missing At Random*) verwendet.

3 Analyse der Verteilung von fehlenden Werten

3.1 Anwendungsbeispiel

Als Anwendungsbeispiel betrachten wir im Kurs mehrfach in Anlehnung an Allison (2002, S. 21) eine Studie zum Ausbildungserfolg an 1302 amerikanischen Colleges im Jahr 1994. Die Daten stehen im Internet auf der folgenden Webseite zur Verfügung:

http://lib.stat.cmu.edu/datasets/colleges/

Es sind die folgenden Variablen beteiligt:

GRADRAT	Prozentsatz der erfolgreichen Absolventen: $\frac{\text{Anzahl der Graduierten}}{\text{Anzahl der Einsteiger vier Jahre zuvor}} \cdot 100$
CSAT	Mittlere kombinierte mathematische und verbale Leistung der College-Bewerber im SAT-Test
MSAT	Mittlere mathematische Leistung der College-Bewerber im SAT-Test
VSAT	Mittlere verbale Leistung der College-Bewerber im SAT-Test
ACT	Mittlere Leistung der College-Bewerber im ACT-Test
ENROLL	Anzahl der Einsteiger
LNENROLL	Logarithmierte Anzahl der Einsteiger
PRIVATE	Trägerschaft: 1 privat 0 öffentlich
STUFAC	Betreuungsverhältnis: $\frac{\text{Anzahl der Studierenden}}{\text{Anzahl der Lehrenden}}$
RMBRD	Jährliche Investitionen in die Ausstattung (hoffentlich relativiert an der Größe)
PCTTOP25	Prozentsatz der Studierenden aus dem Top-25 - Segment der High school

Es soll per linearer Regression untersucht werden, wie GRADRAT von den Prädiktoren CSAT, (LN)ENROLL, PRIVATE, STUFAC und RMBRD abhängt.

Die Variablen MSAT, VSAT, ACT und PCTTOP25 werden später als Hilfsvariablen zur Rekonstruktion fehlender Informationen einbezogen. Sie sind hoch bis sehr hoch korreliert mit den Modellvariablen CSAT und RMBRD, bei denen viele Werte fehlen (siehe unten).

In der Originaldatei sind für die Variable PRIVATE alle Werte vorhanden. Um auch fehlende Werte bei einer nominalskalierten Variablen betrachten zu können, wurden bei der Variablen PRIVATE ca. 20 % der Werte per Zufall gelöscht (MCAR!). Die so entstandene Datei **UsNews mit MCAR-MDs bei PRIVATE.sav** ist an der im Vorwort vereinbarten Stelle zu finden.

3.2 Muster- und MCAR-Analyse mit der Prozedur MVA

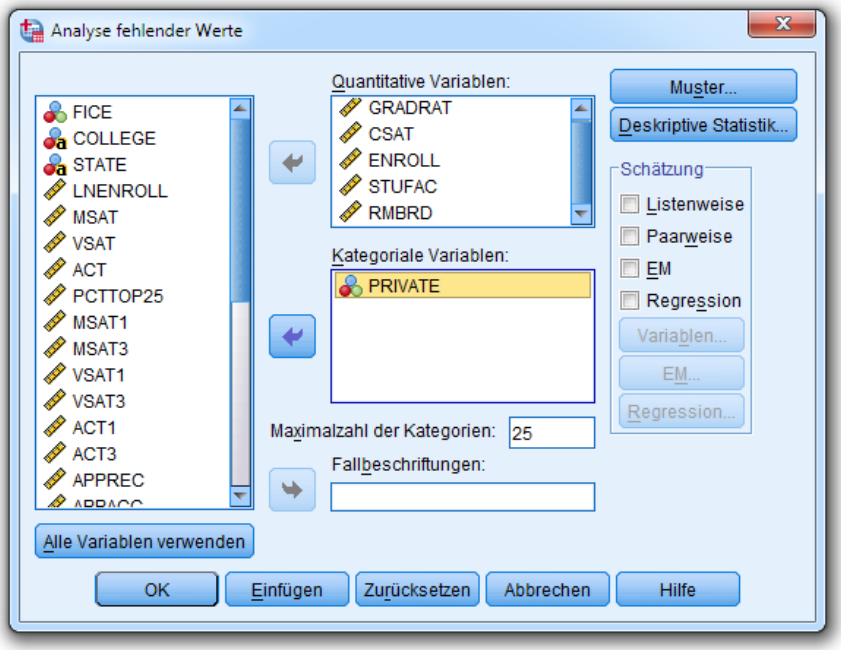
Mit der Prozedur MVA können z.B. folgende Fragestellungen bearbeitet werden:

- Ausmaß des MD-Problems
- Suche nach hauptverantwortlichen Variablen für niedrige Fallzahlen bei multivariaten Analysen
- Welche Variablen sollten normalisierend transformiert werden?
- Suche nach Prädiktoren für das Auftreten fehlender Werte
- Prüfung der MCAR-Bedingung

Wir fordern über den Menübefehl

Analysieren > Analyse fehlender Werte

eine Analyse für die Variablen GRADRAT, CSAT, ENROLL, PRIVATE, STUFAC und RMBRD an:



Vorläufig beschränken wir uns auf den voreingestellten Ausgabeumfang.

3.2.1 Variablen mit fehlenden oder extremen Werten

Beim voreingestellten Ausgabeumfang erscheint eine Tabelle, die u.a. für jede Variable den Anteil fehlender Werte zeigt:

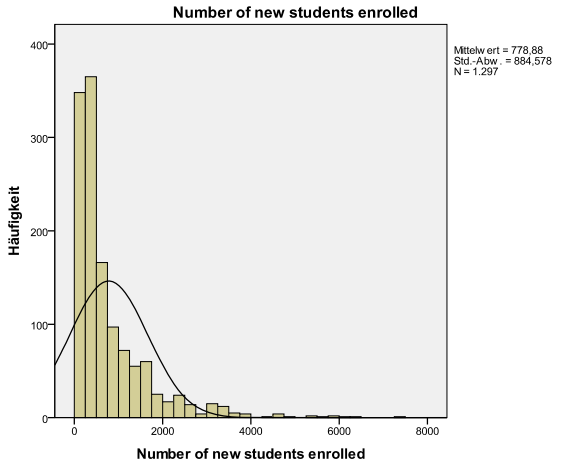
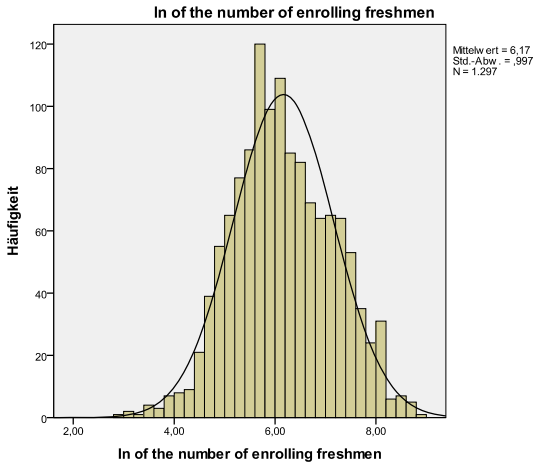
Univariate Statistiken

	N	Mittelwert	Standard-abweichung	Fehlend		Anzahl der Extremwerte[a]	
				Anzahl	Prozent	Niedrig	Hoch
GRADRAT	1204	60,41	18,889	98	7,5	0	1
CSAT	779	967,98	123,577	523	40,2	2	18
ENROLL	1297	778,88	884,578	5	,4	0	103
STUFAC	1300	14,859	5,1864	2	,2	3	19
RMBRD	783	4,1451	1,16959	519	39,9	0	8
PRIVATE	1058			244	18,7		

a. Anzahl der Fälle außerhalb des Bereichs (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

Im Colleges-Beispiel zeigen die Variablen CSAT und RMBRD die höchsten Quoten.

Die **Extremwerte** werden nach Tukey's Box-Kriterium ermittelt und können zur Beurteilung der Verteilungssymmetrie beitragen, die für alle Verfahren unter der Annahme multivariater Normalverteilung relevant ist (z.B. EM-Algorithmus, FIML-Schätzung). Bei der Variablen ENROLL, die zahlreiche extrem hohe Werte besitzt, wirkt sich eine logarithmische Transformation günstig aus:

Verteilung von ENROLL	Verteilung von ln(ENROLL)
	

Bei der logarithmierten Variante LNEROLL sind kaum noch Extremwerte festzustellen:

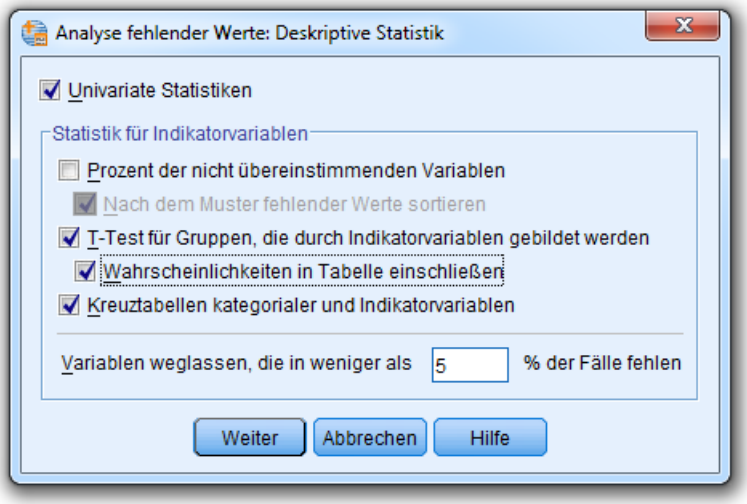
Univariate Statistiken

	N	Mittelwert	Standard-abweichung	Fehlend		Anzahl der Extremwerte[a]	
				Anzahl	Prozent	Niedrig	Hoch
GRADRAT	1204	60,41	18,889	98	7,5	0	1
CSAT	779	967,98	123,577	523	40,2	2	18
LNENROLL	1297	6,1675	,99715	5	,4	4	0
STUFAC	1300	14,859	5,1864	2	,2	3	19
RMBRD	783	4,1451	1,16959	519	39,9	0	8
PRIVATE	1058			244	18,7		

a. Anzahl der Fälle außerhalb des Bereichs (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

3.2.2 Lokale und globale Beurteilung der MCAR-Bedingung

Ist die MCAR-Bedingung erfüllt, müssen zu jeder Variablen die beiden Teilpopulationen mit einem vorhandenen bzw. fehlenden Wert bei jeder anderen Variablen dieselbe Verteilung besitzen. Um diese Bedingung zu prüfen, fordern wir im MVA-Subdialog **Deskriptive Statistik** zusätzlich **T-Tests für Gruppen, die durch Indikatorvariablen gebildet werden**, (inkl. Überschreitungswahrscheinlichkeiten) für metrische Zielvariablen sowie **Kreuztabellen** für kategoriale Variablen an:



Für metrische Variablen zeigen sich im Colleges-Stichprobe viele, teilweise erhebliche Mittelwertsunterschiede. Aus Platzgründen sind hier nur die Ergebnisse für die Variable GRADRAT zu sehen:

T-Tests bei unterschiedlicher Varianz[a]

		GRADRAT	CSAT	LNENROLL	STUFAC	RMBRD
GRADRAT	T	.	6,6	4,2	-1,4	4,3
	df	.	56,8	110,3	99,5	53,6
	P(2-seitig)	.	,000	,000	,153	,000
	Anzahl vorhanden	1204	732	1202	1204	735
	Anzahl fehlend	0	47	95	96	48
	Mittelwert (Vorhanden)	60,41	973,71	6,1994	14,762	4,1893
	Mittelwert (Fehlend)	.	878,68	5,7635	16,077	3,4686

Für jede quantitative Variable werden Gruppenpaare durch Indikatorvariablen gebildet (vorhanden, fehlend).

a. Indikatorvariablen mit weniger als 5% fehlend werden nicht angezeigt.

Wir erhalten z.B. für die Zielvariable CSAT folgende Ergebnisse:

- Von den 1204 Fällen mit einem gültigen GRADRAT-Wert besitzen 732 Fälle auch einen gültigen CSAT-Wert, wobei sich ein Mittelwert von 973,71 ergibt.
- Von den 98 Fällen mit einem fehlenden GRADRAT-Wert besitzen 47 Fälle einen gültigen CSAT-Wert, wobei sich der erheblich niedrigere Mittelwert 878,68 ergibt.
- Daraus ergibt sich der hoch signifikante t-Wert 6,6 ($p < 0{,}001$ bei $df = 56{,}8$).

Offenbar ist die Wahrscheinlichkeit für einen fehlenden GRADRAT-Wert umso höher, je schlechter die Bewerber einer Universität im CSAT-Test abschneiden, so dass eine Verletzung der MCAR-Bedingung anzunehmen ist.

Wer nachvollziehen möchte, wie das t-Test - Ergebnis zustande gekommnen ist, kann so vorgehen:

- MD-Indikator zur Variablen GRADRAT erstellen, z.B. mit der Syntax

```
COMPUTE GradRatMd = MISSING(GRADRAT).
```

- Einen t-Test für unabhängige Stichproben durchführen mit dem Indikator als Gruppen- und CSAT als Testvariable

Es resultieren die bereits bekannten Ergebnisse:

Gruppenstatistiken

	GradRatMd	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
CSAT	,00	732	973,71	123,121	4,551
	1,00	47	878,68	93,394	13,623

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
CSAT	Varianzen sind gleich	5,144	,024	5,195	777	,000	95,031	18,292	59,123	130,939
	Varianzen sind nicht gleich			6,616	56,794	,000	95,031	14,363	66,267	123,794

Die **Kreuztabellen von kategorialen und Indikatorvariablen** zeigen im Colleges-Beispiel für die beiden PRIVATE-Kategorien und erwartungsgemäß auch für die (nach dem MCAR-Prinzip künstlich verursachte) MD-Kategorie nahezu identische Anteile fehlender GRADRAT-Werte (8,3%, 7,9% und 5,3%):

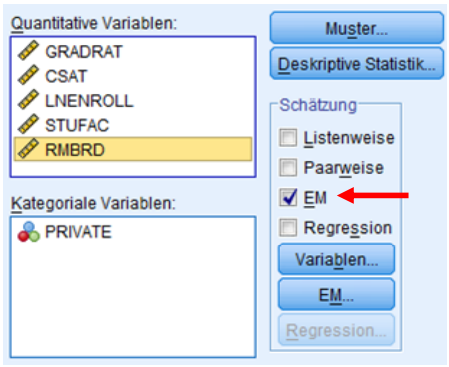
PRIVATE

			Total	public	private	Fehlend SysMis
GRADRAT	Vorhanden	Anzahl	1204	341	632	231
		Prozent	92,5	91,7	92,1	94,7
	Fehlend	% SysMis	7,5	8,3	7,9	5,3
CSAT	Vorhanden	Anzahl	779	205	437	137
		Prozent	59,8	55,1	63,7	56,1
	Fehlend	% SysMis	40,2	44,9	36,3	43,9
RMBRD	Vorhanden	Anzahl	783	205	437	141
		Prozent	60,1	55,1	63,7	57,8
	Fehlend	% SysMis	39,9	44,9	36,3	42,2

Indikatorvariablen mit weniger als 5% fehlend werden nicht angezeigt.

Bei den Variablen CSAT und RMBRD zeigen die öffentlichen Schulen hingegen eine deutliche höhere Rate fehlender Werte als die privaten.

Zur *globalen* Beurteilung der MCAR-Bedingung berechnet die MVA-Prozedur den Test nach Little, wenn man das Schätzen von Verteilungsparametern per EM-Algorithmus anfordert:



Das Testergebnis erscheint als Fußnote zu jeder Tabelle mit EM-Schätzergebnisssen, z.B.:

Geschätzte Randmittel[a]

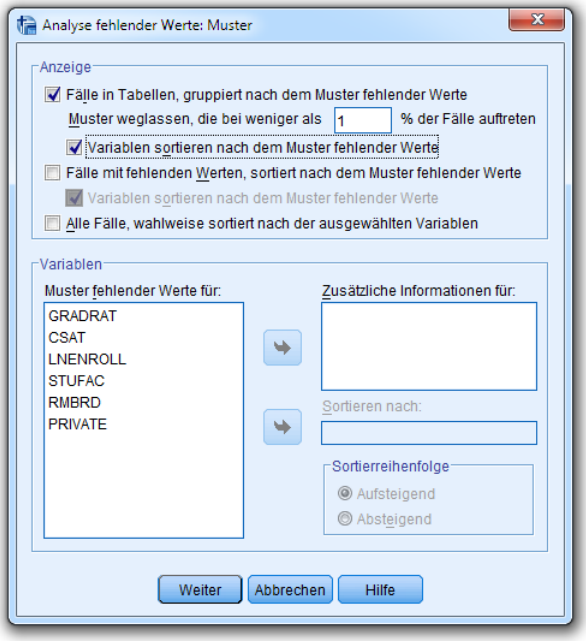
GRADRAT	CSAT	LNENROLL	STUFAC	RMBRD
59,97	959,10	6,1620	14,855	4,0679

a. MCAR-Test nach Little: Chi-Quadrat = 142,487, DF = 33, Sig. = ,000

Wie es aufgrund der zahlreichen signifikanten t-Tests zu erwarten war, verwirft der Test verwirft seine Nullhypothese. Damit sind MCAR-abhängige Verfahren zur Behandlung fehlender Werte (z.B. listenweiser oder paarweiser Ausschluss) unzulässig. Wir hoffen, dass die nicht prüfbare MAR-Bedingung annähernd erfüllt ist, so dass die Maximum-Likelihood - Verfahren und die multiple Imputation zulässig sind.

3.2.3 Muster fehlender Werte

Im MVA-Subdialog **Muster** kann man u.a. eine Tabelle mit Mustern fehlender Werte anfordern:



Die Ergebnistabelle zeigt per Voreinstellung alle Muster, die mindestens ein Prozent der Fälle enthalten:

Muster in Tabellen

Anzahl der Fälle	Muster fehlender Werte[a]						Vollständig, wenn ... [b]
	STUFAC	LNENROLL	GRADRAT	PRIVATE	CSAT	RMBRD	
372							372
20			X				392
26			X		X		641
223					X		595
56				X	X		734
83				X			455
46				X		X	732
231						X	603
145					X	X	971
46				X	X	X	1202
18			X		X	X	1053
18			X			X	641

Muster mit weniger als 1% Fällen (13 oder weniger) werden nicht angezeigt.

a. Variablen sind nach Mustern fehlender Werte sortiert.
b. Anzahl der vollständigen Fälle, wenn die in diesem Muster fehlenden Variablen (mit X gekennzeichnet) nicht verwendet werden.

Es ist z.B. zu erfahren, dass bei 372 Fällen alle Variablen vorhanden sind, und dass bei 223 Fällen ausschließlich der CSAT-Wert fehlt.

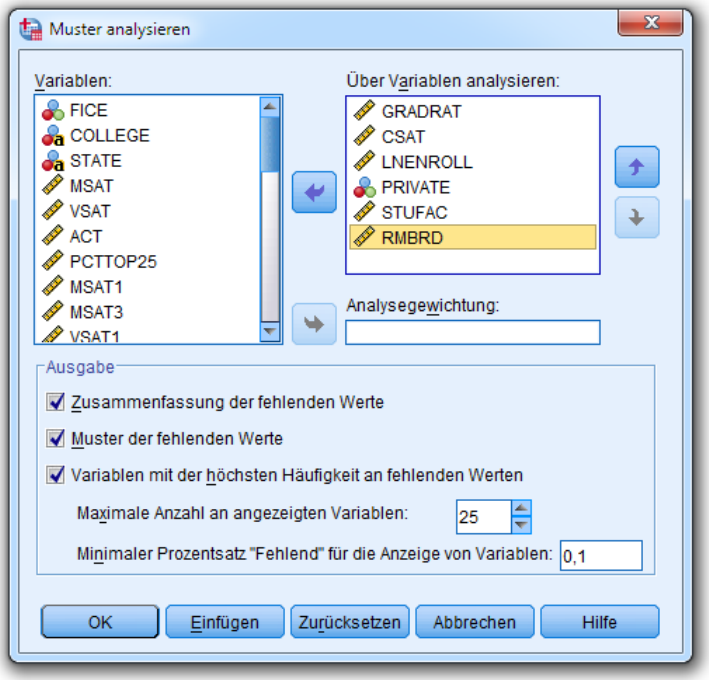
Über die Verwendung des MVA-Moduls zur Berechnung von Verteilungsmomenten (Mittelwerten, Varianzen und Kovarianzen) per EM-Algorithmus und zur Imputation (Ersetzung) fehlender Daten wird später berichtet.

3.3 Musteranalyse mit der Prozedur MULTIPLE IMPUTATION

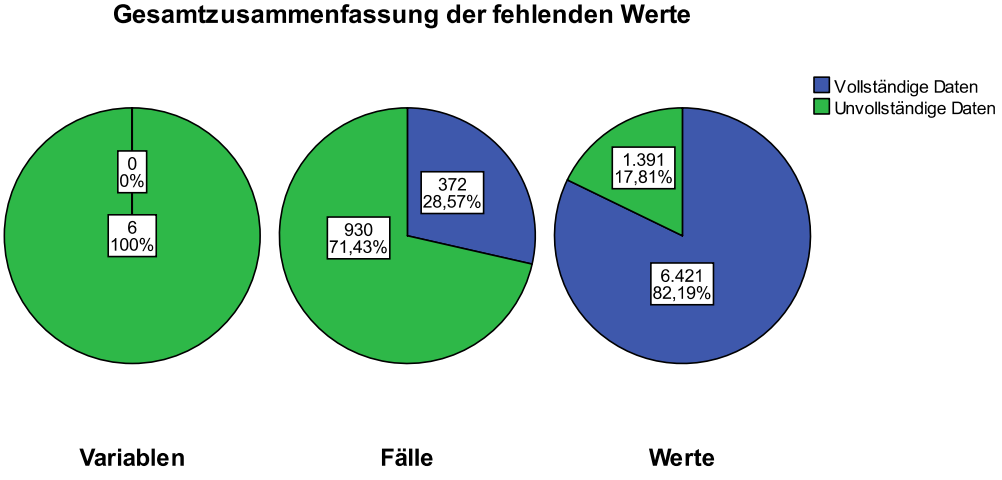
Die Prozedur zur multiplen Imputation bietet über den Menübefehl

Analysieren > Multiple Imputation > Muster analysieren

einige Ausgaben zum Ausmaß und zu den Mustern fehlender Werte. Wir lassen im Colleges-Beispiel (vgl. Abschnitt 3.1) die Variablen des geplanten Regressionsmodells analysieren und ändern im Vergleich zur Voreinstellung nur das Kriterium zur Berücksichtigung von MD-belasteten Variablen ab (**minimaler Prozentsatz** 0,1 statt 10):



Zur zusammenfassenden Beschreibung der MD-Problematik erhalten wir über drei Kreisdiagramme



folgende Ergebnisse:

- Die sechs betrachteten Variablen sind ohne Ausnahme mit fehlenden Werten belastet.
- Von den 1302 Fällen sind nur 372 komplett.
- Von den insgesamt beteiligten 7812 Werten (1302 · 6) fehlen 1391 Werte (17,81%).

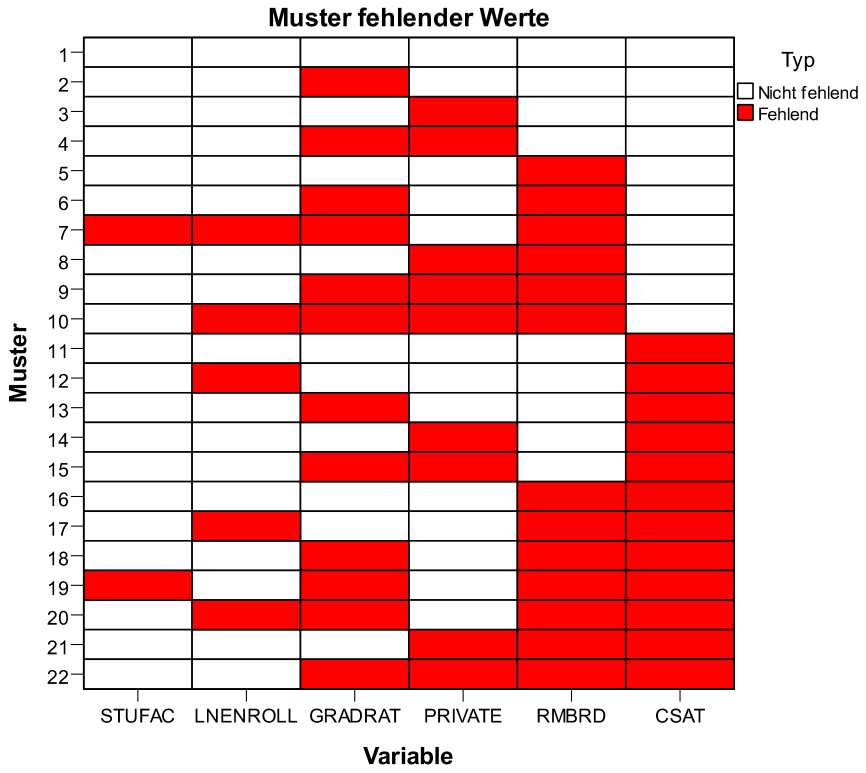
Weil wir das Kriterium für den zu berücksichtigenden MD-Belastungsgrad gesenkt haben, berichtet die folgende Tabelle über alle Variablen u.a. die absolute und die relative Häufigkeit fehlender Werte:

Variablenzusammenfassung

	Fehlend				
	N	Prozent	Gültige N	Mittelwert	Standard-abweichung
CSAT	523	40,2%	779	967,98	123,577
RMBRD	519	39,9%	783	4,1451	1,16959
PRIVATE	244	18,7%	1058		
GRADRAT	98	7,5%	1204	60,41	18,889
ENROLL	5	,4%	1297	778,88	884,578
STUFAC	2	,2%	1300	14,859	5,1864

Dabei sind die Variablen absteigend nach der Anzahl fehlender Werte geordnet.

Im folgenden Diagramm sind die **Muster fehlender Werte** dargestellt:



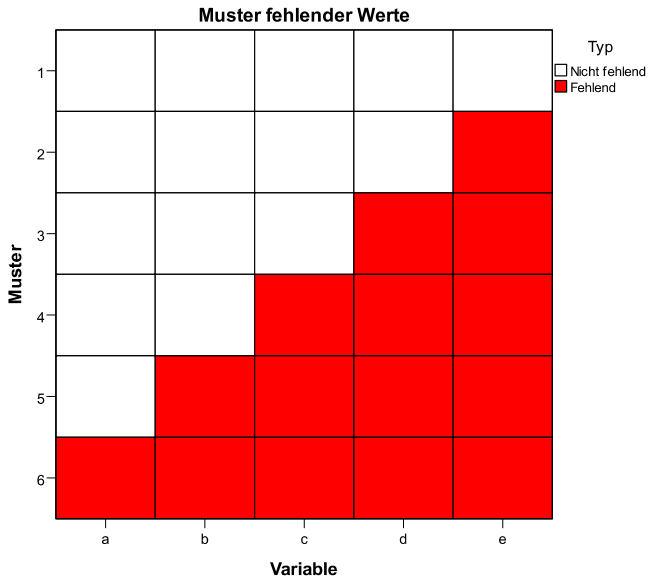
In den Spalten stehen die nach dem Anteil fehlender Werte aufsteigend geordneten Variablen. Am linken Rand erscheint Variable mit dem kleinsten Anteil fehlender Werte (STUFAC) und am rechten Rand die Variable mit dem größten Anteil fehlender Werte (CSAT).

In den Zeilen stehen die Muster fehlender Werte, die in der Stichprobe aufgetreten sind, mit folgender Sortierung:

- Erstes Sortierkriterium ist die Variable mit dem größten Anteil fehlender Werte, wobei die Muster mit vorhandenem Wert vor den Mustern mit fehlendem Wert erscheinen.
- Zweites Sortierkriterium ist die Variable mit dem zweitgrößten Anteil fehlender Werte, wobei die Muster mit vorhandenem Wert vor den Mustern mit fehlendem Wert erscheinen.
- usw.

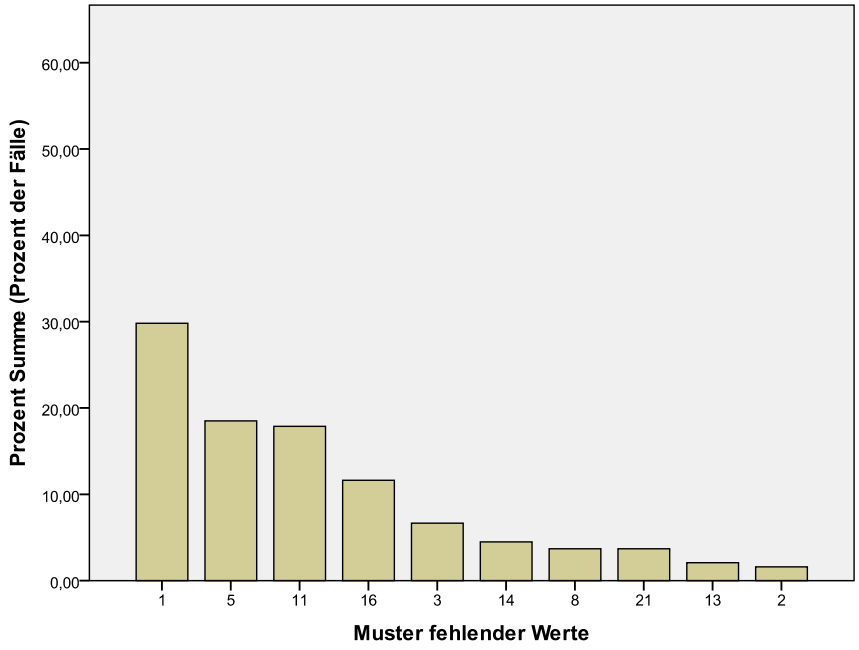
Nach diesem Schema steht in der ersten Zeile das MD-Muster *ohne* fehlende Werte.

Wenn für die MD-Muster eine **monotone** Ordnung besteht, d.h. wenn Fälle mit einem fehlenden Wert bei der Variablen *i* auch bei allen Variablen *j* mit einem größeren Anteil fehlender Werte keinen gültigen Wert besitzen, dann zeigt sich ein Bild wie im folgenden Beispiel (für einen künstlichen Datensatz):



Mit dieser Konstellation, die von SPSS bei der multiplen Imputation durch einen speziellen Algorithmus unterstützt wird, ist vor allem in Längsschnittstudien zu rechnen. Im Colleges-Beispiel liegt keine Monotonie vor, und im Manuskript wird dieser Spezialfall *nicht* behandelt.

Schließlich erhält man noch ein Balkendiagramm mit den zehn häufigsten MD-Mustern, wobei für die Colleges-Studie dieses Ergebnis resultiert:



Die 10 am häufigsten auftretenden Muster werden im Diagramm dargestellt.

Dass nur für 28,57% der Fälle *alle* Werte vorhanden sind, wissen wir bereits.

4 Traditionelle Methoden zur Behandlung fehlender Werte

4.1 Individuelle Mittelwerte aus den vorhandenen Items

Bei der Schätzung einer latenten Variablen durch den Mittelwert aus *k* manifesten Indikatoren (Items) ist es oft vertretbar, einige fehlende Items zu tolerieren (z.B. 10%) und den Mittelwert aus den bei einem Fall vorhandenen Items zu berechnen. Streng genommen müssen dazu die Items *austauschbar* sein, also einem einfaktoriellen Modell genügen, sowie identische Mittelwerte, Ladungen und Fehlervarianzen aufweisen. Wenn sich bei eindimensionalen Items die Mittelwerten und/oder Varianzen deutlich unterscheiden, lässt sich die Austauschbarkeit durch Standardisieren der Items verbessern.

Schafer & Graham (2002, S. 157f) bezeichnen die beschriebene Technik als *ipsative Mittelwerts-Imputation* und halten sie für akzeptabel bei eindimensionalen Items mit einer relativ hohen Reliabilität (Cronbachs $\alpha > 0{,}7$).

Generell bevorzugen die Autoren allerdings die multiple Imputation fehlender Itemwerte (siehe unten). Bei einer Analyse von latenten Variablen mit Amos sind dank FIML-Schätzmethode (*Full Information Maximum Likelihood*, siehe unten) fehlende Einzelitems kein Problem.

SPSS Statistics unterstützt die Berechnung einer Mittelwertsvariablen aus den individuell vorhandenen Items durch die Funktion MEAN, wobei optional hinter dem Funktionsnamen eine Mindestzahl von Argumenten verlangt werden kann, z.B.

Beispiel:
```
compute LZ = mean.8(lz1 to lz10).
```
Wenn für einen Fall bei den Variablen LZ1 bis LZ10, die in der Arbeitsdatei hintereinander stehen, mindestens 8 valide Werte vorliegen, wird deren arithmetisches Mittel der Variablen LZ zugewiesen, ansonsten erhält die Zielvariable den MD-Indikator SYSMIS.

4.2 Ausschluss von Variablen

Sind für das MD-Problem wenige Variablen verantwortlich, kann man das Problem zusammen mit diesen Variablen beseitigen, wenn die Variablen wenig relevant oder durch äquivalente Variablen mit annähernd vollzähligen Werten zu ersetzen sind. Die SPSS-Prozedur MVA informiert über den Anteil fehlender Werte bei den Variablen (siehe Abschnitt 3.2.1).

4.3 Ausschluss von Fällen

Bei der fallweisen Behandlung fehlender Werte werden bei einer Analyse nur Fälle mit gültigen Werten für *alle* beteiligten Variablen berücksichtigt. Fehlt z.B. bei einer multiplen Regressionsanalyse bei einem Fall eine einzige Prädiktorausprägung, wird der komplette Fall ausgeschlossen.

4.3.1 Nachteile des Verfahrens

Diese bei SPSS und vielen anderen vielen Statistikprogrammen voreingestellte Methode hat folgende Nachteile:

- **Potentielle verzerrte Schätzer**
  Bei verletzter MCAR-Bedingung sind verzerrte Schätzer zu befürchten. Bei speziellen Modellen führt die fallweise Behandlung jedoch auch bei MAR- oder gar MNAR-Verhältnissen zu konsistenten (asymptotisch erwartungstreuen) Schätzern (siehe unten).

- **Unvollständige Nutzung der verfügbaren Informationen**
  Die fallweise Behandlung kann auch bei unverzerrten Schätzern (in der MCAR-Situation) inakzeptabel sein, wenn der Informationsverlust zu groß wird. In sozialwissenschaftlichen Studien gehen je nach Anzahl der beteiligten Variablen oft 20% - 50% der Fälle verloren (Acock 2005, S. 1015). Ob die Folgen (vergrößerte Vertrauensintervalle, reduzierte Power der Hypothesentests) zu verschmerzen ist, hängt von der verbliebenen Stichprobengröße und der Effektstärke ab.

Hängt z.B. bei der linearen Regression die MD-Wahrscheinlichkeit eines Prädiktors vom Kriterium ab, dann führt die fallweise Behandlung fehlender Werte zu verzerrten Regressionskoeffizienten. Im folgenden Simulationsbeispiel wird das Kriterium *Y* von den Regressoren *X* und *Z* beeinflusst:[1]

```
compute K = normal(1).
compute X = 0.5*K + normal(1).
compute Z = 0.5*K + normal(1).
compute Y = 0.7*X + 0.7*Z + normal(1).
```

Es wird MCAR-konträr dafür gesorgt, dass die Wahrscheinlichkeit für einen fehlenden *X*-Wert mit zunehmendem Kriteriumswert wächst:

```
compute XM = X.
do if uniform(1) < exp(Y) / (1 + exp(Y)).
  recode XM (lo thru hi = SYSMIS).
end if.
```

Es resultieren unterschätzte Regressionskoeffizienten für *X* und *Z*:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-,416	,027		-15,581	,000	-,468	-,364
	XM	,611	,023	,489	26,453	,000	,566	,656
	Z	,606	,024	,475	25,699	,000	,560	,652

a. Abhängige Variable: Y

In Abhängigkeit von der Struktur der MCAR-Verletzung kann es auch zu überhöhten Schätzungen kommen.

Wie aufgrund der erfüllten MAR-Bedingung zu erwarten, liefert die in Abschnitt 5.1 vorzustellende Regressionsanalyse unter Verwendung der per EM-Algorithmus geschätzten Momente (Korrelationen, Standardabweichungen, Mittelwerte) brauchbare Schätzer für die Regressionskoeffizienten:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,019	,018		1,065	,287	-,016	,055
	XM	,716	,017	,507	43,037	,000	,683	,748
	Z	,707	,017	,495	41,980	,000	,674	,740

a. Abhängige Variable: Y

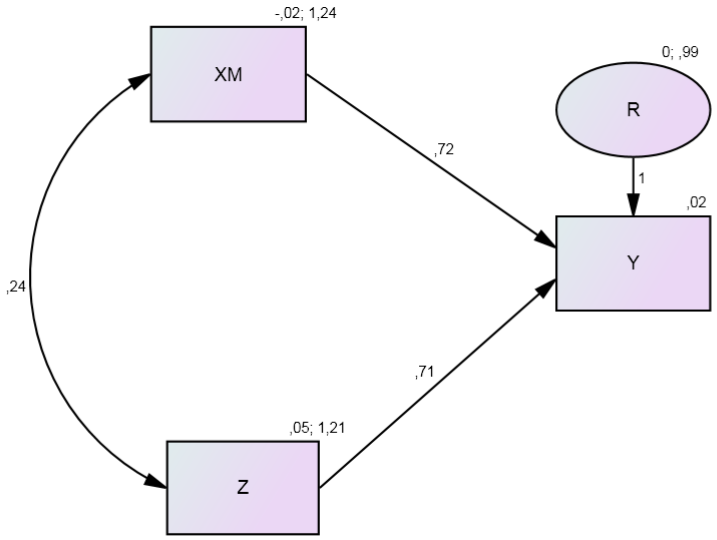
[1] Ein SPSS-Programm, Daten und ein Amos-Projekt zum Simulationsbeispiel sind im Ordner **MC-Ergebnisse bei MAR** an der im Vorwort vereinbarten Stelle zu finden. Es wird eine große Stichprobe simuliert ($N = 3000$), um die Effekte verschiedener Methoden zur Behandlung fehlender Werte mit geringer Stichprobenabhängigkeit beobachten zu können. Selbstverständlich treten die Effektmuster auch in kleineren Stichproben auf, dann aber mit erheblichen Unterschieden zwischen verschiedenen Stichproben.

Praktisch identische Regressionsschätzer liefert IBM SPSS Amos durch Verwendung der in Abschnitt 5.3 vorzustellenden, ebenfalls für MAR-Daten geeigneten FIML - Methode (*Full Information Maximum Likelihood*):

	Estimate	S.E.	C.R.	P	Label
Y <--- XM	,716	,021	34,444	***	
Y <--- Z	,706	,020	35,287	***	

Im Vergleich zum EM-Ergebnis sind allerdings die FIML-Standardfehler zu den Regressionskoeffizienten größer und vertrauenswürdiger, was später noch erläutert wird.

In Amos erfolgt die Modellspezifikation über ein Pfaddiagramm, das sich gut zur Illustration des untersuchten Simulationsmodells eignet:



Auch die kombinierten Regressionsergebnisse aufgrund einer multiplen Imputation mit 20 vervollständigten Datensätzen (siehe Abschnitt 6) liegen nahe bei den Populationsparametern:

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B		Anteil fehlende Info.	Relative Zunahmevarianz	Relative Effizienz
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze			
1	(Konstante)	,017	,024	,730	,467	-,030	,065	,428	,717	,979
	Z	,707	,022	32,206	,000	,664	,751	,424	,705	,979
	XM	,714	,022	33,044	,000	,671	,756	,422	,700	,979

a. Abhängige Variable: Y

Auch dieses Verfahren arbeitet bei erfüllter MAR-Bedingung sehr zuverlässig. Zudem stimmen die Standardfehler aus der FIML- und der MI-Analyse gut überein.

4.3.2 Vorteile des Verfahrens

Als *Vorzüge* der fallweisen Behandlung fehlender Werte sind zu nennen:

- Einfache Anwendung
- Erwartungstreue Schätzer und korrekte Inferenzstatistik in der MCAR-Situation

- Bei der linearen Regressionsanalyse bewährt sich der fallweise Ausschluss oft auch *ohne* MCAR-Voraussetzung:
  - Wenn ausschließlich Kriteriumswerte fehlen und dabei die MAR-Bedingung erfüllt ist, erhält man unverzerrte und effiziente Schätzer (Allison 2002, S. 54; Schafer & Graham 2002, S. 155) sowie eine korrekte Inferenzstatistik. In dieser Situation ist die fallweise Behandlung also die optimale Methode, solange man nur die Variablen des Analysemodells betrachtet. Existieren allerdings Hilfsvariablen mit Informationen über die fehlenden Werte, kommen trotzdem Imputations - oder Maximum Likelihood - Verfahren in Betracht (siehe Abschnitte 5 und 6).
  - Bei fehlenden Regressorwerten ist der fallweise Ausschluss sogar unempfindlich gegenüber Verletzungen der MAR-Bedingung, solange die MD-Wahrscheinlichkeiten der Regressoren nicht vom *Kriterium* abhängen (Allison 2002, S. 6f). In dieser Situation ist der fallweise Ausschluss den später vorzustellenden modernen Methoden (FIML, multiple Imputation) überlegen, welche die MAR-Bedingung voraussetzen.

Um die eben genannte Überlegenheit der fallweisen Behandlung in einer speziellen MNAR-Situation zu demonstrieren, ändern wir in der oben beschriebenen Simulationsstudie den MD-Prozess. Nun hängt die MD-Wahrscheinlichkeit beim Regressor *X* von seiner Ausprägung ab:[1]

```
do if X <= 0.5.
  recode X (lo thru hi = SYSMIS).
end if.
```

Als Ursache für die Schwächen der fallweisen Behandlung wird oft die mangelnde **Repräsentativität** der Reststichprobe genannt. Man benötigt die Repräsentativität selbstverständlich bei der Schätzung univariater Verteilungsaspekte (z.B. Erwartungswert). Bei einer Regressionsanalyse werden aber z.B. *keine* repräsentativen Regressorwerte benötigt. Auch im konkreten Beispiel sind die verbliebenen Fälle (mit einer *X*-Ausprägung größer als 0,5) nicht repräsentativ. Trotzdem erhalten wir brauchbare Schätzwerte:

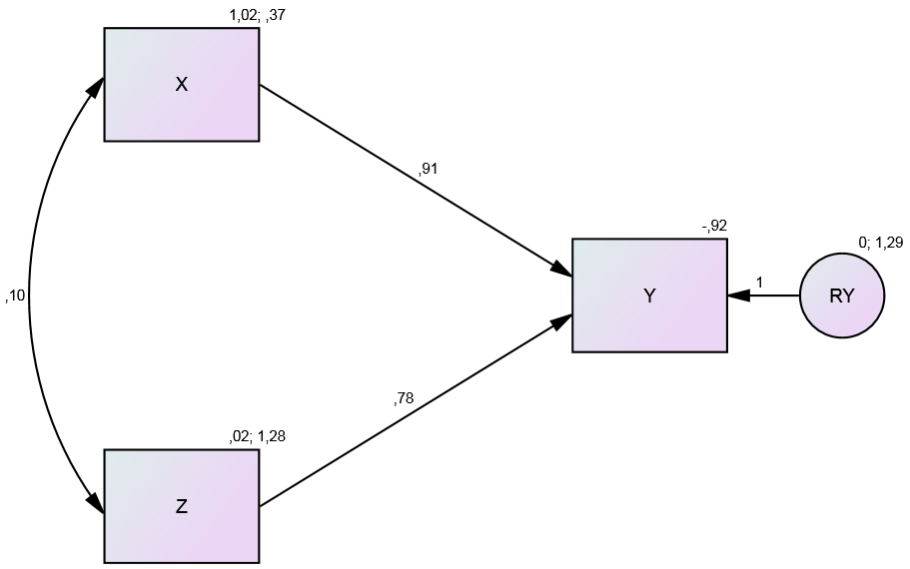
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,072	,073		,988	,323	-,071	,216
	X	,687	,054	,300	12,684	,000	,581	,794
	Z	,693	,028	,585	24,764	,000	,638	,747

a. Abhängige Variable: Y

Demgegenüber reagiert die von Amos ausgeführte FIML - Schätzung sensibel auf die Verletzung der MAR-Bedingung und liefert inflationierte Schätzer (speziell beim Regressionskoeffizienten zu *X*):

[1] Ein SPSS-Programm, Daten und ein Amos-Projekt zum Simulationsbeispiel sind im Ordner **MC-Ergebnisse bei MNAR** an der im Vorwort vereinbarten Stelle zu finden.



Unabhängig von den bisherigen Überlegungen zur fallweisen Behandlung fehlender Werte sollte man sich von Fällen mit einem zu hohen Anteil fehlender Werte tatsächlich trennen, weil hier der Verdacht mangelnder Datenqualität besteht. Als Entscheidungshilfe kann die SPSS-Prozedur MVA für jeden Fall die Zahl, den Anteil und das Muster seiner fehlenden Werte liefern (siehe Abschnitt 3.2).

4.4 Paarweiser Ausschluss fehlender Werte

Bei der paarweisen Behandlung fehlender Werte nutzt man zum Schätzen von Verteilungsparametern (z.B. Mittelwerten, Varianzen, Korrelationen) alle Fälle mit Werten bei den jeweils beteiligten Variablen. Folglich basieren die einzelnen Schätzungen (z.B. in einer Korrelationsmatrix) im Allgemeinen auf unterschiedlichen Teilstichproben.

Dieses Vorgehen nutzt alle verfügbaren Daten (im Unterschied zur fallweisen Behandlung) und liefert immerhin in der MCAR-Situation erwartungstreue Schätzer.

Etwas unklar ist jedoch die anzunehmende Stichprobengröße:

- Nimmt man eine komplette Datenbasis an, wird die Breite der Konfidenzintervalle unterschätzt, und die Hypothesentests sind zu liberal.
- Legt man (wie die REGRESSION-Prozedur in SPSS) das kleinste bivariate *N* zugrunde, wird die Breite der Konfidenzintervalle überschätzt, und die Hypothesentests sind zu konservativ.

4.4.1 Verzerrte Schätzer bei verletzter MCAR-Bedingung

Bei verletzter MCAR-Bedingung resultieren verzerrte Schätzer für Kovarianzen/Korrelationen sowie darauf basierende Parameter (z.B. Regressionskoeffizienten). Bei den schon in Abschnitt 4.3 vorgestellten Simulationsdaten mit dem Kriterium *Y* und den Prädiktoren *X* und *Z* wächst MCAR-konträr die Wahrscheinlichkeit für einen fehlenden *X*-Wert mit zunehmendem Kriteriumswert. Während die fallweise Behandlung zu deutlich unterschätzten Koeffizienten für beide Regressoren geführt hat, liefert die paarweise Behandlung einen inflationierten Wert für *Z*:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,293	,029		10,197	,000	,236	,349
	XM	,721	,026	,482	28,209	,000	,670	,771
	Z	,780	,024	,546	31,928	,000	,732	,827

a. Abhängige Variable: Y

4.4.2 Indefinite Korrelationsmatrizen

In extremen Fällen sind die aus unterschiedlichen (nur teilweise überlappenden) Teilstichproben stammenden Korrelationen so inkonsistent, dass eine *indefinite* Korrelationsmatrix (mit mindestens einem negativen Eigenwert) resultiert, von der z.B. bei einer Regressionsanalyse unsinnige Ergebnisse zu erwarten sind. Dies kann bei sehr kleinen Stichproben zwar auch in der MCAR-Situation passieren, ist jedoch in der MAR- oder MNAR- Bedingung erheblich wahrscheinlicher. Wir erzeugen in einer neuen Simulationsstudie die vollständigen Daten von 100 Fällen durch die folgenden Anweisungen:[1]

```
compute K = normal(1).
compute X = K + normal(1).
compute Z = K + normal(1).
compute Y = 0.7*X + 0.7*Z + normal(1).
```

Es resultiert die folgende Korrelationsmatrix:

Korrelationen

		X	Z	Y
X	Korrelation nach Pearson	1	,437	,760
	Signifikanz (2-seitig)		,000	,000
	N	100	100	100
Z	Korrelation nach Pearson	,437	1	,769
	Signifikanz (2-seitig)	,000		,000
	N	100	100	100
Y	Korrelation nach Pearson	,760	,769	1
	Signifikanz (2-seitig)	,000	,000	
	N	100	100	100

Durch heftige Datenverluste, die bei *X* in Abhängigkeit von der *Z*-Ausprägung erfolgen (MAR), bei *Z* und *Y* hingegen MCAR-konform,

```
compute XM = X.
compute ZM = Z.
compute YM = Y.
do if (Z > -0.5).
  recode  XM (lo thru hi = SYSMIS).
end if.
do if (Uniform(1) > 0.5).
  recode  ZM (lo thru hi = SYSMIS).
end if.
```

[1] Ein SPSS-Programm und die Daten zum Simulationsbeispiel sind im Ordner **Indefinite Korrelationsmatrix** an der im Vorwort vereinbarten Stelle zu finden.

```
do if (Uniform(1) > 0.5).
  recode  YM (lo thru hi = SYSMIS).
end if.
```

entsteht die folgende Korrelationsmatrix:

Korrelationen

		XM	ZM	YM
XM	Korrelation nach Pearson	1	-,155	,781
	Signifikanz (2-seitig)		,502	,000
	N	36	21	20
ZM	Korrelation nach Pearson	-,155	1	,824
	Signifikanz (2-seitig)	,502		,000
	N	21	48	24
YM	Korrelation nach Pearson	,781	,824	1
	Signifikanz (2-seitig)	,000	,000	
	N	20	24	50

Während die Korrelationsmatrix aus 100 vollständigen Fällen drei positive Eigenwerte

2,322 0,563 0,115

besitzt, hat die mit paarweisem Ausschluss fehlender Werte erzeugte Matrix einen negativen Eigenwert und ist damit indefinit:

2,061 1,155 -0,215

Ursache ist die negative Korrelation zwischen den Variablen *X* und *Z*, die im Widerspruch zu den Korrelationen

$r_{xy} = 0{,}781$
$r_{zy} = 0{,}824$

steht. Aus den Korrelationen von *X* und *Z* mit der Drittvariablen *Y* folgen eine obere und eine untere Schranke für r_{xz}:[1]

$$r_{xz} \in [r_{xy}r_{zy} - \sqrt{(1-r_{xy}^2)(1-r_{zy}^2)}; r_{xy}r_{zy} + \sqrt{(1-r_{xy}^2)(1-r_{zy}^2)}]$$

Bei unserer Defektmatrix hat r_{xz} (= -0,155) den zulässigen Wertebereich verlassen:

[0,781 0,824 - 0,354; 0,781 0,824 + 0,354] = [0,290; 0,997]

[1] Die Herleitung der Grenzen macht etwas Mühe und ist für den weiteren Kursverlauf nicht relevant: *X* und *Z* kann man als Summe aus der besten Vorhersage durch *Y* und dem zugehörigen Residuum schreiben. Haben alle Variablen den Mittelwert Null und die Varianz Eins, dann gilt:

$$X = r_{XY}Y + R_X$$
$$Z = r_{ZY}Y + R_Z$$

Wegen der Standardisierung von *X* und *Z* ist ihre Korrelation gleich der Kovarianz:

$$r_{XZ} = \mathrm{Cov}(r_{XY}Y + R_X, r_{ZY}Y + R_Z) = r_{XY}r_{ZY} + \mathrm{Cov}(R_X, R_Z)$$

Die Korrelation von R_X und R_Z kann maximal den Betrag Eins erreichen:

$$1 \geq |r_{R_X R_Z}| = \frac{|\mathrm{Cov}(R_X, R_Z)|}{\sqrt{\mathrm{Var}(R_X)\,\mathrm{Var}(R_Z)}} = \frac{|\mathrm{Cov}(R_X, R_Z)|}{\sqrt{(1-r_{XY}^2)(1-r_{ZY}^2)}}$$

Also hat $\mathrm{Cov}(R_X, R_Z)$ den Maximalbetrag

$$\sqrt{(1-r_{XY}^2)(1-r_{ZY}^2)}\,.$$

Eine Regressionsanalyse mit *fallweiser* Behandlung fehlender Werte bringt trotz einer Stichprobe mit lediglich 11 Fällen ein plausibles Ergebnis:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-,498	,239		-2,078	,071	-1,050	,055
	XM	,586	,090	,734	6,545	,000	,380	,793
	ZM	,746	,129	,651	5,799	,000	,449	1,042

a. Abhängige Variable: YM

Bei *paarweiser* Behandlung erhält man hingegen einen Determinationskoeffizienten von 1,0

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	1,000[a]	1,000	1,000	,00000

a. Einflußvariablen : (Konstante), WM, XM

und sonstigen Unfug:

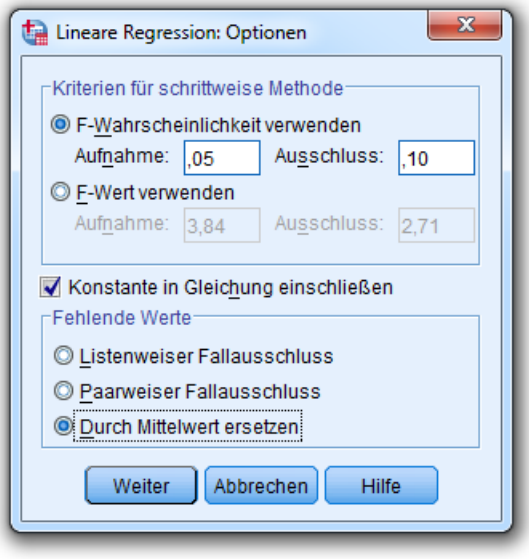
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,870	,000		.	.	,870	,870
	XM	1,261	,000	,931	.	.	1,261	1,261
	ZM	1,367	,000	,969	.	.	1,367	1,367

a. Abhängige Variable: YM

4.5 Ersetzung fehlender Werte durch den Stichprobenmittelwert

Eine Ersetzung fehlender Werte durch den univariaten (unbedingten) Mittelwert der betroffenen Variablen kann z.B. in der SPSS-Regressionsprozedur angefordert werden:



Zwar ändern sich für so behandelte Variablen die Randmittelwerte nicht, doch resultieren verzerrte Schätzer für Varianzen, Kovarianzen und Korrelationen. In der Formel für die geschätzte Varianz einer Variab-

len *X* bleibt bei Aufnahme von Mittelwertsfällen der Zähler unverändert, während der Nenner wächst, so dass der Schätzwert schrumpft:

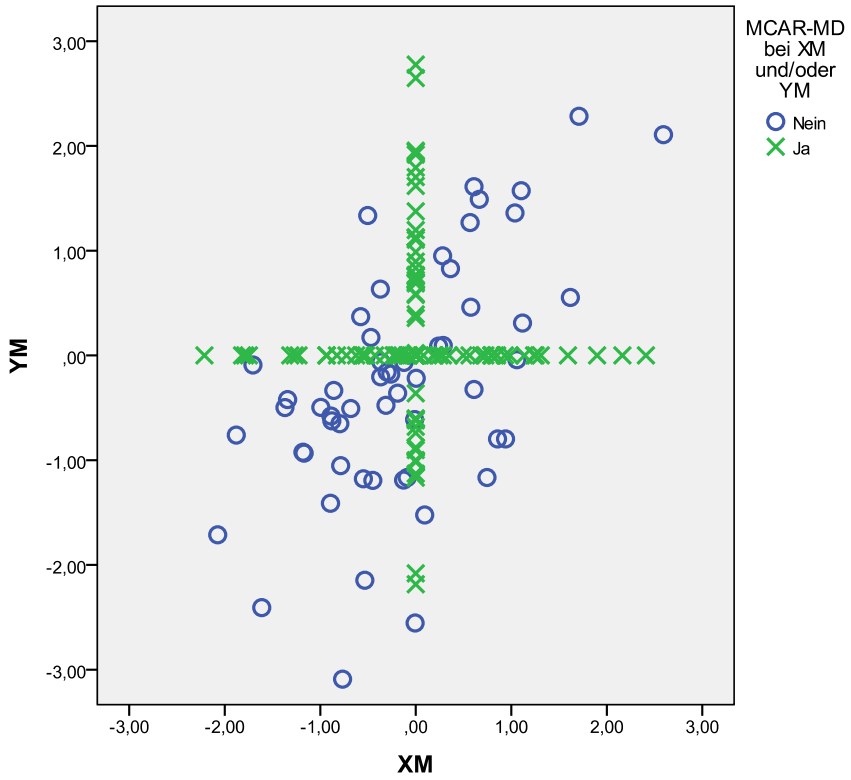
$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

Dass Korrelationen durch die Mittelwertsersetzung schrumpfen, wird an der folgenden Formel für die standardisierten Werte $z^{(x)}$ und $z^{(y)}$ zu zwei Variablen *X* und *Y* ersichtlich:

$$r_{xy} = \frac{\sum_{i=1}^{N} z_i^{(x)} z_i^{(y)}}{N-1}$$

Auch hier bleibt durch die Aufnahme von Mittelwertsfällen der Zähler unverändert, während der Nenner wächst.

Das folgende Streudiagramm zu zwei per Mittelwertsersetzung vervollständigten Variablen *X* und *Y* mahnt eindringlich, diese Technik *nicht* in Betracht zu ziehen:



Im Beispiel sinkt die Korrelation durch Aufnahme der Mittelwertefälle von 0,71 auf 0,30.

Man kann die Ersetzung durch univariate (unbedingte) Mittelwerte als Imputation per Regression (vgl. Abschnitt 4.8) betrachten, wobei aber zwei wichtige Bestandteile fehlen:

- Regressoren
  Weil auf Regressoren zur Schätzung der fehlenden Werte verzichtet wird, resultiert der Kriteriumsmittelwert.
- Zufallskomponente
  Wie sich bald zeigen wird, ist eine deterministische Imputation generell unzulässig.

Zu Vergleichszwecken soll die Mittelwerts-„Methode“ noch auf die Daten aus der bereits in den Abschnitten über die fallweise bzw. paarweise Behandlung fehlender Werte untersuchten Population angewendet werden, in der folgendes Modell gilt:

```
compute K = normal(1).
compute X = 0.5*K + normal(1).
compute Z = 0.5*K + normal(1).
compute Y = 0.7*X + 0.7*Z + normal(1).
```

Diesmal sorgen wir für fehlende *X*-Werte nach der MCAR-Bedingung:[1]

```
compute XM = X.
do if uniform(1) < 0.5.
  recode XM (lo thru hi = SYSMIS).
end if.
```

Es resultiert trotzdem ein deutlich verzerrter Schätzer für den Regressor *Z*:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,006	,021		,308	,758	-,035	,048
	XM	,694	,028	,340	25,171	,000	,640	,748
	Z	,775	,019	,543	40,172	,000	,738	,813

a. Abhängige Variable: Y

Die fallweise und die paarweise Behandlung fehlender Werte liefern bei der gerade analysierten MCAR-Stichprobe erwartungsgemäß korrekte Schätzwerte für die Regressionskoeffizienten.

4.6 MD-Indikatorvariable als Ergänzung eines kontinuierlichen Prädiktors

Cohen et al. (2003, S. 444) empfehlen für Regressionsmodelle, aus einem Prädiktor *X* mit fehlenden Werten die vervollständigte Variable X_V folgendermaßen herzustellen:

$$X_V = \begin{cases} X, & \text{falls } X \text{ einen gültigen Wert besitzt} \\ d, & \text{sonst} \end{cases}$$

Der *X*-Ersatzwert *d* ist prinzipiell beliebig, doch wird meist der Mittelwert der vorhandenen *X*-Werte benutzt. Zusätzlich zu X_V soll die folgendermaßen definierte MD-Indikatorvariable M_X in das Design der Regressionsanalyse aufgenommen werden:

$$M_X = \begin{cases} 1, & \text{falls der Beobachtungswert zu } X \text{ fehlt} \\ 0, & \text{sonst} \end{cases}$$

Damit steht (bei $d = \bar{x}$) der Regressionskoeffizient zu M_X für den Unterschied zwischen dem prognostizierten Kriteriumswert der MD-Fälle und dem prognostizierten Kriteriumswert für einen Fall mit dem *X*-Beobachtungswert $\bar{x}$, das Ganze jeweils bei einer festen Ausprägungskombination der restlichen Prädiktoren.

Bedauerlicherweise kann die Indikatortechnik auch in der MCAR-Situation verzerrte Schätzer liefern (Allison 2009, S. 76). Zur Demonstration verwenden wir dieselben Daten, an denen schon die Mittelwertstechnik gescheitert ist (siehe Abschnitt 4.5), und erhalten ein sehr ähnliches Fehlermuster (überhöhter Schätzer für den Regressor *Z*):

[1] Die vollständige Syntaxdatei **MC-Ergebnisse bei MCAR.sps** befindet sich an der im Vorwort genannten Stelle.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,024	,030		,812	,417	-,034	,083
	XV	,694	,028	,340	25,171	,000	,640	,748
	MX	-,040	,042	-,013	-,946	,344	-,122	,043
	Z	,775	,019	,543	40,158	,000	,737	,813

a. Abhängige Variable: Y

Dies war zu erwarten, weil beide Techniken dieselbe mittelwertsbehandelte *X*-Variante verwenden. Der MD-Indikator spielt in der MCAR-Situation keine Rolle, weil die Fälle mit fehlendem *X*-Wert annähernd denselben Kriteriumsmittelwert haben wie die beobachteten Ausprägungen.

Damit muss von der Indikator- wie von der Mittelwertstechnik *abgeraten* werden.

4.7 Zusatzkategorie bei nominalskalierten Prädiktoren

Bei nominalskalierten Prädiktoren in Regressionsmodellen empfehlen Cohen et al. (2003, S. 435ff), für die Fälle mit unbekanntem Wert eine zusätzliche Kategorie aufzunehmen. Allerdings lassen sich die in Abschnitt 4.6 formulierten Einwände analog übertragen (Allison 2009, S. 76). Es sind also auch in der MCAR-Bedingung verzerrte Schätzer zu befürchten.

Zur Demonstration wird eine Simulationsstudie mit einem dichotomen Regressor *D* und einem mit *D* korrelierten metrischen Regressor *Z* durchgeführt ($r_{DZ} = 0{,}64$):

```
compute K = normal(2).
compute D = (K + normal(1)) > 0.
compute Z = K + normal(1).
compute Y = 0.7*D + 0.7*Z + normal(1).
```

Bei 30% von insgesamt 5000 Fällen wird nach dem MCAR-Prinzip der *D*-Wert entfernt, so dass nunmehr drei *D*-Kategorien vorliegen (0, 1, fehlend), die über zwei Indikatorvariablen kodiert werden:

Indikatorvariable	*D*-Ausprägung		
	0	1	fehlend
D_1	0	1	0
D_2	0	0	1

Die Kategorie 0 dient als Referenz, und die Indikatorvariable D_1 steht für den Kontrast zwischen den beiden ursprünglichen *D*-Kategorien.

Bei fallweisem Ausschluss resultieren korrekte Schätzergebnisse (MCAR!):

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,020	,028		,712	,477	-,035	,075
	D	,693	,045	,167	15,374	,000	,605	,782
	Z	,701	,010	,756	69,540	,000	,681	,721

a. Abhängige Variable: Y

Bei Verwendung der beiden Indikatorvariablen wird der Kontrast zwischen den ursprünglichen *D*-Kategorien unterschätzt (Koeffizient zu D_1: 0,579), der Effekt des metrischen Regressors hingegen überschätzt (Koeffizient zu *Z*: 0,74):

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,077	,027		2,868	,004	,024	,129
	D1	,579	,041	,132	14,030	,000	,498	,660
	D2	,301	,037	,066	8,025	,000	,227	,374
	Z	,740	,008	,802	96,780	,000	,725	,755

a. Abhängige Variable: Y

In der Simulationsstudie spielt der korrelierte Regressor *Z* eine entscheidende Rolle. Bei einer einfaktoriellen Varianzanalyse (mit einem Faktor als *einzigem* Regressor) kann die vorgeführte Verzerrung *nicht* auftreten, und es spricht nichts dagegen, über eine MD-Zusatzkategorie die Kriteriumswerte der Fälle mit unbekannten Prädiktorwert vergleichend zu analysieren.

Bei einem Design mit zusätzlichen Regressoren ist von einer MD-Zusatzkategorie jedoch abzuraten. Weil die von SPSS Statistics zur multiplen Imputation verwendete FCS-Methode (*Fully Conditional Specification*) auch fehlende Werte bei *kategorialen* Variablen behandeln kann, ist eine sinnvolle Alternative verfügbar (siehe Abschnitt 6.2.4.2).

4.8 Regressionsimputation

Wenn ein statistisches Auswertungsverfahren vollständige Datensätze verlangt, und zahlreiche Fälle nur einen lückenhaften Wertevektor anbieten, führt der traditionell dominierende fallweise Ausschluss fehlender Werte oft zu einem inakzeptablen Verlust an statistischer Substanz. Hier liegt der Gedanke nahe, die unvollständigen Fälle zu komplettieren und für fehlende Werte plausible Schätzungen (**Imputationen**) zu verwenden. Dabei sind regressionsanalytische Techniken der in Abschnitt 4.5 beschriebenen Mittelwertersetzung deutlich überlegen. Stehen für eine Variable *Z* mit fehlenden Werten andere Variablen X_1, .. X_k mit kompletten Wertevektoren und prognostischer Relevanz für *Z* zur Verfügung, kann man aus den Fällen *mit Z*-Wert ein multiples Prognosemodell ermitteln und auf Fälle *ohne Z*-Wert anwenden. Wir wollen dieses Verfahren anschließend als *Regressionsimputation* (RI) bezeichnen.

Werden die Prognosen des Imputationsmodells direkt als Ersatz für fehlende Werte verwendet, liegt eine **deterministische Regressionsimputation** vor. Diese führt zur Verzerrungen bei vielen Verteilungsaspekten (z.B. Varianzen, Korrelationen) und insbesondere zu einem inflationierten Determinationskoeffizient im Analysemodell, wenn fehlende Kriteriumswerte unter Mitverwendung von Regressoren ersetzt werden.

Man muss zu den Prognosen des Imputationsmodells unbedingt eine Residualkomponente mit geeigneter Varianz addieren und erhält so eine **stochastische Regressionsimputation**. Ein Imputationswert kann als Zufallsziehung aus der bedingten Verteilung der behandelten Variablen gegeben die verwendeten Regresoren aufgefasst werden.

In einem Imputationsmodell können (und sollten) auch **Hilfsvariablen** als Prädiktoren eingesetzt werden, die im primär betrachteten Analysemodell *nicht* auftauchen, aber mit MD-belasteten Modellvariablen korrelieren.

Während z.B. Acock (2005, s. 1026) und Allison (2002, S. 54) ausdrücklich empfehlen, sowohl bei Regressoren wie auch beim Kriterium fehlende Werte zu ersetzen, sprechen sich andere Autoren dagegen aus, fehlende Kriteriumswerte unter Verwendung von Regressoren zu ersetzen (z.B. Cohen et al. 2003, S. 446). Grundsätzlich kann man Acock und Allison zustimmen. Wenn allerdings ausschließlich Kriteriumswerte fehlen, die MAR-Bedingung erfüllt ist und keine Hilfsvariablen verfügbar sind, lohnt sich das

Imputieren der Kriteriumswerte *nicht*. In dieser Situation bietet die fallweise Behandlung konsistente und effiziente Schätzer sowie korrekte Standardfehler (siehe Abschnitt 4.3.2).

Umgekehrt spricht beim Imputieren von fehlenden Regressorwerten nichts dagegen, auch das Kriterium heranzuziehen (Little & Rubin 2002, S. 66).

Nach Enders (2010, S. 46ff) liefert die stochastische Regressionsimputation unverzerrte Schätzer in der MAR-Bedingung.

Gegen die stochastische RI ist aber einzuwenden, dass bei den Anschlussanalysen die imputierten Werte wie beobachtete verwendet werden, obwohl in den Imputationsmodellen statt der Populationsparameter nur Stichprobenschätzer verwendet werden konnten, was für eine erhöhte Unsicherheit sorgt. Es resultieren unterschätzte Standardfehler zu den Regressionskoeffizienten und eine zu liberale Inferenzstatistik (Allison 2002, S. 12). Das Ausmaß der Unterschätzung von Standardfehlern hängt direkt von Umfang der fehlenden Information ab. Vermutlich kann man die Einfachimputation akzeptieren, wenn lediglich 5 % der Daten (auf möglichst intelligente Weise) ersetzt werden. Später werden zwei Methoden zur Behandlung fehlender Werte vorgestellt, die unter der MAR-Bedingung zu konsistenten Schätzern *und* zu einer korrekten Inferenzstatistik führen:

- Die von Strukturgleichungsprogrammen wie Amos angebotene *Full Information Maximum Likelihood* - Methode (siehe Abschnitt 5.3).
- Die *multiple* Imputation (siehe Abschnitt 6).

Bei Verwendung der **SPSS-Prozedur MVA** zur **stochastischen RI** ist leider mit **unplausiblen Ergebnissen** in der MAR-Bedingung zu rechnen. Wir verwenden zur Demonstration das von der fallweisen und der paarweisen Behandlung fehlender Werte (vgl. Abschnitte 4.3 und 4.4) bekannte Modell

```
compute K = normal(1).
compute X = 0.5*K + normal(1).
compute Z = 0.5*K + normal(1).
compute Y = 0.7*X + 0.7*Z + normal(1).
```

und simulieren als MD-Prozess sowohl MCAR

```
compute XM = X.
do if uniform(1) < 0.5.
  recode  XM (lo thru hi = SYSMIS).
end if.
```

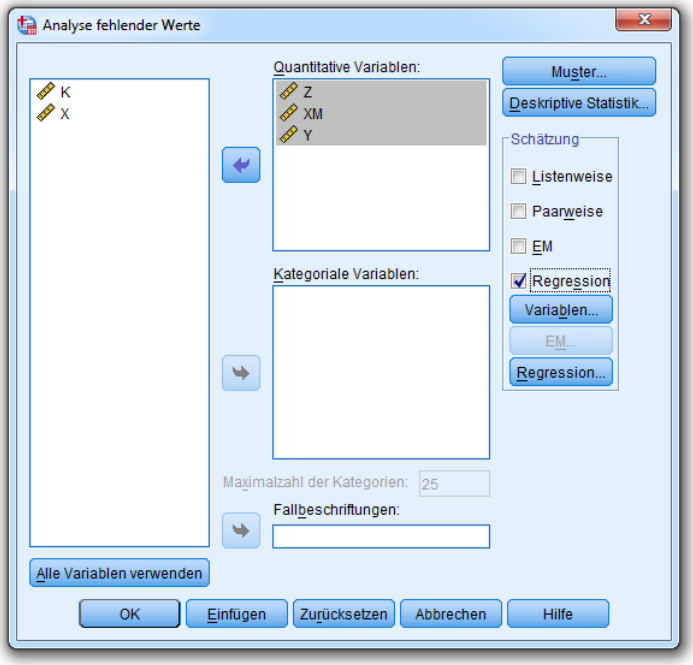
als auch MAR:

```
compute XM = X.
do if uniform(1) < exp(Y) / (1 + exp(Y)).
  recode XM (lo thru hi = SYSMIS).
end if.
```

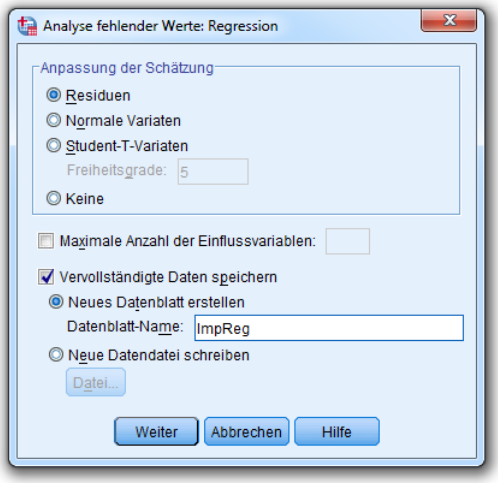
Nach dem Menübefehl

Analysieren > Analyse fehlender Werte

fordern wir im folgenden Dialog die **Schätzung** von Mittelwerten, Kovarianzen und Korrelationen mit der Methode **Regression** an:



Daraufhin wird der Schalter Regression... verfügbar, und seine Verwendung führt zur folgenden Subdialogbox mit Optionen zur Produktion von Imputationswerten:



Für die zu einem Prognosewert zu addierende Zufallskomponente bietet die Prozedur MVA folgende Optionen:

- **Residuen**
  Zufällige Wahl aus den beobachteten Residuen. Dies ist die Voreinstellung, wenn mindestens 50% der Werte vorhanden sind.
- **Normale Variaten**
  Ein normalverteilter Zufallswert mit Mittelwert Null und der im Imputationsmodell geschätzten Residualvarianz. Dies ist die Voreinstellung, wenn weniger als 50% der Werte vorhanden sind.
- **Student-T-Variaten**
  Ein t-verteilter Zufallswert
- **Keine**
  Verzicht auf eine Zufallskomponente (deterministische RI)

Außerdem wird in dieser Subdialogbox festgelegt, wo die **vervollständigten Daten** gespeichert werden sollen, wobei man eine Datei oder ein Datenblatt angeben kann. In der resultierenden Datenmatrix

landen alle in der MVA-Dialogbox benutzten Variablen, auch die als kategorial definierten. Es kann also sinnvoll sein, (kategoriale) Variablen in den MVA-Dialog aufzunehmen, damit sie in die Ausgabedatenmatrix gelangen.

Die per stochastischer Regressionsimputation vervollständigte Datenmatrix liefert in der MCAR-Situation plausible Schätzergebnisse:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,025	,018		1,330	,184	-,012	,061
	XM	,709	,017	,496	41,391	,000	,676	,743
	Z	,701	,017	,491	40,944	,000	,668	,735

a. Abhängige Variable: Y

Allerdings sind die Standardfehler (identische Werte wie beim vollständigen Datensatz!) zu klein. Die multiple Imputation (vgl. Abschnitt 6) liefert korrekte Werte:

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B		Anteil fehlende Info.	Relative Zunahmevarianz	Relative Effizienz
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze			
1	(Konstante)	,020	,020	,981	,327	-,020	,060	,189	,228	,991
	Z	,699	,020	35,131	,000	,660	,738	,274	,368	,986
	XM	,716	,022	31,901	,000	,671	,760	,438	,745	,979

a. Abhängige Variable: Y

In der MAR-Situation (die Wahrscheinlichkeit für einen fehlenden *X*-Wert wächst mit dem Kriteriumswert) führen die MVA-Imputationswerte zu verzerrten Regressionskoeffizienten:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,171	,021		8,172	,000	,130	,212
	XM	,594	,019	,395	30,699	,000	,556	,632
	Z	,800	,018	,560	43,497	,000	,764	,836

a. Abhängige Variable: Y

Der Koeffizient zum MD-belasteten Regressor *X* wird unterschätzt, und der Koeffizient zum unbelasteten (mit *X* korrelierten) Regressor *Z* wird überschätzt.

Verwendet man die Koeffizienten aus der Regression der MD-belasteten Variablen *X* auf *Y* und *Z*

```
COMPUTE  XM = -0.028 + 0.534*Y - 0.255*Z + NORMAL(0.859).
```

zur manuellen Imputation der fehlenden *X*-Werte, resultieren hingegen plausible Schätzungen der Regressionskoeffizienten im Analysemodell:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-,002	,018		-,126	,899	-,038	,033
	XMC	,723	,016	,514	44,045	,000	,691	,755
	Z	,706	,017	,494	42,342	,000	,674	,739

a. Abhängige Variable: Y

Erwartungsgemäß sind die Standardfehler zu klein.

Korrekt geschätzte Regressionskoeffizienten erhält man für die Beispieldaten mit MAR-Bedingung auch über die folgenden Methoden:

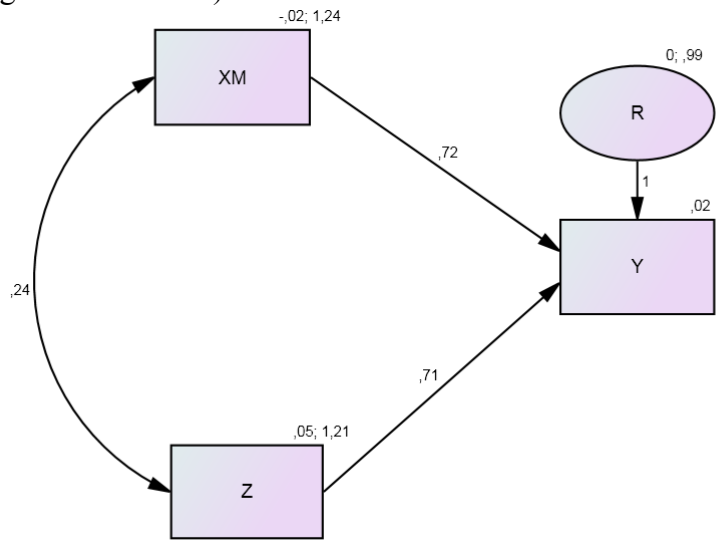
- Regressionsanalyse mit den per EM-Algorithmus geschätzten Verteilungsparametern (vgl. Abschnitt 5.1)

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	,019	,018		1,065	,287	-,016	,055
	XM	,716	,017	,507	43,037	,000	,683	,748
	Z	,707	,017	,495	41,980	,000	,674	,740

a. Abhängige Variable: Y

- FIML mit Amos (vgl. Abschnitt 5.3)



Regression Weights: (Group number 1 - Default model)

		Estimate	S.E.	C.R.	PLabel
Y	<--- XM	,7161	,0208	34,4439	***
Y	<--- Z	,7065	,0200	35,2866	***

- Multiple Imputation mit SPSS (vgl. Abschnitt 6)

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B	
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze
1	(Konstante)	,017	,024	,730	,467	-,030	,065
	Z	,707	,022	32,206	,000	,664	,751
	XM	,714	,022	33,044	,000	,671	,756

a. Abhängige Variable: Y

Die Standardfehler zu den Regressionskoeffizienten sind der FIML- und der MI-Methode korrekt, beim EM-basierten Verfahren hingegen zu klein (identisch mit den Werten für eine vollständige Stichprobe).

Offenbar können bei der stochastischen Regressionsimputation per MVA-Prozedur gravierende Fehler auftreten. Scheffer (2002, S. 160) kommt nach einer Simulationsstudie zum Verhalten verschiedener Programme mit Imputationsoptionen zur Empfehlung:

> If Single regression must be used, use EM or Regression Imputation, although not SPSS MVA REG imputation, as this gives VERY odd results.

Daher muss man empfehlen, auf die stochastische Regressionsimputation per MVA-Prozedur zu verzichten.

Neben der Regressionsimputation bietet die Prozedur MVA noch eine Einfachimputation auf der Basis von zuvor per EM-Algorithmus geschätzten Verteilungsparametern an, die sich in Abschnitt 5.2 als wenig tauglich erweisen wird (wegen fehlender Residualkomponente). Offenbar plant die Firma IBM SPSS nicht, die Mängel bei der Einfachimputation durch die Prozedur MVA zu beheben, und rät stattdessen zur multiplen Imputation (IBM Corp. 2012, S. 1):

> Note that multiple imputation is generally considered to be superior to single imputation.

5 Maximum Likelihood - Methoden

Das Maximum Likelihood - Prinzip zur Schätzung von Parametern spielt bei vielen statistischen Verfahren eine herausragende Rolle. Gegeben ein Modell und beobachtete Daten werden Parameterausprägungen so geschätzt, dass die Wahrscheinlichkeit der beobachteten Daten unter dem parametrisch spezifizierten Modell maximal wird. Eine ausführliche und gut lesbare Einführung in die ML-Schätzung bietet Enders (2010, S. 56ff).

Im aktuellen Abschnitt 5 werden zwei Maximum Likelihood - Methoden zur Analyse von Daten mit fehlenden Werten beschrieben:

- EM-Algorithmus (*Expectation Maximization*)
  Dabei werden zunächst Verteilungsmomente (Mittelwerte, Varianzen und Kovarianzen) geschätzt, die später als Eingabe für traditionelle Statistikmethoden dienen (z.B. für die multiple Regression). Man erhält unter der MAR-Bedingung konsistente Schätzer, aber unterschätzte Standardfehler und mithin eine ungültige Inferenzstatistik.
- Direkte ML-Schätzung mit FIML-Technik (*Full Information Maximum Likelihood*)
  Bei diesem Verfahren werden alle vorhandenen Daten genutzt, ohne dass fehlende Daten imputiert werden müssten. Man erhält unter der MAR-Bedingung konsistente Schätzer und korrekte Standardfehler.

Während der EM-Algorithmus in SPSS Statistics verfügbar ist, wird die attraktivere FIML-Technik in der IBM SPSS - Produktfamilie vom Strukturgleichungsanalyseprogramm Amos realisiert. Viele SPSS-Anwender(innen) müssen sich also in ein zusätzliches Programm einarbeiten, um die FIML-Technik nutzen zu können.

Auf der Suche nach einer angemessenen Behandlung fehlender Werte steht oft eine Entscheidung an zwischen der von Amos angebotenen FIML-Methode und der in SPSS Statistics verfügbaren multiplen Imputation (siehe Abschnitt 6). Trotzdem betrachten wir zunächst das EM-Verfahren, um die Nutzungsmöglichkeiten und Beschränkungen dieser Technik kennen zu lernen.

5.1 ML-Schätzung von Verteilungsparametern per EM-Algorithmus

Mit dem EM-Verfahren (*Expectation Maximization*) lassen sich bei erfüllter MAR-Bedingung valide Maximum-Likelihood – Schätzer für Mittelwerte, Varianzen und Kovarianzen ermitteln (Allison 2002, S. 18; von Hippel, 2004, S. 162). Mit den geschätzten Verteilungsparametern lassen sich diverse lineare Modelle analysieren (z.B. lineare Regression, Faktorenanalyse). Das skizzierte Verfahren besteht also aus zwei getrennten Phasen:

- ML-Schätzung der Verteilungsparameter
  In der ersten Phase werden per EM-Algorithmus Mittelwerte, Varianzen und Kovarianzen geschätzt. Wie bei jeder ML-Schätzung ist eine Verteilungsannahme erforderlich. Meist wird die multivariate Normalverteilung der Variablen angenommen, wobei diese Annahme für Variablen *ohne* fehlende Werte aber irrelevant ist. Bei gültiger MCAR-Bedingung erweist sich die ML-Schätzung per EM-Algorithmus auch dann noch als robust, wenn fehlende Werte bei Variablen ohne Normalverteilung (z.B. bei Indikatorvariablen) vorliegen (Allison 2002, S. 18).
- Anwendung einer Auswertungsprozedur, die Parameterschätzungen als Eingabe akzeptiert
  Viele Auswertungsverfahren (z.B. lineare Regression, Faktorenanalyse) können statt mit Rohdaten auch mit geschätzten Verteilungsparametern gefüttert werden.

Anschließend wird der EM-Algorithmus zur Schätzung von Verteilungsparametern unter der Annahme der multivariaten Normalität nach Allison (2002, 19f) beschrieben.

Zunächst werden Startwerte für die Normalverteilungsparameter (Mittelwerte, Varianzen und Kovarianzen) mit konventionellen Methoden ermittelt, also z.B. bei fallweisem oder paarweisem Ausschluss fehlender Werte.

Dann beginnt ein iteratives Verfahren mit zwei Teilen pro Schritt, die auch für den Namen des Verfahrens verantwortlich sind:

1. **Expectation (Ersatzwerte schätzen)**
   Aufgrund der aktuellen Schätzungen für die Normalverteilungsparameter und der vorhandenen Beobachtungen werden die bedingten Erwartungen für die fehlenden Werte mit Regressionstechniken ermittelt. Um bei einem konkreten Fall den fehlenden Wert einer Variablen durch deterministische Regressionsimputation zu ersetzen, kommen alle Variablen mit vorhandenen Werten zum Einsatz. Dabei wird nicht zwischen abhängigen und unabhängigen Variablen unterschieden.
2. **Maximization (ML-Schätzung der Normalverteilungsparameter)**
   Aus den beobachteten und konstruierten Daten werden neue Schätzer für die Mittelwerte, Varianzen und Kovarianzen nach dem Maximum-Likelihood-Prinzip ermittelt. Über zusätzliche Terme für die Residualvarianzen und –kovarianzen wird vermieden, dass durch die deterministische Regressionsimputation im 1. Teilschritt Verzerrungen entstehen.

Dann beginnt die nächste Iteration mit einer neuen Berechnung der bedingten Erwartungen für die fehlenden Werte unter Verwendung der aktuellen Normalverteilungsparameter (Teilschritt 1). Der Algorithmus endet, wenn sich die Schätzungen der Normalverteilungsparameter nicht mehr ändern.

Die EM-Schätzung von Verteilungsparametern mit anschließender Linearmodellierung ist in der MAR-Situation folgendermaßen zu bewerten:

- Valide ML-Schätzer der Verteilungsparameter
  Man erhält valide ML-Schätzer der Verteilungsparameter, also konsistente (asymptotisch erwartungstreue), asymptotisch effiziente und asymptotisch normalverteilte Schätzer, sofern …
  - die Verteilungsannahme (z.B. multivariate Normalverteilung) akzeptabel ist,
  - die Stichprobe hinreichend groß ist.
- Konsistente Schätzer der Parameter in den Linearmodellen
  Die per EM-Verfahren ermittelten Momentmatrizen werden anschließend mit konventionellen Methoden analysiert, wobei konsistente Schätzer der Modellparameter resultieren.
- Fehlerhafte Inferenzstatistik zu den Modellparametern
  Weil die zur Analyse der geschätzten Verteilungsparameter verwendeten Methoden in der Regel von kompletten Daten ausgehen, sind die ermittelten Standardfehler und Überschreitungswahrscheinlichkeiten zu klein.

In der MCAR-Situation nutzt das EM-Verfahren die vorhandenen Informationen besser als der fallweise Ausschluss. Im Vergleich zu den besten aktuell verfügbaren Techniken (FIML und multiple Imputation, siehe Abschnitte 5.3 und 6) ist die fehlerhafte Inferenzstatistik zu bemängeln.

Durchaus zu empfehlen ist die Zwei-Phasen - Prozedur (Verteilungsmomente per EM-Algorithmus schätzen, anschließende Analyse durch Verfahren mit der Fähigkeit zum Import von Verteilungsmomenten), wenn bei der statistischen Analyse Signifikanztests und Vertrauensintervalls wenig relevant sind, z.B.:

- Explorative Faktorenanalyse
  Eine Schätzung von Faktorwerten ist allerdings aufgrund einer Momentenmatrix *nicht* möglich.
- Reliabilitätsschätzung über die interne Konsistenz (Cronbachs α)

Leider muss man bei SPSS Statistics in Phase 2 mit einigem Aufwand Tabellen mit den geschätzten Verteilungsmomenten aus dem Ausgabefenster in eine Datendatei überführen. Wir unterziehen uns der Mühe

und verwenden die Zwei-Phasen - Prozedur zur Analyse des in Abschnitt 3.1 beschriebenen Modells zur Vorhersage des Ausbildungserfolgs an amerikanischen Hochschulen:[1]

- Kriterium: GRADRAT
- Regressoren: CSAT, LNENROLL, PRIVATE, STUFAC und RMBRD

Bei einer linearen Regressionsanalyse mit der voreingestellten fallweisen Behandlung fehlender Werte resultiert die folgende Koeffiziententabelle:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-37,298	8,430		-4,424	,000	-53,876	-20,719
	CSAT	,067	,007	,442	9,249	,000	,053	,081
	LNENROLL	2,803	1,055	,138	2,656	,008	,728	4,879
	PRIVATE	14,957	2,192	,366	6,825	,000	10,648	19,267
	STUFAC	-,046	,142	-,014	-,322	,748	-,325	,233
	RMBRD	1,789	,784	,105	2,284	,023	,249	3,330

a. Abhängige Variable: GRADRAT

Von den 1302 Fällen in der Stichprobe verbleiben nur 372, so dass ein sehr erheblicher Teil der verfügbaren Information verloren geht. Für den Regressor STUFAC wird ein betragsmäßig sehr kleiner Regressionskoeffizient geschätzt (-0,46) mit einem sehr großen, weit von Signifikanz entfernten, *p*-Level (0,748). Vermutlich führen zwei unerwünschte Ursachen zu dieser ungünstigen, von der theoretischen Erwartung abweichenden, Bewertung:

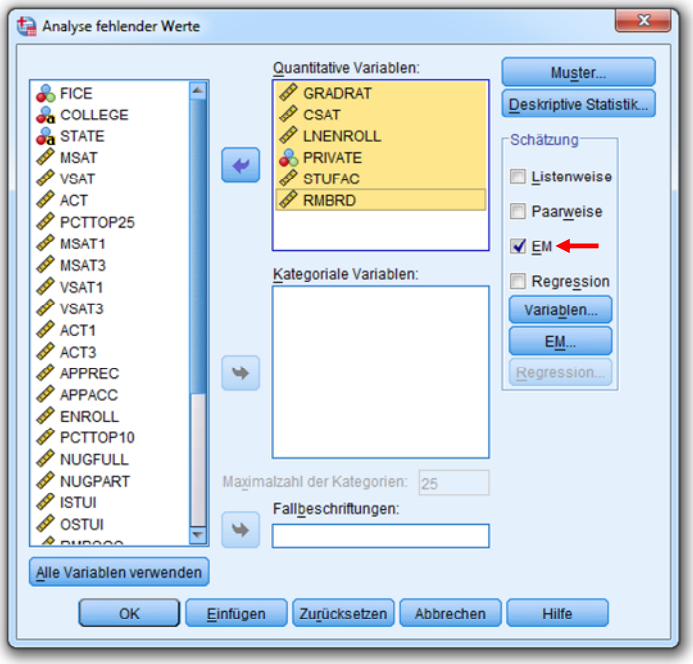
- Verzerrung
  Im Beispiel ist die MCAR-Bedingung nicht erfüllt (siehe Abschnitt 3.2.2), so dass beim fallweisen Ausschluss fehlender Werte Verzerrungen zu befürchten sind.
- Drastische Reduktion der Stichprobengröße
  Dies reduziert die Power der Hypothesentests.

Nun lassen wir die Mittelwerte, Kovarianzen und Korrelationen von der SPSS-Prozedur MVA, die über den Menübefehl

Analysieren > Analysieren fehlender Werte

verfügbar ist, per EM-Algorithmus schätzen:

[1] Wie in Abschnitt 3.1 beschrieben, verwenden wir im Kurs die Datei **UsNews mit MCAR-MDs bei PRIVATE.sav**, die im Vergleich zum Original (siehe http://lib.stat.cmu.edu/datasets/colleges/) auch bei der kategorialen Variablen PRIVATE fehlende Wert aufweist. Diese Datei ist an der im Vorwort vereinbarten Stelle zu finden.



Von der im EM-Algorithmus enthaltenen Annahme der multivariaten Normalverteilung sind nur Variablen mit fehlenden Werten betroffen (Allison 2002, S. 18). Während die dichotome (also sicher *nicht* normalverteilte) Variable PRIVAT in der Originalversion der Beispieldatei komplett vorhanden ist, haben wir ca. 20 % der Werte per Zufall (also unter der freundlichen MCAR-Bedingung) gelöscht. Beruhigenderweise haben Simulationsstudien gezeigt, dass die ML-Schätzer bei gültiger MCAR-Bedingung robust arbeiten, wenn fehlende Werte bei Variablen ohne Normalverteilung (z.B. bei dichotomen Variablen) auftreten (Allison 2002, S. 18). Daher ist es zu vertreten, im MVA-Dialog den dichotomen Regressor PRIVATE als **quantitative Variable** zu deklarieren und so in das EM-Verfahren einzubeziehen.

Wir erhalten im Ausgabefenster unter der Überschrift

EM-geschätzte Statistiken

die gewünschten Schätzergebnisse:

Geschätzte Randmittel[a]

GRADRAT	CSAT	LNENROLL	PRIVATE	STUFAC	RMBRD
59,94	959,08	6,1696	,64	14,863	4,0836

a. MCAR-Test nach Little: Chi-Quadrat = 221,975, DF = 73, Sig. = ,000

EM-Korrelationen[a]

	GRADRAT	CSAT	LNENROLL	PRIVATE	STUFAC	RMBRD
GRADRAT	1					
CSAT	,595	1				
LNENROLL	-,027	,187	1			
PRIVATE	,405	,168	-,616	1		
STUFAC	-,320	-,307	,267	-,375	1	
RMBRD	,479	,480	-,022	,344	-,281	1

a. MCAR-Test nach Little: Chi-Quadrat = 221,975, DF = 73, Sig. = ,000

Unter beiden Tabellen wird das Ergebnis des MCAR-Tests nach Little protokolliert (vgl. Abschnitt 2.1). Der Test verwirft seine Nullhypothese, und wir hoffen, dass die nicht prüfbare MAR-Bedingung annähernd erfüllt ist.

Mit den EM-Schätzergebnissen lässt sich die geplante Regressionsanalyse durchführen. Leider ist dazu in SPSS einige Handarbeit nötig, um z.B. mit der folgenden Syntax eine geeignete Matrix-Datendatei zu erstellen:[1]

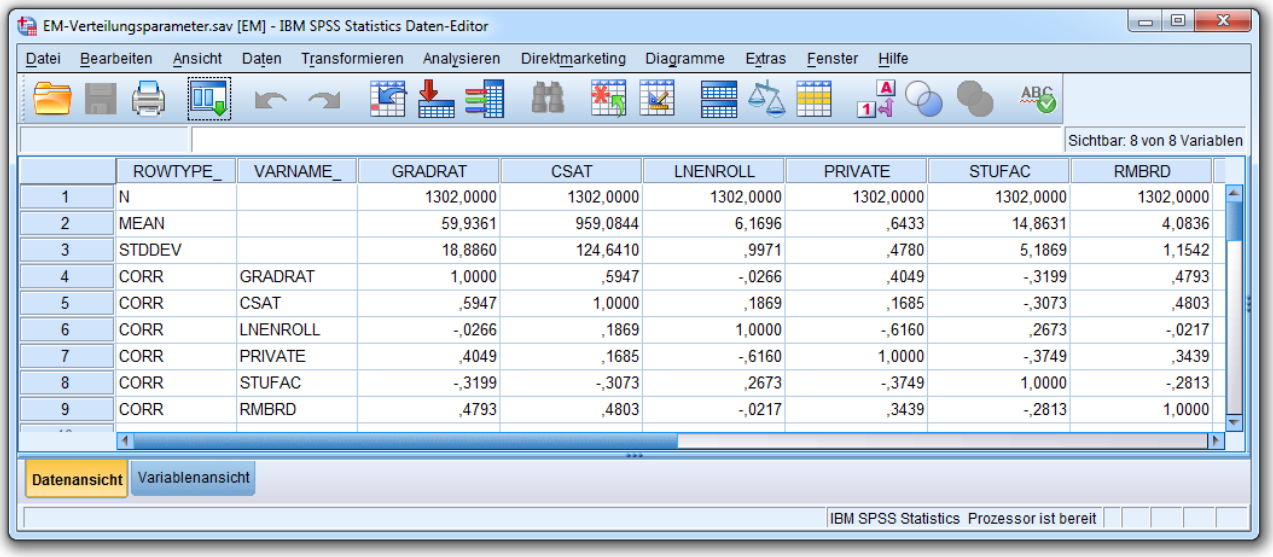
```
MATRIX DATA VARIABLES=ROWTYPE_ GRADRAT CSAT LNENROLL PRIVATE STUFAC RMBRD.
BEGIN DATA
MEAN   59,93609  959,08443  6,16960   ,64330  14,86309  4,08355
STDDEV 18,886    124,641     ,99711   ,478     5,1869   1,15424
N      1302      1302       1302      1302     1302     1302
CORR   1,00000
CORR    ,59466  1,00000
CORR   -,02659   ,18694  1,00000
CORR    ,40487   ,16847  -,61596  1,00000
CORR   -,31985  -,30730   ,26731  -,37487  1,00000
CORR    ,47932   ,48026  -,02172   ,34385  -,28128  1,00000
END DATA.
```

Die EM-Schätzer der Standardabweichungen stammen aus der folgenden MVA-Ergebnistabelle:

Zusammenfassung der geschätzten Standardabweichungen

	GRADRAT	CSAT	LNENROLL	PRIVATE	STUFAC	RMBRD
Alle Werte	18,889	123,577	,99715	,478	5,1864	1,16959
EM	18,886	124,641	,99711	,478	5,1869	1,15424

Um mit dem resultierenden Datenblatt



	ROWTYPE_	VARNAME_	GRADRAT	CSAT	LNENROLL	PRIVATE	STUFAC	RMBRD
1	N		1302,0000	1302,0000	1302,0000	1302,0000	1302,0000	1302,0000
2	MEAN		59,9361	959,0844	6,1696	,6433	14,8631	4,0836
3	STDDEV		18,8860	124,6410	,9971	,4780	5,1869	1,1542
4	CORR	GRADRAT	1,0000	,5947	-,0266	,4049	-,3199	,4793
5	CORR	CSAT	,5947	1,0000	,1869	,1685	-,3073	,4803
6	CORR	LNENROLL	-,0266	,1869	1,0000	-,6160	,2673	-,0217
7	CORR	PRIVATE	,4049	,1685	-,6160	1,0000	-,3749	,3439
8	CORR	STUFAC	-,3199	-,3073	,2673	-,3749	1,0000	-,2813
9	CORR	RMBRD	,4793	,4803	-,0217	,3439	-,2813	1,0000

eine lineare Regressionsanalyse zu rechnen, ist erneut Syntax erforderlich, weil die zuständige Dialogbox das Einlesen aus einer Matrix-Datendatei nicht unterstützt:

[1] Leider akzeptiert die SPSS-Prozedur REGRESSION in einer Eingabematrix keine Kovarianzen, so dass man die Korrelationen *und* die Standardabweichungen übergeben muss. Die Datei **Regression mit Matrix-Datei (EM).sps** mit der Syntax findet sich neben anderen Dateien zum Beispiel im Ordner **America's Best Colleges 1994** an der im Vorwort vereinbarten Stelle.

```
regression matrix=in(*)
 /statistics = defaults ci(95)
 /dependent=GRADRAT
 /enter=CSAT LNENROLL PRIVATE STUFAC RMBRD.
```

Nach der recht umständlichen Prozedur erhalten wir endlich die Regressionsanalyse basierend auf den EM-Schätzergebnissen:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-31,574	4,281		-7,376	,000	-39,972	-23,176
	CSAT	,065	,004	,431	17,315	,000	,058	,073
	LNENROLL	2,191	,533	,116	4,111	,000	1,146	3,237
	PRIVATE	13,195	1,146	,334	11,518	,000	10,948	15,443
	STUFAC	-,190	,084	-,052	-2,280	,023	-,354	-,027
	RMBRD	2,378	,398	,145	5,980	,000	1,598	3,158

a. Abhängige Variable: GRADRAT

Im Vergleich zur fallweisen Behandlung fehlender Werte sind die Standardfehler deutlich keiner, so dass nun ein signifikanter Effekt der Variablen STUFAC gemeldet wird. Leider sind die Standardfehler mit Vorsicht zu genießen, weil sie unter der Annahme vollständiger Daten berechnet wurden, also mehr oder weniger deflationiert sind. Ähnlich wie beim paarweisen Ausschluss ist es unklar, welche Stichprobengröße man angeben soll, weil die einzelnen Momente unterschiedlich präzise geschätzt sind (Enders 2010, S. 132).

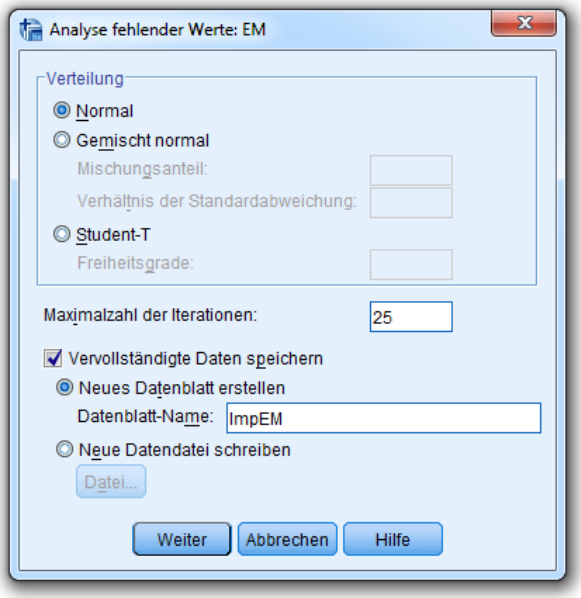
Wie später im Zusammenhang mit der FIML-Schätzmethode in Strukturgleichungsanalysen (siehe Abschnitt 5.3.2) noch ausführlich diskutiert wird, profitiert die ML-Schätzung von der Verwendung von **Hilfsvariablen**, die zwar ohne Bedeutung im intendierten Modell sind, aber Informationen enthalten über die Auftretenswahrscheinlichkeit und die Ausprägung fehlender Werte bei Modellvariablen:

- Man reduziert potentiell Verletzungen der MAR-Bedingung und damit Verzerrungen bei der Parameterschätzung.
- Man steigert die Präzision beim Schätzen und Testen, weil fehlende Informationen teilweise kompensiert werden.

Bei der Zwei-Phasen - Prozedur lassen sich diese Hilfsvariablen in praktisch beliebiger Zahl ohne großen Aufwand in die erste Phase integrieren, um die Schätzungen der Momente (Mittelwerte, Varianzen, Kovarianzen) von Modellvariablen zu verbessern. Wenn in Phase 2 die EM-geschätzten Momente analysiert werden, lässt man die Hilfsvariablen weg, so dass hier keine zusätzliche Komplexität auftritt.

5.2 Einfache Imputation nach EM-Schätzung der Verteilungsmomente

Die eben beschriebene Analyse von EM-geschätzten Verteilungsmomenten ist etwas umständlich und nur realisierbar, wenn bei der intendierten Analyse Verteilungsmomente als Datenbasis verarbeitet werden können. Eine vervollständigte Datenmatrix bietet mehr Flexibilität. Neben der in Abschnitt 4.8 beschriebenen stochastischen Regressionsimputation, deren SPSS-Implementierung nur unter der MCAR-Bedingung einsetzbar ist, bietet SPSS Statistics auch eine Einfachimputation per EM-Algorithmus, die auch unter der MAR-Bedingung funktionieren sollte. Die SPSS-Prozedur MVA kann die Ersatzwerte aus der letzten Iteration des EM-Verfahrens (siehe Abschnitt 5.1) als neue Datendatei oder neues Datenblatt abspeichern. Diese Leistung wird in folgender Subdialogbox angefordert, die vom Hauptdialog aus bei markiertem **EM**-Kontrollkästchen über den Schalter EM... erreichbar ist:



Leider versäumt es SPSS, durch Addition einer Zufallskomponente die Residualvarianzen der Imputationsmodelle zu berücksichtigen, so dass bei einer Analyse der vervollständigten Daten mit verzerrten Ergebnissen zu rechnen ist (siehe von Hippel 2004). Werden z.B. Kriteriumswerte in nennenswertem Umfang imputiert, ist mit folgenden Fehlern zu rechnen:

- Unterschätzte Fehlervarianz und infolgedessen unterschätzte Standardfehler zu den Regressionskoeffizienten
- Überschätzter Determinationskoeffizient

Die im **Verteilungs**-Rahmen der obigen Dialogbox wählbaren Optionen wirken sich auf den EM-Algorithmus aus, haben aber keine Imputationswerte mit Zufallskomponente zur Folge. Die von SPSS Statistics gelieferten EM-Kovarianz- bzw. Korrelationsmatrizen sind in Ordnung, weil im M-Teilschritt die Unsicherheit in den Ersatzwerten durch Terme für die Residualvarianzen und –kovarianzen berücksichtigt wird (siehe Abschnitt 5.1; vgl. Enders 2010, S. 105-112). Während der EM-Iterationen benötigt SPSS Statistics also keine Zufallszahlen. Leider wird es versäumt, diese beim Erstellen der Imputationswerte zu ergänzen. Daher muss davor gewarnt werden, die von SPSS Statistics gelieferten EM-Imputationswerte zu verwenden.

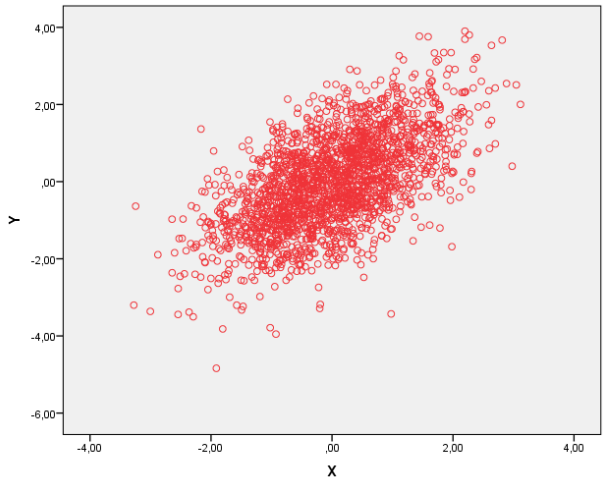
Die Simulation einer einfachen Regression von *Y* auf *X* mit fehlenden *Y*-Werten für alle Fälle mit einem unterdurchschnittlichen *X*-Wert[1]

```
compute X = normal(1).
compute Y = 0.7*X + normal(1).
compute YM = Y.
do if (x < 0).
  recode YM (lo thru hi = SYSMIS).
end if.
```

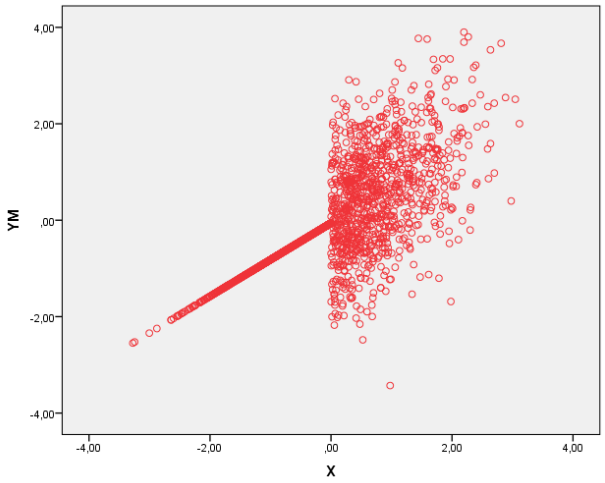
kann man das Fehlverhalten der MVA-Prozedur veranschaulichen:

[1] Die vollständige Syntaxdatei **Fehlerhafte EM-Ersatzwerte.sps** befindet sich an der im Vorwort genannten Stelle im Ordner **MC-Ergebnisse bei MAR\EM\Fehlerhafte EM-Ersatzwerte**.

komplette Werte



MVA-Ersatzwerte (EM)



Aus den EM-imputierten Werten errechnet sich ein erheblich inflationierter Wert von 0,72 für die Korrelation zwischen *X* und *Y* (wahrer Wert: 0,573) sowie ein deflationierter **Standardfehler des Schätzers** (geschätzte Fehlerstandardabweichung, wahrer Wert: 1). Dies zeigt sich in der folgenden Tabelle zur linearen Regression von YM auf X:

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,738[a]	,545	,544	,69907

a. Einflußvariablen : (Konstante), X

Der Regressionskoeffizient (wahrer Wert: 0,7) wird passabel geschätzt, doch sind Standardfehler und Konfidenzintervall zu klein:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-,058	,016		-3,683	,000	-,088	-,027
	X	,760	,016	,738	48,880	,000	,730	,791

a. Abhängige Variable: YM

Bei fallweiser Behandlung fehlender Werte sind die Schätzer für den Regressionskoeffizienten und für den Standardfehler in Ordnung:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-,058	,051		-1,135	,257	-,159	,043
	X	,761	,051	,428	14,985	,000	,661	,861

a. Abhängige Variable: YM

Dies war trotz MAR-Bedingung zu erwarten, weil ausschließlich Kriteriumswerte fehlen (vgl. Abschnitt 4.3.2). Allerdings wird der Determinationskoeffizient (wahrer Wert: 0,329) aufgrund der Varianzeinschränkung bei *X* deutlich unterschätzt:

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,428[a]	,183	,182	,98715

a. Einflußvariablen : (Konstante), X

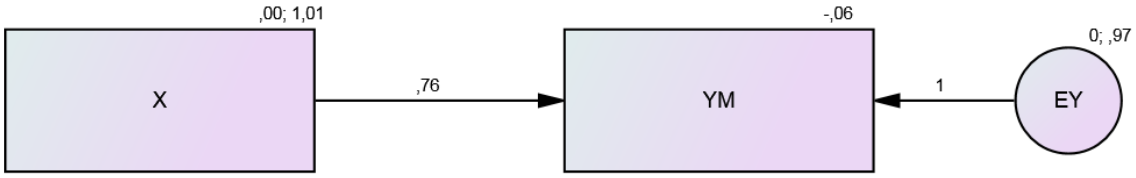
Der EM-Algorithmus ist vom Problem der EM-Einfachimputation *nicht* betroffen und kann die Korrelation von *Y* und *X* recht gut rekonstruieren (, wobei die voreingestellte Anzahl von 25 Iterationen nicht reichte, um die Konvergenz zu erreichen):

EM-Korrelationen[a]

	X	YM
X	1	
YM	,612	1

a. MCAR-Test nach Little: Chi-Quadrat = 1292,613, DF = 1, Sig. = ,000

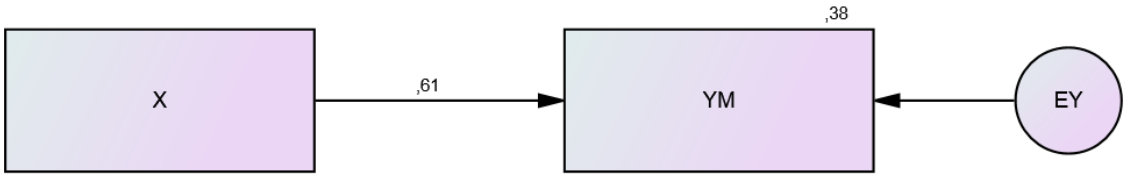
Nun kommen wir zu einer Methodologie, die das Problem fehlender Werte in ihrem Anwendungsbereich überzeugend löst, indem sie konsistente Schätzer *und* eine korrekte Inferenzstatistik bietet. Für die eben diskutierte Simulationsstudie liefert Amos per FIML-Technologie sinnvolle Schätzer für den Regressionskoeffizienten



und seinen Standardfehler:

	Estimate	S.E.	C.R.	P	Label
YM <--- X	,7611	,0310	24,5687	***	

Beim standardisierten Regressionskoeffizienten (identisch mit der Korrelation zwischen *X* und *Y*) stimmen FIML- und EM-Schätzung gut überein:



5.3 Direkte ML-Schätzung in Strukturgleichungsmodellen

Übergibt man die per EM-Algorithmus geschätzten Normalverteilungsparameter (Mittelwerte, Varianzen, Kovarianzen) z.B. an eine Regressionsprozedur, erhält man konsistente Schätzungen für die Regressionskoeffizienten. Allerdings sind die zugehörigen Standardfehler deflationiert, weil das Informationsdefizit gegenüber Verteilungsmomenten aus vollständig beobachteten Daten nicht in Rechnung gestellt wird. Statt die generell für MAR-Verhältnisse gut geeignete Maximum Likelihood - Technologie lediglich auf Verteilungsmomente anzuwenden, kann man bei vielen Modellen die letztlich interessierenden Parameter (z.B. Regressionskoeffizienten, Faktorladungen) auch *direkt* mit ML-Methoden schätzen. Bei diesem meist als *Full Information Maximum Likelihood* (FIML) bezeichneten Verfahren werden alle vorhandenen Daten genutzt, ohne dass fehlende Daten imputiert werden müssten. Man erspart sich einigen Auf-

wand sowie das Problem deflationierter Standardfehler. Wie der EM-Algorithmus ist auch das FIML-Verfahren unter der MAR-Bedingung zulässig (siehe z.B. Allison 2002, S. 18).

Mittlerweile wird das FIML-Verfahren von den meisten Strukturgleichungsanalyseprogrammen unterstützt (z.B. Amos, LISREL, Mplus). In SPSS Statistics wird das Verfahren aber nicht angeboten, so dass wir im Abschnitt 5.3 mit Amos arbeiten. Zumindest im Hochschulbereich steht Amos bei der aktuellen Lizenzpolitik des Herstellers IBM SPSS in der Regel zusammen mit SPSS Statistics zur Verfügung.

Zur Berechnung der Standardfehler zu den FIML-Schätzern sind nach Enders (2010, S. 97ff) zwei Verfahren im Einsatz, wobei entweder die *erwartete* oder die *beobachtete* Informationsmatrix verwendet wird. Nur bei Verwendung der *beobachten* Informationsmatrix erhält man korrekte Vertrauensintervalle unter der MAR-Bedingung, während die alternative Berechnungsmethode MCAR-Verhältnisse voraussetzt. Wer nun eine Erläuterung der Informationsmatrix erwartet, wird enttäuscht. Der Begriff wurde erwähnt für den Fall, dass Sie zwischen den beiden Algorithmen wählen können bzw. müssen. Nach Enders (2010, S. 231) verwendet Amos das empfehlenswerte, MAR-taugliche Verfahren.

Bei MNAR-Verhältnissen kann auch die FIML-Methode eine Verzerrung von Parametern nicht verhindern. Im Unterschied zu traditionellen MD-Techniken schafft es die FIML-Methode (wie die im Abschnitt 6 vorzustellende multiple Imputation) jedoch, die Verzerrungen auf die von fehlenden Werten betroffenen Variablen zu beschränken (Enders 2010, S. 96 und 125).

5.3.1 FIML-Lösung zum Colleges-Beispiel

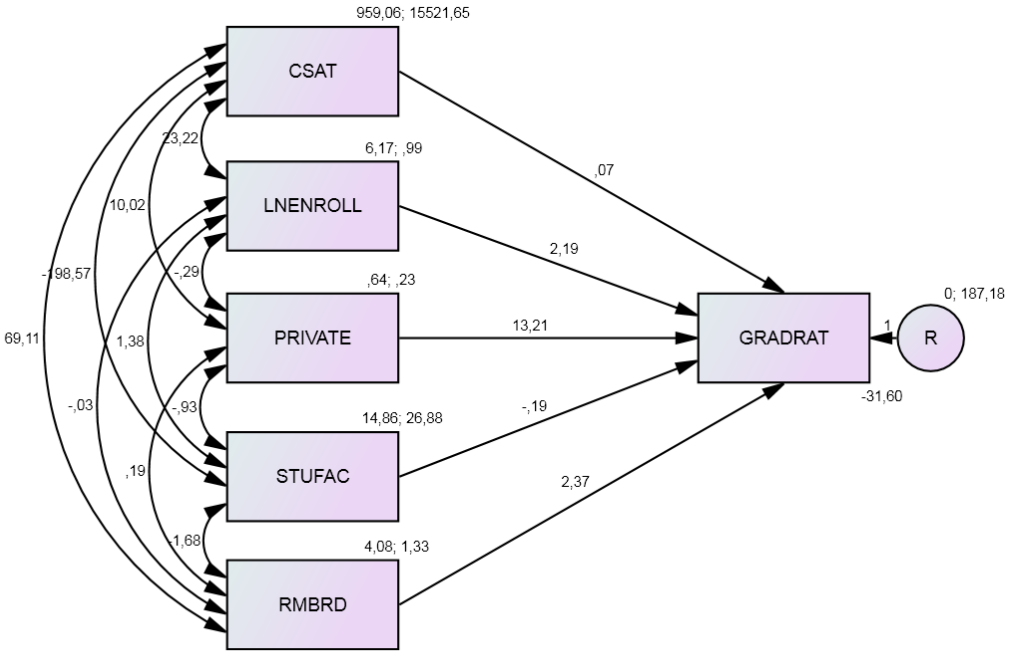
Zum Colleges-Beispiel aus Abschnitt 3.1 liefert Amos 21 folgende Schätz- und Testergebnisse zu den Regressionskoeffizienten:

Regression Weights: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
GRADRAT <--- CSAT	,0653	,0049	13,2528	***	
GRADRAT <--- LNENROLL	2,1919	,6274	3,4934	***	
GRADRAT <--- PRIVATE	13,2142	1,4186	9,3150	***	
GRADRAT <--- STUFAC	-,1902	,0948	-2,0061	,0448	
GRADRAT <--- RMBRD	2,3732	,5664	4,1903	***	

Wir erhalten praktisch dieselben Parameterschätzer wie bei der Regressionsanalyse mit EM-geschätzten Verteilungsmomenten (siehe Abschnitt 5.1). Erwartungsgemäß sind die geschätzten FIML-Standardfehler größer (und vertrauenswürdiger), doch kommen alle Tests zum selben Ergebnis, insbesondere auch der Test zum Regressor STUFAC.

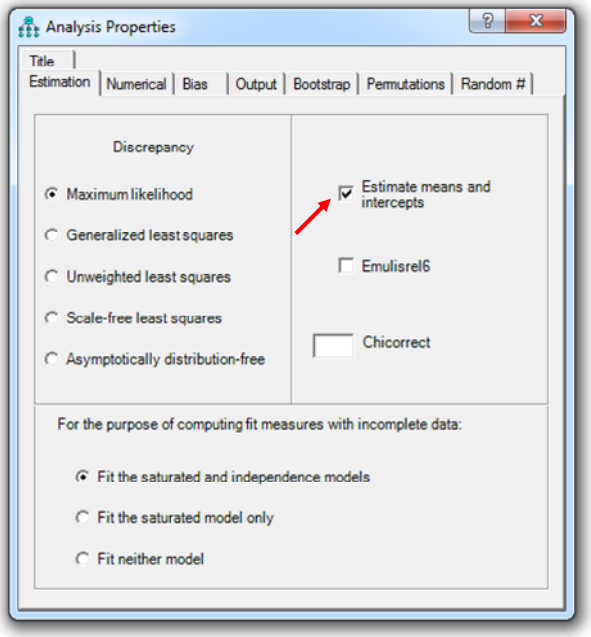
Das in Amos zur Modellspezifikation verwendete Pfaddiagramm eignet sich auch gut zur Präsentation der Schätzergebnisse:



Eine FIML-Schätzung mit fehlenden Werten benötigt ein Modell mit Mittelwertstruktur. Folglich muss bei der Modellspezifikation nach dem Menübefehl

View > Analysis Properties

in folgender Dialogbox



ein Modell mit Mittelwerten und Ordinatenabschnitten angefordert werden. Zur generellen Amos-Bedienung siehe z.B. Baltes-Götz (2010).

Offenbar hat die in Amos (und anderen Strukturgleichungsprogrammen) mögliche FIML-Schätzung gegenüber der in Abschnitt 5.1 beschriebenen Vorgehensweise (Regression nach EM-Schätzung der Verteilungsparameter) deutliche Vorteile bei vergleichbaren Voraussetzungen:

- korrekte Standardfehler (bei hinreichender Stichprobengröße)
- geringerer Aufwand

Wenn Sie SPSS Statistics gut kennen, mit Amos bisher noch keine Erfahrungen gesammelt haben und ein rekursives Modell mit ausschließlich manifesten Variablen betrachten (wie im Colleges-Beispiel), dann verursacht die von SPSS Statistics angebotene multiple Imputation (siehe Abschnitt 6) bei vergleichbarer Präzision weniger Aufwand als die mit Amos realisierte FIML-Schätzung. Ein FIML-Vorteil gegenüber der multiplen Imputation besteht allerdings darin, dass ein festes Ergebnis existiert, während das Ergebnis einer multiplen Imputation aufgrund der beteiligten Zufallsziehung variabel ist.

5.3.2 Hilfsvariablen

Beim Einsatz der FIML-Schätzmethode in Strukturgleichungsanalysen kann die Erweiterung des Modells um Hilfsvariablen zwei wichtigen Zielen dienen:

- **Verzerrungen beim Schätzen vermeiden bzw. reduzieren**
  Hilfsvariablen mit Einfluss auf die Wahrscheinlichkeit fehlender Werte machen die MAR-Bedingung plausibler. Somit wird die Gefahr verzerrter Schätzer reduziert.
- **Präzision beim Schätzen und Testen steigern**
  Durch Hilfsvariablen mit einer korrelativen Beziehung zu unvollständigen Modellvariablen werden Informationsdefizite kompensiert, was zu einer gesteigerten Präzision führt (kleinere Vertrauensintervalle beim Schätzen, größere Power beim Testen). Enders (2010, S. 131) gibt als Faustregel für die Nützlichkeit von Korrelaten zu MD-belasteten Variablen eine betragsmäßige Korrelationsuntergrenze von 0,4 an.

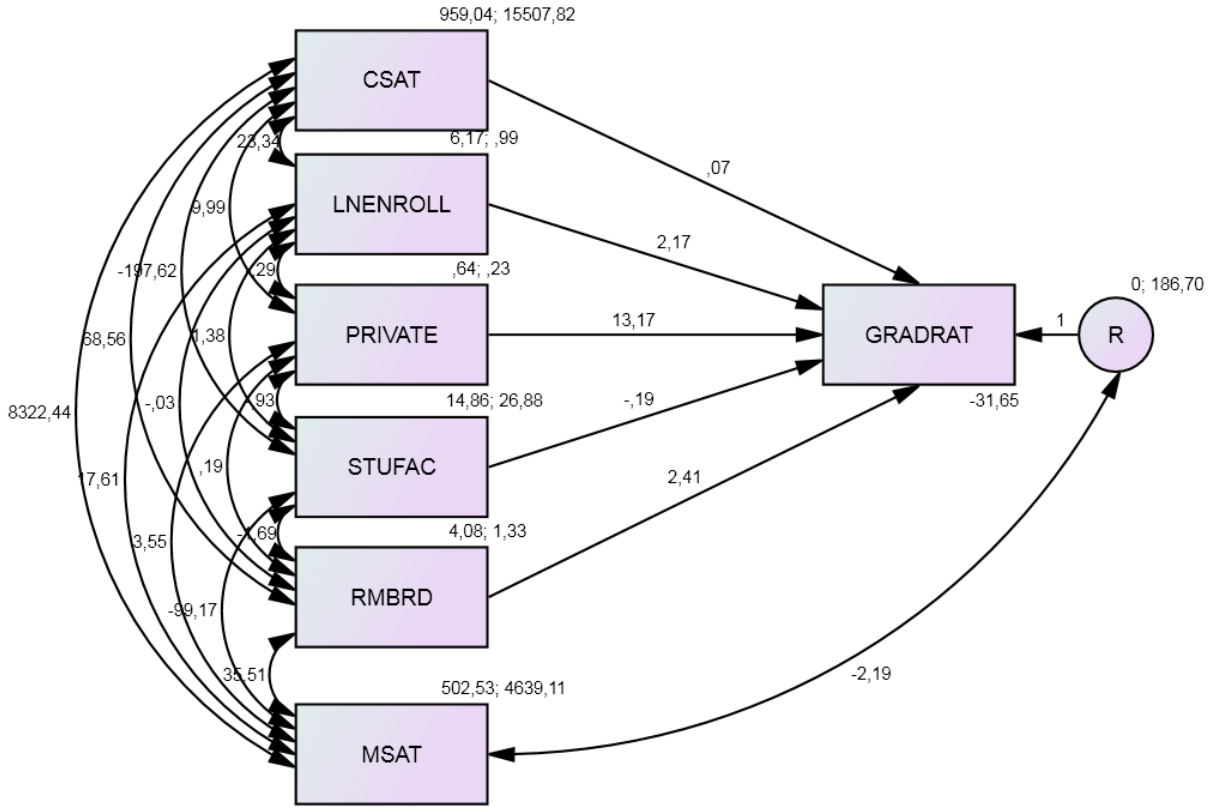
Oft wird auf Hilfsvariablen verzichtet, weil diese in keinem direkten Bezug zum Analysemodell stehen (Acock 2005, S. 2021). Außerdem müssen die Hilfsvariablen mit einem gewissen Aufwand so in ein Strukturgleichungsmodell integriert werden, dass die Bedeutungen der bisher vorhandenen Parameter unverändert bleiben. Bei einer solchen „aussagenneutral" vorgenommenen Erweiterung entsteht in der Terminologie von Enders (2010, S. 134f) ein **Modell mit saturierten Korrelaten** (engl.: *Saturated Correlates Model*). Bei der Aufnahme in ein Modell für manifeste Variablen muss nach Graham (2003) für eine Hilfsvariable jeweils eine frei schätzbare Kovarianz vorgesehen werden ...

- mit jeder anderen Hilfsvariablen
- mit jeder exogenen Modellvariablen
- mit jedem Residuum zu einer endogenen Modellvariablen

Eine Hilfsvariable kann auch dann nützlich sein, wenn sie selbst durch fehlende Werte „geschwächt" ist. Enders (2010, S. 137) fand in einer Simulationsstudie eine Verzerrungsminderung auch bei Hilfsvariablen mit bis zu 50 % fehlenden Werten.

Mit der Anzahl der Hilfsvariablen steigt allerdings auch das Risiko von Schätzproblemen (z.B. gescheiterte Konvergenz, irreguläre Lösungen) (Enders 2010, S. 139).

Wir erweitern im Colleges-Beispiel das Modell um die Hilfsvariable MSAT (mittlere mathematische Leistung der Bewerber), die sehr hoch mit dem Regressor CSAT korreliert (0,98 bei den 776 Fällen mit Werten für beide Variablen). Nach Grahams Regeln sind frei schätzbare Kovarianzen mit allen exogenen Variablen (inkl. Residuum zum Kriterium) erforderlich:



Die Aufnahme der Variablen MSAT hat nur geringe, aber überwiegend günstige Auswirkungen auf die Standardfehler und die Signifikanztests:

Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
GRADRAT	<---	CSAT	,0654	,0049	13,3304	***	
GRADRAT	<---	LNENROLL	2,1700	,6268	3,4621	***	
GRADRAT	<---	PRIVATE	13,1682	1,4167	9,2948	***	
GRADRAT	<---	STUFAC	-,1906	,0947	-2,0129	,0441	
GRADRAT	<---	RMBRD	2,4058	,5641	4,2648	***	

In der folgenden Tabelle werden die mit verschiedenen Analysemethoden ermittelten Standardfehler der Regressionskoeffizienten gegenüber gestellt:

	Fallweise Behandlung (N = 372)	Regr. mit EM-Momenten (N = 1302)	FIML ohne Hilfsvariable (N = 1302)	FIML mit Hilfsvar. MSAT (N = 1302)
CSAT	0,07	0,04	0,005	,005
LNENROLL	1,055	0,533	0,627	0,627
PRIVATE	2,192	1,146	1,419	1,417
STUFAC	0,142	0,084	0,095	0,095
RMBRD	0,784	0,398	0,566	0,564

Die FIML-Methode liefert kleinere Standardfehler als die (bei MAR-Daten unzulässige) fallweise Behandlung und zuverlässigere als die lineare Regression mit EM-geschätzten Verteilungsmomenten.

Auch in ein Modell mit *latenten* Variablen lassen sich saturierte Korrelate zur Linderung von MD-Problemen integrieren (siehe Enders 2010, S. 135ff). Dabei ändert sich nichts am Identifikationszustand des Modells. Beim χ^2 - Modellgültigkeitstest ...

- bleibt die Anzahl der Freiheitsgrade gleich,
- ändert sich aber i.A. die Prüfgröße, was z.B. aufgrund einer Reduktion von MNAR-bedingten Verzerrungen durchaus zu erhoffen ist.

Bei inkrementellen Fit-Indizes führt die Aufnahme von Hilfsvariablen allerdings zu einer Verzerrung in positiver Richtung, also zu einer übertrieben günstigen Bewertung. Ein Beispiel ist der *Comparative Fit Index* (CFI) nach Bentler (1990):

$$\mathrm{CFI} := \frac{\max\left(\chi^2_U - df_U, 0\right) - \max\left(\chi^2_M - df_M, 0\right)}{\max\left(\chi^2_U - df_U, 0\right)}$$

Hier wird für das zu beurteilende Modell (M) und das in der Regel sehr unplausible Referenzmodell kompletter Unabhängigkeit (U) jeweils die Differenz aus dem χ^2 - Fehlerwert und der Freiheitsgradzahl ermittelt, wobei eine (selten zu beobachtende) negative Differenz ggf. durch den Wert 0 ersetzt wird. Wie Enders (2010, S 137f) überzeugend darlegt, wird das Unabhängigkeitsmodell durch die Aufnahme der Hilfsvariablen noch fehlerhafter, so dass der CFI-Quotient notwendigerweise ansteigt. Zur Korrektur schlägt er die Verwendung eines liberalisierten Unabhängigkeitsmodells vor, das alle Kovarianzen mit Beteiligung der Hilfsvariablen zulässt.

5.3.3 Optionen bei ungültiger Normalverteilungsannahme

Alle ML-Verfahren (inkl. EM-Algorithmus, vgl. Abschnitt 5.1) basieren auf der Annahme multivariater Normalverteilung der manifesten Variablen. Ist sie verletzt, bleiben die ML-Schätzer zwar konsistent (asymptotisch erwartungstreu), aber die Standardfehler und der Modellgültigkeitstest werden fehlerhaft (siehe Bollen 1989, S. 415ff). Die ML-Techniken haben sich in Simulationen als robust gegenüber mäßigen Verletzungen der Normalverteilungsvoraussetzung erwiesen (siehe z.B. Allison 2002, S. 32ff).

Will man bei einer erheblichen Verletzung der Normalverteilungsannahme nicht auf die Robustheit der ML-Technologie vertrauen, muss man sich um Alternativen bemühen, wobei Amos bei der Analyse *vollständiger* Daten folgende Optionen bietet:

- ADF-Schätzer (*Asymptotically Distribution Free*)
- Bootstrapping zur empirischen Ermittlung von Stichprobenverteilungen zu den ML-Schätzern
- Bootstrapping nach Bollen & Stine (1993) zur Beurteilung des Modellgültigkeit
  Bei verletzter Normalverteilungsannahme fällt die χ^2-Prüfgröße tendenziell zu groß aus, was zur falschen Entscheidung gegen ein akzeptables Modell führen kann (Kline 2005, S. 157).

Leider stehen diese Techniken bei der Analyse unvollständiger Datensätze (mit fehlenden Werten) *nicht* zur Verfügung. Das ist insbesondere beim Bootstrapping zu bedauern, weil diese Technik bei verletzter Normalverteilungsvoraussetzung sehr präzise Standardfehler für die ML-Schätzer liefern kann (siehe Enders 2010, S. 161).

Die von anderen Strukturgleichungsanalyseprogrammen (z.B. EQS, LISREL, Mplus) unterstützten *robusten Standardfehler* für ML-Schätzer (siehe z.B. Enders 2010, S. 141ff) bietet Amos generell nicht an.

6 Multiple Imputation

Die im Abschnitt 5.3 vorgestellte FIML-Technologie bietet eine elegante Lösung für das Problem fehlender Werte, ist aber nicht bei jeder Fragestellung anwendbar. Ein vervollständigter Datensatz hat den Vorteil, dass *jede* ursprünglich geplante Analysemethode genutzt werden kann.

Wird ein *einzelner* Datensatz mit imputierten Werten analysiert, resultieren geminderte Standardfehler zu den Parametern des Analysemodells und damit eine ungültige Inferenzstatistik. Hier werden imputierte Daten behandelt wie beobachtete, obwohl in den Imputationsmodellen statt der eigentlich benötigten Populationsparameter nur Stichprobenschätzer verfügbar waren, woraus eine zusätzliche Fehlervarianzquelle beim Schätzen der Analysemodellparameter resultiert.

Die von Rubin (1987) vorgeschlagene *multiple* Imputation vermeidet den Fehler der Einfachimputation, indem sie *mehrere* (z.B. 20) komplettierte Datensätze erstellt und parallel auswertet, wobei für einen Parameter des Analysemodells in der Regel verschiedene Schätzwerte resultieren. So kommt die per Imputation eingeschleuste Unsicherheit zum Ausdruck und kann bei der Berechnung gültiger Standardfehler berücksichtigt werden.

6.1 Grundprinzip und Phasen

Eine statistische Analyse mit multipler Imputation fehlender Werte verläuft in drei Phasen:

- **Imputation**
  Zum Komplettieren eines Datensatzes wird für jede MD-belastete Variable ein passendes Imputationsmodell mit Regressionstechnik verwendet. Auf dem Weg zu korrekten Standardfehlern der Analysemodellparameter muss unbedingt die Unsicherheit bzgl. der Parameter der Imputationsmodelle berücksichtigt werden. Dazu erstellt man *mehrere* (z.B. 20) komplettierte Datensätze. Statt für jeden vervollständigten Datensatz in den Imputationsmodellen dieselben Parameterschätzungen zu verwenden, wählt man jeweils zufällig aus der **Verteilung potentieller Parameterausprägungen**. Zur Konstruktion dieser Verteilung verwendet man nach den Prinzipien der **Bayes-Statistik** (siehe Abschnitt 6.2.2) Vorwissen über die Parameter und Informationen aus der beobachteten Stichprobe (Maximum Likelihood - Kalkulationen). Wie in einem iterativen Verfahren eine Zufallsauswahl von potentiellen Ausprägungen der Imputationsmodellparameter zu gewinnen ist, beschreibt der Abschnitt 6.2.4.
- **Analyse**
  In der Analysephase kommen dieselben Methoden zum Einsatz, die man auch bei einem vollständigen Datensatz verwendet hätte. Allerdings werden diese Methoden nun auf *jeden* Imputationsdatensatz angewendet, so dass z.B. 20 Regressionsanalysen mit demselben Analysemodell zurechnen sind, und für jeden Regressionskoeffizienten 20 Parameterschätzungen und 20 geschätzte Standardfehler resultieren.
- **Zusammenfassung** (engl.: *Pooling*)
  Mit jedem vervollständigten Datensatz werden die Parameter des Analysemodells geschätzt. Die Schätzungen zu jedem Parameter werden gemittelt und ergeben so die gesuchte Punktschätzung. Die *M* vervollständigten Datensätzen liefern zu jedem Parameter auch jeweils einen geschätzten Standardfehler. Aus dem Mittelwert der *M* quadrierten Standardfehler und der Varianz der *M* Parameterschätzungen wird ein *korrigierter* Standardfehler berechnet. Damit lassen sich Vertrauensintervalle und Hypothesentests konstruieren (siehe Abschnitt 6.3). So erhält man erwartungstreue Parameterschätzungen sowie korrekte Standardfehler für die Inferenzstatistik. Leider existieren für manche statische Ergebnisse aus den Imputationsstichproben (z.B. Determinationskoeffizient und globaler F-Test der multiplen Regression) noch keine allgemein anerkannten Methoden zur Zusammenfassung.

Der Gesamtaufwand dieser Drei-Phasen-Prozedur ist nicht unerheblich. SPSS Statistics kann seit der Version 17 das Verfahren bei vielen Analysemodellen (z.B. lineare Regression, logistische Regression) weitgehend automatisieren. Außerdem lassen sich die *M* Imputationsstichproben für *mehrere* Analysen verwenden.

Für die in einer Phase zu erledigenden Aufgabe können in relativer Unabhängigkeit von den anderen Phasen Lösungsmethoden entwickelt und angewendet werden (van Buuren 2007, S. 220). Allerdings besteht für die beiden ersten Phasen eine unbedingt zu beachtende Restriktion: Alle im Analysemodell enthaltenen Variablen und statistischen Beziehungen müssen unbedingt auch in den Imputationsmodellen enthalten sein. Andenfalls sorgen die Imputationswerte für eine systematische Schwächung der zu untersuchenden Beziehungen (siehe Abschnitt 6.2.1).

Analysieren verschiedene Arbeitsgruppen dieselben Daten, werden sie in der Regel die vervollständigten Datensätze jeweils selbst erzeugen. Weil bei der multiplen Imputation (Pseudo)zufall im Spiel ist, erhalten die Arbeitsgruppen bei der Anwendung derselben Auswertungsprozedur leicht abweichende Ergebnisse.

Bei der multiplen Imputation wird wie bei der stochastischen Einfachimputation und bei den Abschnitt 5 vorgestellten ML - basierten Verfahren für die fehlenden Daten das **MAR-Modell** voraussetzt.

6.2 Imputationsphase

6.2.1 Zu berücksichtigende Variablen und Beziehungen

Für die in der Imputationsphase zu berücksichtigenden Variablen und Beziehungen gelten analog zu den ML-basierten Verfahren (vgl. Abschnitt 5) folgende Empfehlungen:

- Auf keinen Fall dürfen Variablen des Analysemodells in der Imputationsphase fehlen, weil es in diesem Fall (bei beliebigem MD-Mechanismus!) zu verzerrten Schätzern kommt (Enders 22010, S. 229).
- Alle im Analysemodell enthaltenen statistischen Beziehungen müssen unbedingt auch in den Imputationsmodellen enthalten sein. Andenfalls sorgen die Imputationswerte für eine systematische Schwächung der untersuchten Beziehungen. Wenn z.B. laut Analysemodell der Effekt eines Regressors *X* auf das Kriterium *Y* durch die Variable *Z* moderiert wird, bei der Imputation aber eine multivariate Normalverteilung unterstellt wird, dann resultieren modellkonträre Imputationswerte. Im ungünstigsten Fall wird der untersuchte Interaktionseffekt überdeckt, und alle Investitionen in eine moderne Behandlung fehlender Werte mit dem Ziel, verzerrte Parameterschätzungen zu vermeiden sowie die statistische Präzision (beim Schätzen und Testen) zu steigern, waren nicht nur vergebens, sondern sogar schädlich.
- Es ist oft sinnvoll, in der Imputationsphase neben den Variablen des Analysemodells auch **Hilfsvariablen** zu verwenden (siehe Abschnitt 5.3.2). Es kommen in Frage:
    - Variablen, die das Auftreten fehlender Werte erklären und damit die MAR-Bedingung plausibel machen können.
    - Variablen, die hoch mit lückenhaften Variablen korrelieren, also viel Information über fehlende Werte enthalten.
- Abhängige und unabhängige Variablen aus dem Analysemodel sollten bei der Imputation weitgehend identisch behandelt werden (siehe z.B. Allison 2002, S. 54f; King et al. 2001, S. 56). Es stellt z.B. kein Problem dar, im Imputationsmodell das Kriterium bei der Schätzung von fehlenden Regressorwerten zu verwenden. Weil bei der multiplen Imputation stets eine Residualkomponente addiert wird, kommt es *nicht* zu verfälschten Schätzungen. Wenn ausschließlich Kriteriumswerte fehlen, und keine Hilfsvariablen verfügbar sind, sollte man allerdings auf die Ersetzung fehlender Werte verzichten und die fallweise Behandlung wählen (vgl. Abschnitt 4.3.2).

Beim Erstellen der Imputationsstichproben können Hilfsvariablen mit geringem Aufwand und folglich bei Bedarf auch in größerer Anzahl beteiligt werden. Bei den eigentlichen Analysen spielen sie dann *keine* Rolle mehr, belasten also das Verfahren nicht durch Komplexität. Hier hat die multiple Imputation (wie im Übrigen auch die in Abschnitt 5.1 beschriebene EM-Technik) einen Vorteil gegenüber der FIML-Analyse per Strukturgleichungsprogramm, wo die Hilfsvariablen in das Analysemodell aufgenommen werden müssen (siehe Abschnitt 5.3.2 zum Modell mit saturierten Korrelaten**)**.

6.2.2 Proper Multiple Imputations und Bayes-Statistik

Für *proper multiple imputations* im Sinn von Rubin (1987) genügt es *nicht*, bei der Produktion der vervollständigten Datensätze jeweils *dieselben* Punktschätzungen der Imputationsmodellparameter (z.B. Mittelwerte, Varianzen und Kovarianzen der multivariaten Normalverteilung) zu verwenden und die resultierenden MD-Prognosen jeweils durch frischen Residualzufall zu ergänzen. Dabei würden stichprobenfehlerabhängige Schätzer wie bekannte Populationsparameter behandelt. Die Ungenauigkeit beim Schätzen der Imputationsmodellparameter sorgt für zusätzliche Unsicherheit in den MD-Ersatzwerten, und diese muss in die einzelnen Imputationsdatensätze einfließen, damit schlussendlich eine korrekte Inferenzstatistik für die Parameter des Analysemodells resultiert.

Einen attraktiven Lösungsansatz bietet die Bayes-Statistik, die für Modellparameter (gesammelt im Vektor $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, der Werte aus dem Parameterraum Θ annehmen kann) eine *Verteilung* annimmt und die Modellparameter zusammen mit den Variablen (gesammelt im Vektor $\mathbf{X} = (X_1, X_2, \ldots, X_k)$) in einem gemeinsamen Wahrscheinlichkeitsraum betrachtet.

Eine halbwegs präzise Darstellung der Bayes-Statistik ist technisch recht aufwändig und für anwendungsorientierte Leser eine vermeidbare Belastung. Wer an den Details momentan weniger interessiert ist, sollte eine wichtige Idee und ein technisches Hilfsmittel zur Kenntnis nehmen und dann den Rest von Abschnitt 6.2.2 überspringen:

- Mit dem Bayes-Ansatz lassen sich aus den beobachteten Daten und eventuell vorhandenen Voraannahmen (a-priori - Verteilung) Informationen über die a-posteriori - Verteilung der Parameter gewinnen. Dies ermöglicht es, im Rahmen der multiplen Imputation für jeden einzelnen vervollständigten Datensatz aus der a-posteriori - Verteilung potentieller Parameterausprägungen zufällig zu wählen.
- Die Anwendung der Bayes-Statistik wurde lange durch enorme technische Probleme behindert. In den letzten Jahren hat die so genannte *Markov-Chain-Monte-Carlo* - Technik (MCMC) zusammen mit leistungsfähigen Computer-Systemen einen Durchbruch gebracht.

Beim Bayes-Ansatz zur Parameterschätzung wird generell (also unabhängig vom Problem fehlender Werte) versucht, die **a-posteriori - Verteilung** der Modellparameter, nämlich die *bedingte Verteilung* $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ der Parameter gegeben die beobachtete Datenmatrix **x** zu ermitteln, wobei eine *vor* der Datenerhebung bestehende Annahme über die **a-priori - Parameterverteilung** $\pi_0(\boldsymbol{\theta})$ berücksichtigt wird. Die von den Daten gelieferte Information steckt in der Likelihoodfunktion $\mathrm{L}(\mathbf{x} \mid \boldsymbol{\theta})$. Im Nenner der folgenden **Grundgleichung der Bayes-Schätzung** sorgt ein Normierungsfaktor, der als Randwahrscheinlichkeit der Datenmatrix **x** aufgefasst werden kann, dafür, dass $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ eine Wahrscheinlichkeitsverteilung über dem Parameterraum Θ ist:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\mathrm{L}(\mathbf{x} \mid \boldsymbol{\theta}) \cdot \pi_0(\boldsymbol{\theta})}{\int_{\Theta} \mathrm{L}(\mathbf{x} \mid \boldsymbol{\theta}) \cdot \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

Es resultiert eine auf den ersten Blick abschreckende Formel über eine wichtige Beziehung im gemeinsamen Wahrscheinlichkeitsraum von Parametern und Daten. Weil in allgemeinen *stetige* Verteilungen vorliegen, kommt auch noch die Integrationstheorie ins Spiel. Beim zweiten Blick ist hinter der Formel

noch der famose und zugleich triviale Satz des britischen Mathematikers und Theologen Thomas Bayes (1702-1761) über bedingte Wahrscheinlichkeiten von Ereignissen zu erkennen:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Ausgehend von der Definition der bedingten Wahrscheinlichkeit

$$\mathrm{P}(A \mid B) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

ist der Satz von Bayes mit elementaren Mitteln zu beweisen.

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Als einfaches Anwendungsbeispiel betrachten wir die Schätzung einer Proportion (z.B. Anteil der Befürworter eines Standpunkts in der Population) anhand einer Zufallsstichprobe von N unabhängigen Beobachtungen X_1, X_2, X_N (Zustimmung sei mit Eins kodiert, Ablehnung mit Null). Wenn wir den Anteil der Befürworter in der Population mit γ bezeichnen, folgt die Zufallsvariable S mit der Anzahl zustimmender Äußerungen in der Stichprobe einer Binomialverteilung mit den Parametern γ und N:

$$S := \sum_{i=1}^{N} X_i \sim \mathrm{B}(\gamma, N)$$

In der Bayes-Statistik betrachtet man die gemeinsame Verteilung von S und γ, die das Produkt aus der a-priori - Parameterverteilung $\pi_0(\gamma)$ und der bedingten Verteilung von S bei gegebener Parameterausprägung γ ist. Ist *kein* Vorwissen über die Verteilung des Parameters γ vorhanden, setzt man eine geeignete „Nullinformationsverteilung“ ein. Im Beispiel eignet sich die Gleichverteilung U(0, 1) auf dem Intervall von Null bis Eins:

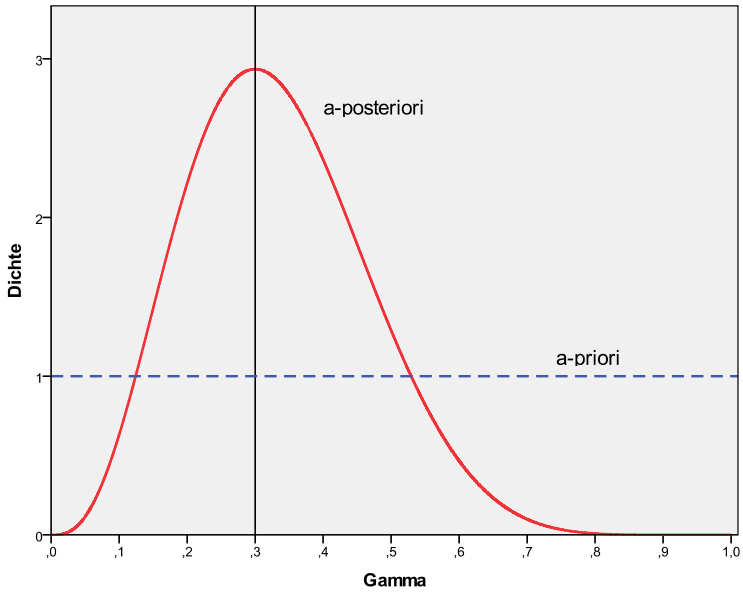
$$\gamma \sim \mathrm{U}(0, 1)$$

Als a-posteriori - Verteilung von γ für ein Stichprobenergebnis $S = s$ ergibt sich (siehe Grynaviski 2003):

$$\pi(\gamma \mid s) = \begin{cases} \dfrac{(n+1)!}{s!(n-s)!}\gamma^{s}(1-\gamma)^{n-s} & 0 \le \gamma \le 1 \\ 0 & \text{sonst} \end{cases}$$

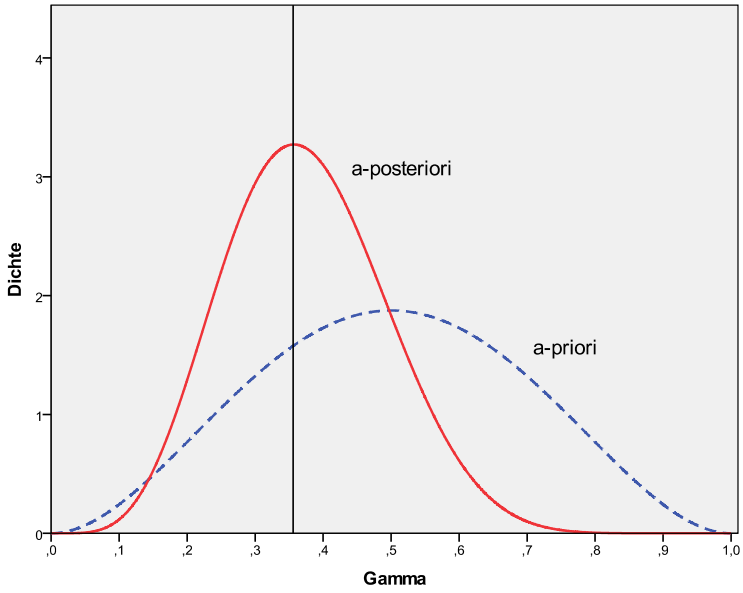
Dies ist eine Betaverteilung mit den Parametern (s + 1) und (N - s + 1).

Bei einem Stichprobenergebnis mit N = 10 und s = 3 erhält man die folgende a-posteriori - Verteilung, wobei der Modus gerade identisch ist mit der Maximum-Likelihood - Punktschätzung für γ (= relative Häufigkeit):



Das Intervall vom 0,025 - bis zum 0,975 - Quantil der a-posteriori - Verteilung (im Beispiel: [0,109; 0,610]) kann man als die Bayes-Variante des Vertrauensintervalls interpretieren (engl.: *Bayes Credible Intervals*).

Wenn man den Parameter γ in der Nähe des Werts 0,5 vermutet und an Stelle der Gleichverteilung U(0, 1) die Beta(3, 3) - Verteilung als a-priori - Verteilung verwendet, resultiert eine andere a-posteriori - Verteilung (Beta(6, 10), nach Grynaviski 2003):



Insbesondere fällt der sich als γ-Punktschätzer anbietende a-posteriori - Modus deutlich größer aus. Speziell bei einer kleinen Stichprobe (wie im Beispiel) kann das in einer a-priori - Verteilung enthaltene Vorwissen über den Parameter die Schätzung verbessern. Bei der multiplen Imputation kommt jedoch meist eine Nullinformations - a-priori - Verteilung zum Einsatz.

Eine gut verständliche generelle Einführung in die Grundgedanken der Bayes-Statistik inklusive Vergleich mit zentralen Begriffen der konventionellen (von den Bayes-Vertretern als *frequentistisch* bezeichneten) Statistik (z.B. Stichprobenkennwerteverteilung, Erwartungswert, Standardfehler, Vertrauensintervall, Hypothesentest) bieten z.B. Arbuckle (2012, S. 385ff) oder Enders (2010, Kap. 6).

6.2.3 Zufallsziehung aus der a-posteriori - Verteilung per Markoff Chain Monte Carlo (MCMC)

Bei weniger trivialen Problemen sind zahlreiche Parameter beteiligt (bei 10 multivariat normalverteilten Variablen z.B. 10 Erwartungswerte, 10 Varianzen und 45 Kovarianzen), und bei der Analyse der a-posteriori - Verteilung machen mehrdimensionale Integrale einen erheblichen Aufwand, der lange Zeit die Verbreitung von Bayes-Methoden behindert hat. Seit einiger Zeit wird die so genannte **Markov-Chain-Monte-Carlo** (**MCMC**) - Technik erfolgreich zur numerischen Analyse der a-posteriori - Verteilung eingesetzt. Über ein iteratives Verfahren gelingt es, eine Zufallsstichprobe aus dieser Verteilung zu gewinnen. Wenn zur Beschreibung einer schwer berechenbaren Verteilung eine Zufallsstichprobe aus dieser Verteilung verwendet wird, spricht man von einer Simulations- oder *Monte-Carlo* - Methode. Aus der Bayes - a-posteriori - Verteilung gewinnt man die Zufallsstichprobe mit Hilfe einer speziellen Zeitreihe aus Zufallsvariablen Z_1, Z_2, Z_3, … (siehe Abschnitt 6.2.4). Weil bei dieser Zeitreihe für die Verteilung der Variablen Z_k aus der Vergangenheit ausschließlich der unmittelbare Vorgänger Z_{k-1} relevant ist, liegt eine so genannte *Markov-Kette* vor.

Bei der multiplen Imputation (siehe Abschnitt 6.2.4) sind wir an mehreren unabhängigen Ziehungen aus der a-posteriori - Verteilung der Imputationsmodellparameter interessiert. Aus jeder gezogenen Parameterrealisation werden Schätzungen für die fehlenden Werte ermittelt (und durch frischen Residualzufall ergänzt). Man kann das Endergebnis auch als Ziehung aus der a-posteriori - Verteilung der fehlenden Werte beschreiben. Die zur multiplen Imputation verwendeten MCMC-Verfahren werden gleich beschrieben.

6.2.4 Imputationsalgorithmen mit MCMC-Technik

Für jeden vervollständigten Datensatz wird eine unabhängige Zufallsziehung aus der a-posteriori - Verteilung der Imputationsmodellparameter gegeben die beobachteten Werte benötigt. Anschließend werden zwei MCMC-Techniken zur Beschaffung solcher Zufallsziehungen vorgestellt:

- JM (*Joint Modeling*) im Normalverteilungsmodell
  Bei diesem Verfahren wird eine gemeinsame Modellierung aller Variablen einer Analyse vorgenommen und zur Erleichterung dieses ambitionierten Unterfangens in der Regel die mathematisch angenehme multivariate Normalverteilung vorausgesetzt, woraus u.a. die Beschränkung auf metrische Variablen resultiert. Dieses Verfahren kann als traditioneller Standard im Rahmen der multiplen Imputation gelten (vgl. Enders 2010, Kapitel 7), wird aber von SPSS Statistics *nicht* verwendet. Daher beschränken wir uns in Abschnitt 6.2.4.1 auf eine kurze Beschreibung.
- FCS (*Fully Conditional Specification*)
  Das von SPSS Statistics seit der Version 17 unterstützte Verfahren behandelt jede Variable mit MD-Problemtatik separat, wobei dem jeweiligen Skalenniveau angemessene Imputationspartialmodelle (z.B. multiple oder logistische Regression) Verwendung finden. Weil in jedem Iterationsschritt die Variablen nacheinander behandelt werden, wird auch die Bezeichnung *Chained Equations* verwendet. Aus *Multiple Imputation by Chained Equations* ergibt sich die leicht zu merkende Kurzbezeichnung *MICE*.

6.2.4.1 JM-Imputation mit Normalverteilungsmodell und IP-Algorithmus (Data Augmentation)

In der Imputationsphase wird (für jeden von den *M* vervollständigten Datensätzen) eine unabhängige Zufallsziehung aus der a-posteriori - Verteilung der Imputationsmodellparameter benötigt, um Schätzungen (bedingte Erwartungen) der fehlenden Werte zu liefern, die dann noch ein stochastisches Residuum erhalten. Die angenommenen Parameterwerte müssen aus einer potentiell möglichen multivariaten Verteilung der Analysevariablen stammen. Dies ist gewährleistet, wenn im Verlauf der MCMC-Iterationen mit der gemeinsamen Verteilung aller Analysevariablen gearbeitet wird. Lange Jahre hat bei den meisten mathematischen Ausarbeitungen und Software-Implementationen der multiplen Imputation der JM-Ansatz

(*Joint Modeling*) dominiert. Um die enorme Komplexität zu beherrschen, wurde in der Regel die **multivariate Normalverteilung** aller Variablen unterstellt.

Im Normalverteilungsmodell liefert die lineare Regression für jede Variable die bedingten Erwartungswerte gegeben eine Teilmenge der restlichen Variablen, wobei der Prognosefehler mit homogener Varianz normalverteilt ist. Man gewinnt also für jedes MD-Muster die Imputationswerte über eine lineare Regression und Addieren einer normalverteilten Zufallskomponente mit der geschätzten Residualvarianz. Die benötigten Regressionsparameter lassen sich direkt aus den Normalverteilungsparametern (Mittelwerten, Varianzen und Kovarianzen) berechnen.

Das Verfahren zur Produktion multipler Imputationen bei unterstellter gemeinsamer Normalverteilung aller Variablen wird als *Imputation-Posterior - Algorithmus* (Enders 2010, S. 190ff; King et al 2001) oder als *Data Augmentation* bezeichnet (z.B. Allison 2002, S. 32ff; Schafer 1997). Zunächst werden Startwerte für die Normalverteilungsparameter ermittelt. Dann folgen Iterationen, die jeweils aus einem I- und einem P-Schritt bestehen:

1. **Imputation**
   Für jedes Muster fehlender Werte wird die Regressionsprognose ermittelt unter Verwendung der beobachteten Werte und der aktuellen Normalverteilungsparameter ($\tilde{\mu}$, $\tilde{\Sigma}$), die nach den üblichen Formeln die benötigten Regressionskoeffizienten liefern. Zur Regressionsprognose wird ein zufällig aus der passenden Normalverteilung gewähltes Residuum addiert.
2. **Posterior-Parameterziehung**
   Für die Normalverteilungsparameter ($\tilde{\mu}$, $\tilde{\Sigma}$) werden neue Werte aus ihrer a-posteriori - Verteilung gegeben die vervollständigte Datenmatrix gezogen.

Dann beginnt die nächste Iteration mit der einer neuen Imputation unter Verwendung der aktuellen Normalverteilungsparameter (Teilschritt 1). Nach einer Einbrennphase erreicht der Algorithmus eine Verteilungskonvergenz und liefert brauchbare Imputationswerte. Zwischen zwei Entnahmen für verschiedene Imputationsdatensätze müssen hinreichend viele Iterationsschritte liegen, um die Unabhängigkeit zu gewährleisten. Enders (2010, S. 211) empfiehlt für die Einbrennphase und für den Abstand zwischen zwei Entnahmen jeweils 200 Iterationen.

Bei der multiplen Imputation im Normalverteilungsmodell per IP-Algorithmus verwendet man in der Regel eine Nullinformations - a-priori - Verteilung, wobei wir uns technische Details hier sparen (siehe z.B. Allison 2002, S. 35f).

Insbesondere bei Beteiligung von dichotomen oder ordinalen Variablen ist die Annahme einer gemeinsamen multivariaten Normalverteilung unplausibel. Aufgrund von Simulationsstudien gehen viele Autoren davon aus, dass eine Verletzung der Normalitätsannahme in Abhängigkeit vom Ausmaß, von der Stichprobengröße und vom Anteil fehlender Werte oft unschädlich ist (z.B. Allison 2002, S. 48; Enders 2010, S. 259; King et al 2001, S. 53; Schafer 1997). Man diskutiert eher darüber, ob und wie bei dichotomen oder ordinalen Variablen die Imputationswerte gerundet werden sollen (siehe z.B. Enders 2010, 261ff).

6.2.4.2 FCS-Imputation mit CE-Algorithmus (Chained Equations)

Beim FCS-Algorithmus (*Fully Conditional Specification*) wird nicht versucht, *ein* Imputationsmodell für die gemeinsame Verteilung aller Variablen zu erstellen. Stattdessen wird für jede Variable ein selbständiges Imputationsmodell in Abhängigkeit vom Messniveau gewählt, wobei in der Regel zum Einsatz kommen:

- bei metrischen Variablen ein lineares Regressionsmodell
- bei kategorialen Variablen ein logistisches Regressionsmodell

Mit der Bezeichnung *Fully Conditional Specification* wird zum Ausdruck gebracht, dass für jede zu behandelnde Variable eine bedingte Verteilung gegeben die restlichen Variablen bestimmt wird. Potentiell besteht das Risiko, dass die bedingten Verteilungen der isolierten Imputationsmodelle zu den einzelnen Variablen **inkompatibel** sind, wobei Konvergenzprobleme beim MCMC-Prozess zu erwarten sind. Im Unterschied zum JM-Ansatz (*Joint Modeling*) ist also bei der FCS-Methode nicht garantiert, dass die imputierten Daten aus einer realen multivariaten Verteilung stammen. In Simulationsstudien hat sich aber gezeigt, dass in der Regel keine Probleme zu erwarten sind (van Buuren & Groothuis-Oudshoorn 2011, S. 7).

Als Vorteile der FCS-Technik sind zu nennen:

- Flexibilität der Modellierung
  Bei den Imputationsmodellen zu den einzelnen Variablen gibt es kaum Einschränkungen. Im JM-Ansatz ist es hingegen schwer, eine gemeinsame Verteilung für alle Variablen zu finden. Meist wird die mathematisch angenehme multivariate Normalverteilung verwendet, die jedoch Einschränkungen auferlegt:
  - Beschränkung auf lineare und additive Modelle
  - unbefriedigende Behandlung von Variablen mit diskreter Verteilung
- Auch der FCS-Algorithmus kann als MCMC-Verfahren beschrieben werden, doch werden meist nur 5-10 Iterationen für jeden vervollständigten Datensatz benötigt (van Buuren & Groothuis-Oudshoorn 2011, S. 7), während für die JM-Technik z.B. von Enders (2010, S. 211) ca. 200 Iterationen vor dem ersten Datensatz und zwischen zwei unabhängigen Datensätzen empfohlen werden.
- Weil zum Imputieren fehlender Werte bei kategorialen Variablen eine logistische Regressionsanalyse durchgeführt wird, entfällt die Frage nach dem Runden von Imputationswerten.

Die JM- und die FCS-Imputation haben als wichtige Gemeinsamkeit, dass für jeden vervollständigten Datensatz eine unabhängige Ziehung von Verteilungsparametern aus der a-posteriori - Verteilung im Sinne der Bayes-Theorie vorgenommen wird.

Auch verwenden beide Verfahren einen iterativen Algorithmus, der als MCMC-Prozess beschrieben werden kann. Der auffälligste Unterschied zum IP-Algorithmus der JM-Imputation (siehe Abschnitt 6.2.4.1), besteht beim Algorithmus der FCS-Imputation darin, dass die die fehlerbelasteten Variablen sukzessiv behandelt werden. Wir verwenden daher die Bezeichnung **CE-Algorithmus** (*Chained Equations*).

Beim CE-Algorithmus werden die fehlenden Werte initial durch einfache Ziehungen aus den beobachteten Randverteilungen der Variablen ersetzt. Dann startet ein iteratives Verfahren, wobei in jedem Schritt die Variablen *sukzessive* behandelt werden (siehe z.B. bei van Buuren & Groothuis-Oudshoorn 2011, S. 6ff). SPSS Statistics orientiert sich per Voreinstellung an der Anzahl fehlender Werte und startet mit der komplettesten Variablen. Wir gehen anschließend der Einfachheit halber davon aus, dass die Variablen gemäß ihrem MD-Belastungsrangplatz nummeriert sind. Im Iterationsschritt t werden ...

- für die Parameter $\boldsymbol{\theta}^{(1)}$ des Imputationsmodells zu X_1 neue Realisationen aus ihrer a-posteriori - Verteilung gegeben die vervollständigte Datenmatrix aus dem Iterationsschritt (t-1) gezogen,
- mit Hilfe des neuen Parametervektors und der vervollständigten Datenmatrix aus dem Iterationsschritt (t-1) werden neue Imputationen für die fehlenden Werte in X_1 ermittelt.

Dann wird die Variable X_2 analog behandelt, wobei die frischen Imputationswerte von X_1 bereits Verwendung finden, usw.

Ist die Konvergenz gegen die a-posteriori - Verteilung der Imputationsmodellparameter gelungen, können die zugehörigen Imputationswerte verwendet werden.

Im Unterschied zur JM-Imputation kommt die Konvergenz bei der FCS-Imputation meist schnell zustande. Während man bei der JM-Imputation alle Imputationsdatensätze aus einem Prozess extrahiert (mit

ausreichendem Abstand zwischen zwei Entnahmen) verwendet man bei der FCS-Imputation für jeden Imputationsdatensatz einen eigenen Prozess.

6.2.5 Technische Details

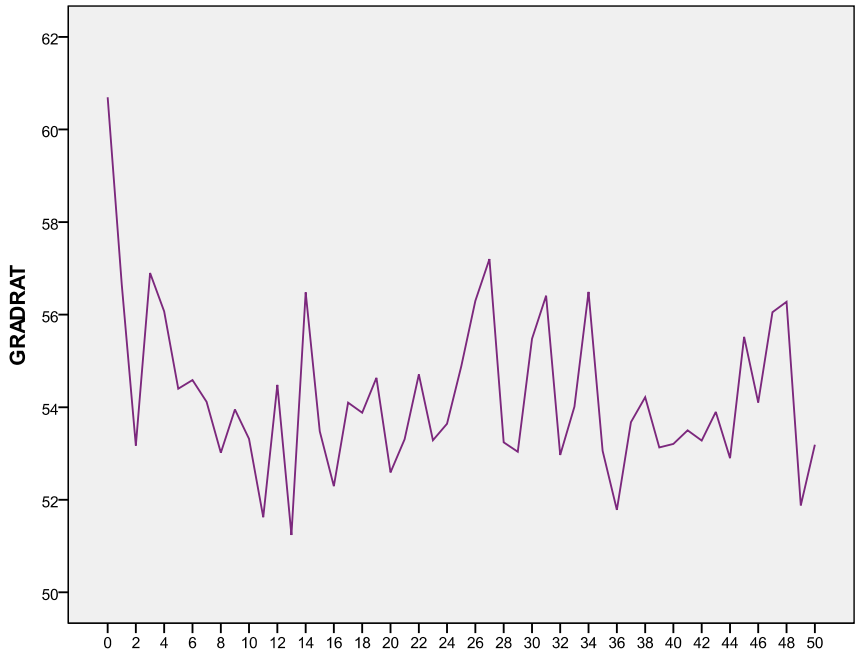
6.2.5.1 Anzahl der Imputationsstichproben

Für die multiple Imputation werden mehrere *unabhängige* Ziehungen aus der a-posteriori - Verteilung der fehlenden Werte gegeben die beobachteten Werte benötigt. Enders (2010, S. 214, 232) empfiehlt zugunsten einer hohen Teststärke bei den Signifikanztests zu den Modellparametern, mindestens 20 Imputationsstichproben zu verwenden. Im Zusammenhang mit Rubins Regeln zur Kombination der multiplen Schätzergebnisse ist zu erkennen, wie sich die Anzahl der Imputationsstichproben auf die Größe der kombinierten Standardfehler auswirkt (siehe Abschnitt 6.3).

6.2.5.2 Konvergenz

Bei einem MCMC-Prozess mit JM-Imputation (z.B. im Normalverteilungsmodell) werden relativ viele Iterationen bis zur Verteilungskonvergenz benötigt (manchmal Tausende) und man verwendet oft aus Zeitgründen für alle Imputationsstichproben denselben Prozess, wobei aber hinreichend viele Takte zwischen zwei Ziehungen liegen müssen, damit dieses als unabhängig betrachtet werden können (zu beurteilen über die Autokorrelationsfunktion). Ein MCMC-Prozess mit FCS-Imputation erreicht hingegen die Verteilungskonvergenz viel schneller, so dass man für jede Imputationsstichprobe einen eigenen MCMC-Prozess verwendet.

Grundsätzlich ist für alle Imputationsmodellparameter die Konvergenz in jeder Imputationsstichprobe zu prüfen. SPSS liefert im Iterationsprotokoll (siehe Abschnitt 6.4.1) aber ausschließlich Informationen zu den Mittelwerten und Standardabweichungen der metrischen Variablen. Bei einem FCS-Prozess stellt sich in der Regel schnell ein unauffällig schwankender Plot ohne erkennbaren Trend ein, z.B.:



6.3 Kombination der multiplen Schätzergebnisse

6.3.1 Rubins Regeln

Aus den *M* vervollständigten Datensätzen gewinnt man für einen Analysemodellparameter θ die *M* Punktschätzungen $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_M$ sowie *M* geschätzte Varianzen der Stichprobenkennwerteverteilung (quadrierte Standardfehler) $\hat{\sigma}^2_1(\theta), \hat{\sigma}^2_2(\theta), ..., \hat{\sigma}^2_M(\theta)$. Nach dem Vorschlag von Rubin (1987) verwendet man den Mittelwert der *M* Schätzungen als **MI-Schätzung** von θ:

$$\hat{\theta}_{MI} := \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

und ermittelt den zugehörigen **MI-Standardfehler** folgendermaßen:

$$\hat{\sigma}(\hat{\theta}_{MI}) := \sqrt{V_W + (1+\frac{1}{M})V_B}$$

Dabei steht V_W für den Mittelwert der datensatzinternen Varianzschätzungen

$$V_W := \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2_m(\theta)$$

und V_B für die datensatzübergreifende Varianz der $\hat{\theta}_m$-Schätzungen:

$$V_B := \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{MI})^2$$

Wenn die Anzahl *M* von Imputationsstichproben wächst, schrumpft der Standardfehler $\hat{\sigma}(\hat{\theta}_{MI})$, und die damit konstruierten Signifikanztests (siehe nächsten Abschnitt) gewinnen an Teststärke. Daher empfiehlt Enders (2010, S. 214, 232), mindestens 20 Imputationsdatensätze zu verwenden.

6.3.2 Tests zu einzelnen Parametern

Die Statistik

$$\frac{\hat{\theta}_{MI} - \theta}{\hat{\sigma}(\hat{\theta}_{MI})}$$

ist approximativ t-verteilt mit

$$\nu_M := (M-1)\left[1 + \frac{V_W}{(1+\frac{1}{M})V_B}\right]^2$$

Freiheitsgraden und ermöglicht daher einen Test zum Hypothesenpaar:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

Dieser Test ist das direkte Analogon zum Wald-Test in Strukturgleichungsmodellen mit ML-Technik (Enders 2010, S. 231).

6.3.3 Durch fehlende Werte bedingter Präzisionsverlust bei der Parameterschätzung

Es sind verschiedene Quantifizierungen für den durch fehlende Werte bedingten Präzisionsverlust bei der Parameterschätzung vorgeschlagen worden.

6.3.3.1 Relativer Anstieg der Varianz

Mit dem folgendermaßen definierten relativen Anstieg der Varianz (engl.: *Relative Increase of Variance*, *RIV*)

$$RAV := \frac{(1+\frac{1}{M})V_B}{V_W}$$

wird für jeden Parameter im Analysemodell die durch fehlende Werte bedingte zusätzliche Unsicherheit im Stichprobenschätzer ins Verhältnis gesetzt zur normalen Unsicherheit aus einer vollständigen Stichprobe.

6.3.3.2 Anteil fehlender Informationen

Der folgende Ausdruck gibt für einen Parameter des Analysemodells an, welcher Anteil seiner Stichprobenvarianz durch fehlende Werte bedingt ist (Formel nach Enders 2010, S. 225):

$$AFI := \frac{(1+\frac{1}{M})V_B}{V_W + (1+\frac{1}{M})V_B}$$

Man spricht hier vom *Anteil der fehlenden Informationen* (engl.: *Fraction of Missing Information*, *FMI*).

6.3.3.3 Relative Effizienz

Mit der relativen Effizienz wird beurteilt, wie effektiv die auf *M* Imputationen basierende Schätzung im Vergleich zum theoretischen Optimum einer auf unendlich vielen Imputationen basierenden Schätzung bereits ist:

$$RE := \frac{1}{1+\frac{AFI}{M}}$$

Das Ergebnis hängt wesentlich vom Anteil fehlender Informationen ab (siehe Abschnitt 6.3.3.2).

6.3.4 Mehrparametertests

Zur Prüfung von komplexeren Hypothesen (z.B. über die Identität zweier Parameter) sind mehrere Teststatistiken vorgeschlagen worden, die jedoch von vorhandenen Programmen schlecht unterstützt werden und/oder sehr große Stichproben erfordern (siehe Allison 2002, S. 65ff; Enders 2010, S. 233ff), so dass auf eine Darstellung verzichtet wird.

6.4 Beispiel

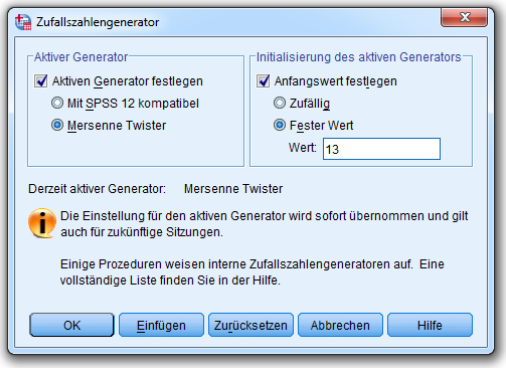
6.4.1 Imputationsstichproben erstellen

6.4.1.1 Pseudozufallszahlengenerator präparieren

Bei der Erstellung von Imputationsstichproben ist der Pseudozufallszahlengenerator beteiligt. Sollen die Daten reproduzierbar sein, setzt man für den Generator nach dem Menübefehl

Transformieren > Zufallszahlengeneratoren

den Typ (kompatibel mit alten SPSS-Versionen oder Mersenne Twister) und einen beliebig wählbaren Startwert fest:

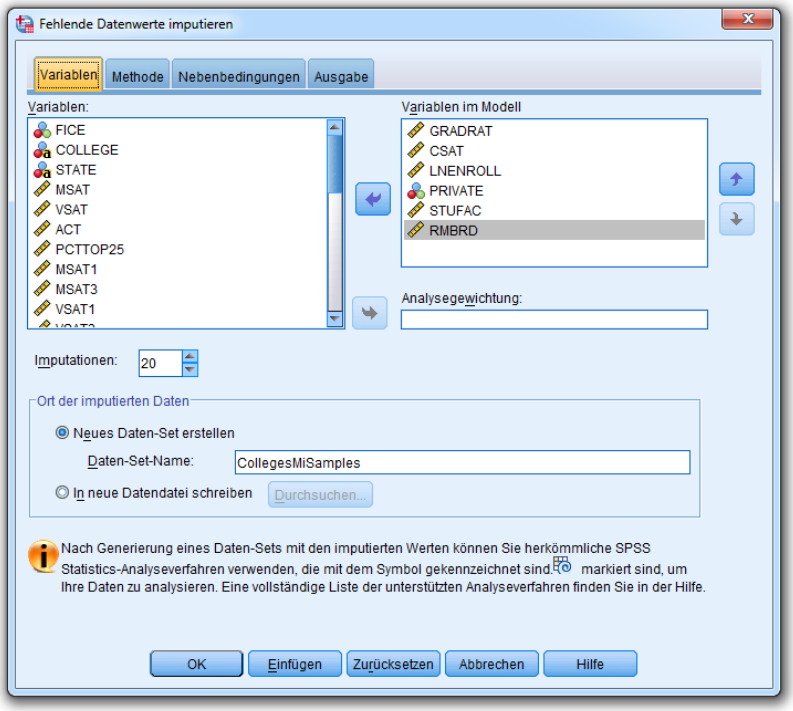


6.4.1.2 Multiple Imputation anfordern

Wir fordern für das in Abschnitt 3.1 beschriebene Colleges-Beispiel (zunächst noch ohne Beteiligung von Hilfsvariablen) über den Menübefehl

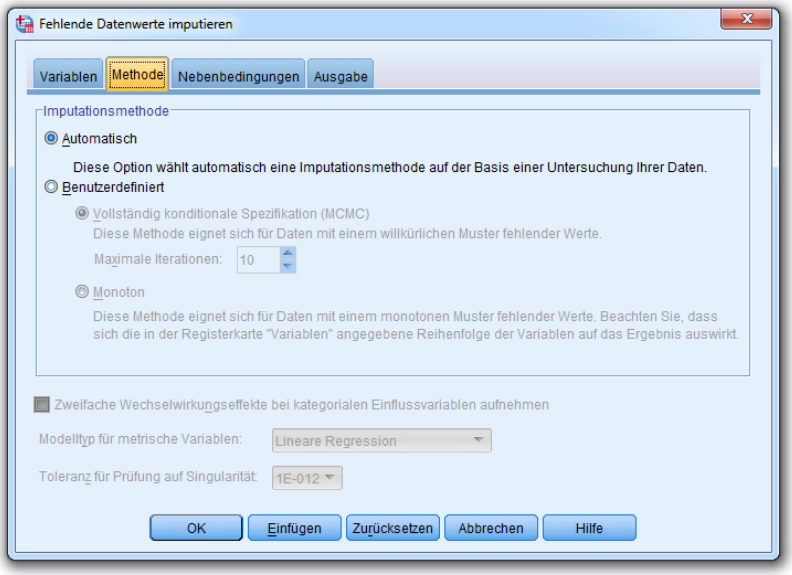
Analysieren > Multiple Imputation > Fehlende Datenwerte imputieren

in der folgenden Dialogbox



das Erstellen von 20 Imputationsstichproben in einem **neuen Daten-Set** an, das den Namen **CollegesMiSamples** erhält. Weil SPSS Statistics bei der multiplen Imputation die Messniveaus der Variablen berücksichtigt, müssen die Deklarationen eventuell im Dateneditor kontrolliert bzw. korrigiert werden.

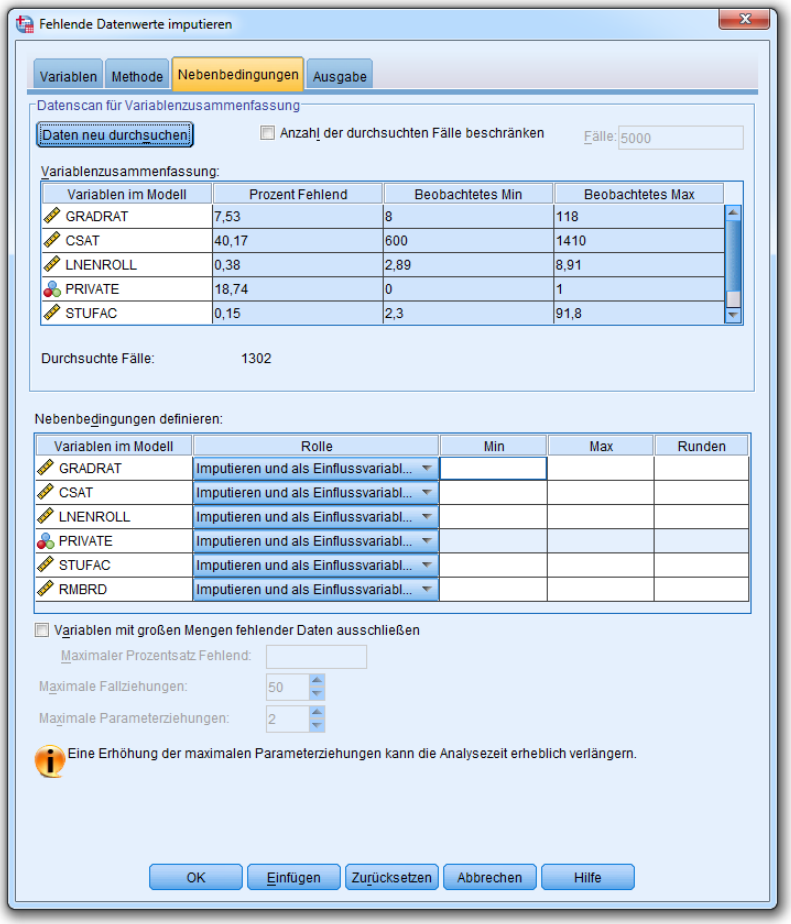
Wird auf der Registerkarte **Methode** die **Imputationsmethode** vom **Benutzer definiert**, sind einige Einstellungen modifizierbar (z.B. die Anzahl der MCMC-Iterationen für den FCS-Algorithmus (**vollständig konditionale Spezifikation**)). Allerdings geht auf diese Weise im FCS-Algorithmus die sinnvolle Imputationssequenz nach aufsteigender Anzahl fehlender Werte verloren (siehe unten), so dass man entweder die Automatik beibehalten oder die Variablen (auf dem Registerblatt **Variablen**) in der gewünschten Imputationssequenz angeben muss. In unserem Beispiel ist gegen die Automatik nichts einzuwenden:



Weil nur *eine* kategoriale Variable beteiligt ist (PRIVATE), müssen keine **Wechselwirkungseffekte bei kategorialen Einflussvariablen** einbezogen werden. Leider werden Interaktionen unter Beteiligung metrischer Variablen nicht unterstützt. Wenn ein Analysemodell solche Interaktionen enthält, liegt nach den Empfehlungen aus Abschnitt 6.2.1 über die in Imputationsmodellen zu berücksichtigenden Variablen und Beziehungen ein ernst zu nehmendes Problem vor.

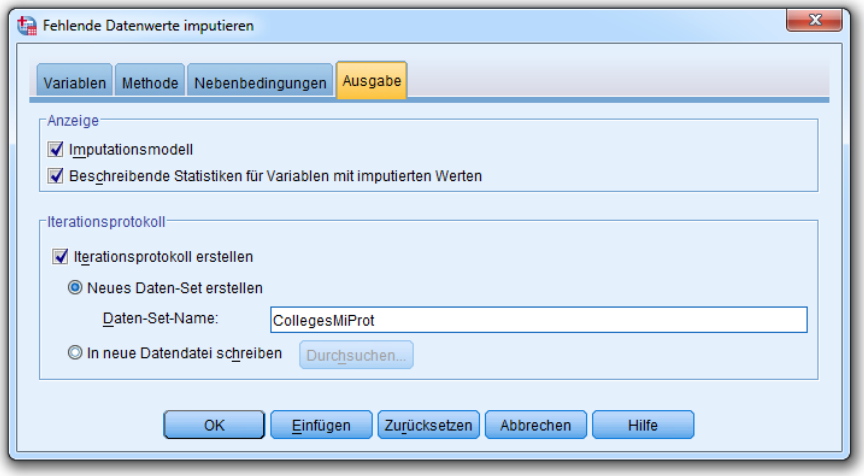
Auf der Registerkarte **Nebenbedingungen** kann man die **Daten durchsuchen** lassen und für einzelne Variablen

- die erlaubten Rollen festlegen
- den Wertebereich für imputierte Werte beschränken
- das Runden der imputierten Werte veranlassen.



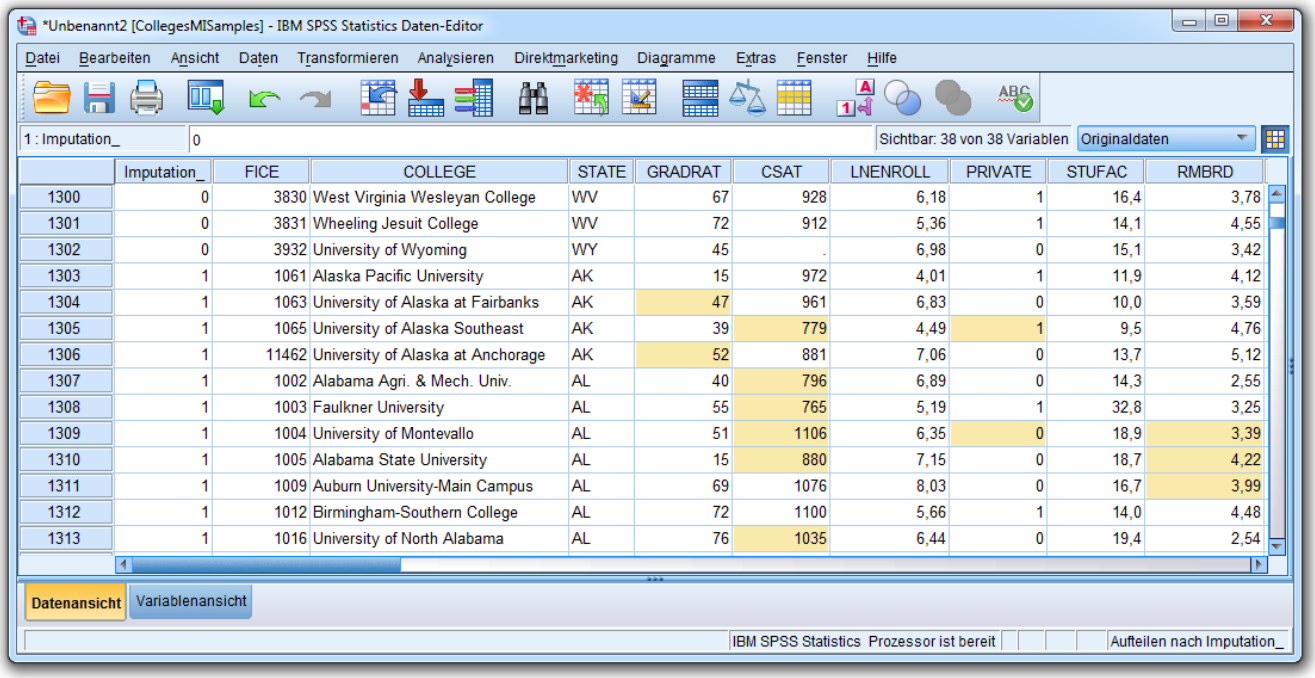
Bei der Variablen GRADRAT fällt das **beobachtete Max** von 118 auf. Dieser Wert gehört zum Cazenovia College, wo offenbar die Anzahl der Graduierten höher war als die Anzahl der Einsteiger vier Jahre zuvor. Nicht minder erstaunlich ist das **beobachtete Min** von 8, das zur University of Houston - Downtown gehört.

Auf der Registerkarte **Ausgabe** verlangen wir die zusätzliche Ausgabe von **beschreibende Statistiken für Variablen mit imputierten Werten** und ordern außerdem ein **Iterationsprotokoll** zur Konvergenzuntersuchung, das in ein neues Datenblatt geschrieben werden soll:



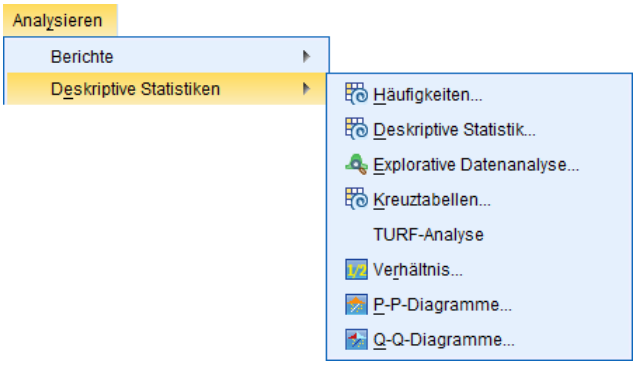
6.4.1.3 Ergebnisdatenblatt der multiplen Imputation

Nach Ausführung der multiplen Imputation erhalten wir ein Mehrstichproben-Datenblatt mit der Originalstichprobe am Anfang und 20 dahinter liegenden Imputationsstichproben. Für Ordnung sorgt die Variable **Imputation_** mit dem Wert 0 für die Originalstichprobe und den Werten 1 bis 20 für die Imputationsstichproben. Es ist eine Fallaufteilung nach der Variablen **Imputation_** aktiv (siehe Statuszeile).



	Imputation_	FICE	COLLEGE	STATE	GRADRAT	CSAT	LNENROLL	PRIVATE	STUFAC	RMBRD
1300	0	3830	West Virginia Wesleyan College	WV	67	928	6,18	1	16,4	3,78
1301	0	3831	Wheeling Jesuit College	WV	72	912	5,36	1	14,1	4,55
1302	0	3932	University of Wyoming	WY	45	.	6,98	0	15,1	3,42
1303	1	1061	Alaska Pacific University	AK	15	972	4,01	1	11,9	4,12
1304	1	1063	University of Alaska at Fairbanks	AK	47	961	6,83	0	10,0	3,59
1305	1	1065	University of Alaska Southeast	AK	39	779	4,49	1	9,5	4,76
1306	1	11462	University of Alaska at Anchorage	AK	52	881	7,06	0	13,7	5,12
1307	1	1002	Alabama Agri. & Mech. Univ.	AL	40	796	6,89	0	14,3	2,55
1308	1	1003	Faulkner University	AL	55	765	5,19	1	32,8	3,25
1309	1	1004	University of Montevallo	AL	51	1106	6,35	0	18,9	3,39
1310	1	1005	Alabama State University	AL	15	880	7,15	0	18,7	4,22
1311	1	1009	Auburn University-Main Campus	AL	69	1076	8,03	0	16,7	3,99
1312	1	1012	Birmingham-Southern College	AL	72	1100	5,66	1	14,0	4,48
1313	1	1016	University of North Alabama	AL	76	1035	6,44	0	19,4	2,54

Imputierte Werte sind farblich gekennzeichnet. Ist ein solches Imputations-Datenblatt die Arbeitsdatei (das aktive Datenblatt), dann sind im **Analysieren**-Menü die Verfahren mit einer Unterstützung für die multiple Imputation am Symbol zu erkennen, z.B.:



6.4.1.4 Ausgaben zu den Imputationsmodellen und -ergebnissen

In der folgenden Ausgabetabelle

Imputationsergebnisse

Imputationsmethode		Vollständig konditionale Spezifikation
Iterationen der vollständig konditionalen Spezifikationsmethode		10
Abhängige Variablen	Imputiert	GRADRAT,CSAT,LNENROLL,PRIVATE, STUFAC,RMBRD
	Nicht imputiert (zu viele fehlende Werte)	
	Nicht imputiert (keine fehlenden Werte)	
Imputationssequenz		STUFAC,LNENROLL,GRADRAT,PRIVATE, RMBRD,CSAT

sind dokumentiert:

- verwendete Imputationsmethode
- (nicht) behandelte Variablen
- Imputationssequenz (vgl. Abschnitt 6.2.4.2)
  Im Beispiel sind die Variablen nach aufsteigender Anzahl fehlender Werte geordnet. Wird auf dem Registerblatt **Methode** des Dialogs zur multiplen Imputation die Automatik abgeschaltet (siehe oben), wird stattdessen die Reihenfolge aus dem Anforderungsdialog verwendet.

Anschließend werden die **Imputationsmodelle** zu den behandelten Variablen beschrieben. Als **Typ** kommt bei metrischen Variablen die lineare Regression und bei kategorialen Variablen die logistische Regression zum Einsatz. Als Regressoren (**Effekte**) werden jeweils alle anderen Variablen verwendet:

Imputationsmodelle

	Modell		Fehlende Werte	Imputierte Werte
	Typ	Effekte		
STUFAC	Lineare Regression	PRIVATE,LNENROLL,GRADRAT,RMBRD,CSAT	2	40
LNENROLL	Lineare Regression	PRIVATE,STUFAC,GRADRAT,RMBRD,CSAT	5	100
GRADRAT	Lineare Regression	PRIVATE,STUFAC,LNENROLL,RMBRD,CSAT	98	1960
PRIVATE	Logistische Regression	STUFAC,LNENROLL,GRADRAT,RMBRD,CSAT	244	4880
RMBRD	Lineare Regression	PRIVATE,STUFAC,LNENROLL,GRADRAT,CSAT	519	10380
CSAT	Lineare Regression	PRIVATE,STUFAC,LNENROLL,GRADRAT,RMBRD	523	10460

In der Tabelle erscheinen die Variablen in der Imputationssequenz.

Insbesondere bei Variablen mit einem beschränkten Wertebereich lohnt sich ein Blick auf die Tabellen mit **deskriptiven Statistiken**. Im Beispiel zeigt sich für die Variable GRADRAT (Anteil der Graduierten) in einem Imputationsdatensatz ein irregulärer Prozentwert unter 0:

GRADRAT

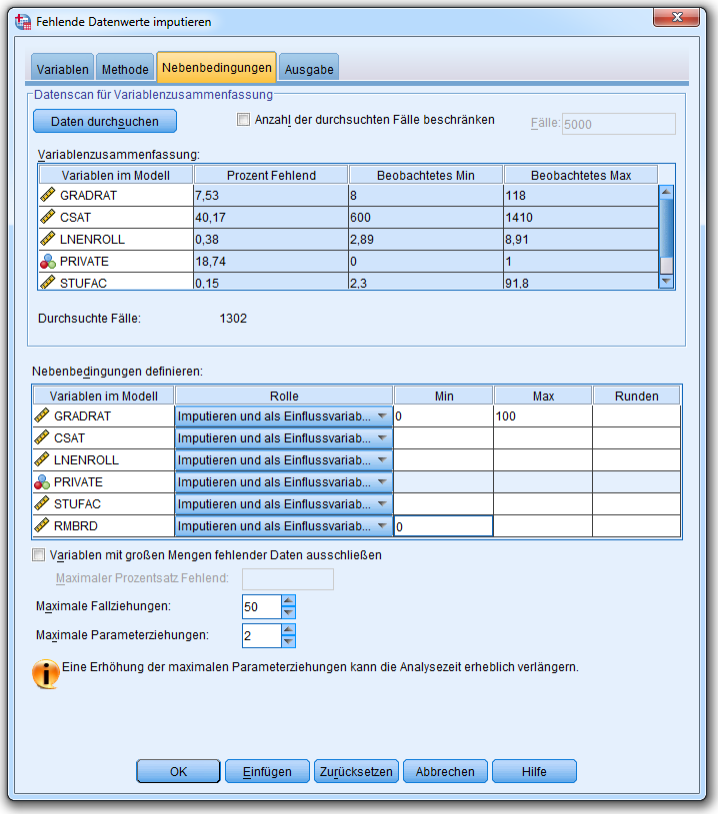
Daten: Imputierte Werte

	N	Mittelwert	Standardabweichung	Minimum	Maximum
1	98	56,89	19,638	3,66	104,69
2	98	56,48	16,931	10,10	107,22
3	98	53,43	15,672	19,64	106,91
4	98	55,31	19,474	11,71	100,83
5	98	52,56	15,686	9,18	79,50
6	98	55,12	17,000	16,55	95,38
7	98	56,48	16,840	17,37	105,28
8	98	53,94	18,567	10,74	99,90
9	98	54,88	18,548	10,21	108,69
10	98	55,40	19,091	-10,01	97,57
11	98	52,69	17,208	7,16	91,30
12	98	55,05	17,243	13,17	109,23
13	98	54,62	18,846	10,61	99,49
14	98	53,95	16,735	7,96	99,76
15	98	51,84	18,584	1,83	97,64
16	98	54,54	16,050	5,90	99,75
17	98	54,50	18,576	7,42	103,61
18	98	54,63	17,522	21,56	95,30
19	98	54,22	17,637	7,80	91,73
20	98	55,09	17,984	7,93	101,21

Das Maximum liegt zwar mehrfach über 100, aber nie über dem beobachteten Maximum von 118 (siehe oben).

6.4.1.5 Nebenbedingungen für Imputationswerte formulieren

Wir aktivieren das Datenblatt mit den Originaldaten und öffnen erneut den Dialog zur multiplen Imputation fehlender Werte. Auf dem Registerblatt **Nebenbedingungen** verhindern wir negative GRADRAT-Werte sowie negative Investitionen (Variable RMBRD):



Ist eine Wertbegrenzung verletzt, wird ein neuer Fall gezogen. Das Verfahren wiederholt sich nötigenfalls, bis ein zulässiger Fall gefunden oder die angegebene Maximalzahl von **Fallziehungen** erreicht ist. Nach einer erfolglosen Suche werden neue Parameter aus der a-posteriori - Verteilung gezogen und zur Suche nach einem zulässigen Fall verwendet. Auch die Parameterziehung wird nötigenfalls wiederholt, bis ein zulässiger Fall gefunden oder die angegebene Maximalzahl von **Parameterziehungen** erreicht ist. Scheitert das Verfahren mit einer Fehlermeldung wie im folgenden Beispiel,

Warnungen

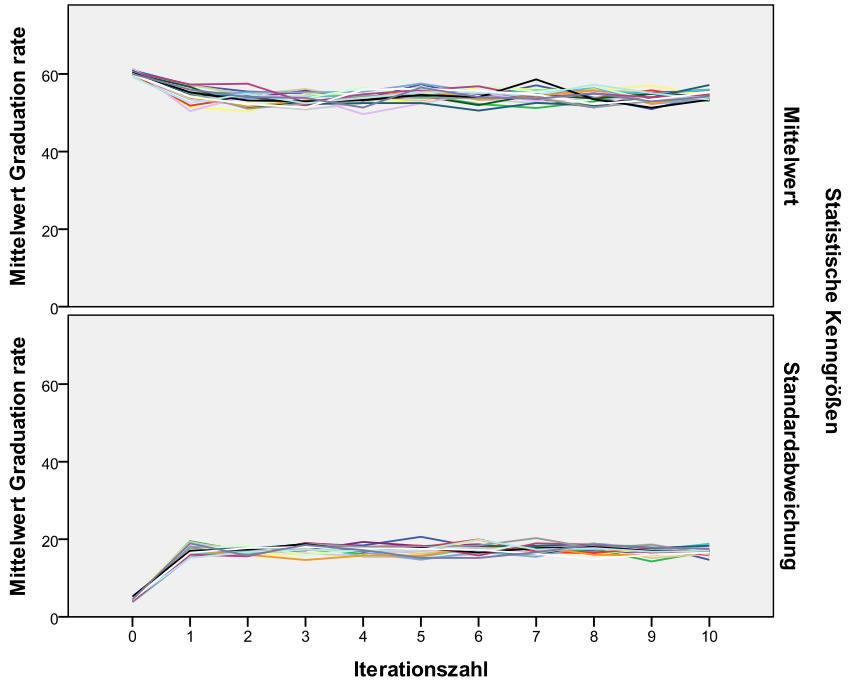
Nach 100 Ziehungen kann der Imputationsalgorithmus keinen imputierten Wert in den Nebenbedingungen für Variable GRADRAT finden. Prüfen Sie die angegebenen Minimum- und Maximumwerte, um festzustellen, ob sie angemessen sind, und ziehen Sie in Erwägung, die Anzahl der zulässigen Ziehungen zu erhöhen.

Die Ausführung dieses Befehls wurde unterbrochen.

kann man die Anzahl der erlaubten Fall- oder Parameterziehungen erhöhen, was im Fall der Parameterziehungen allerdings einen merklich erhöhten Zeitaufwand zur Folge hat. Statt die Anzahl die Parameterziehungen wesentlich zu erhöhen, sollte man besser die Nebenbedingungen einer kritischen Prüfung unterziehen.

6.4.2 Konvergenzbeurteilung

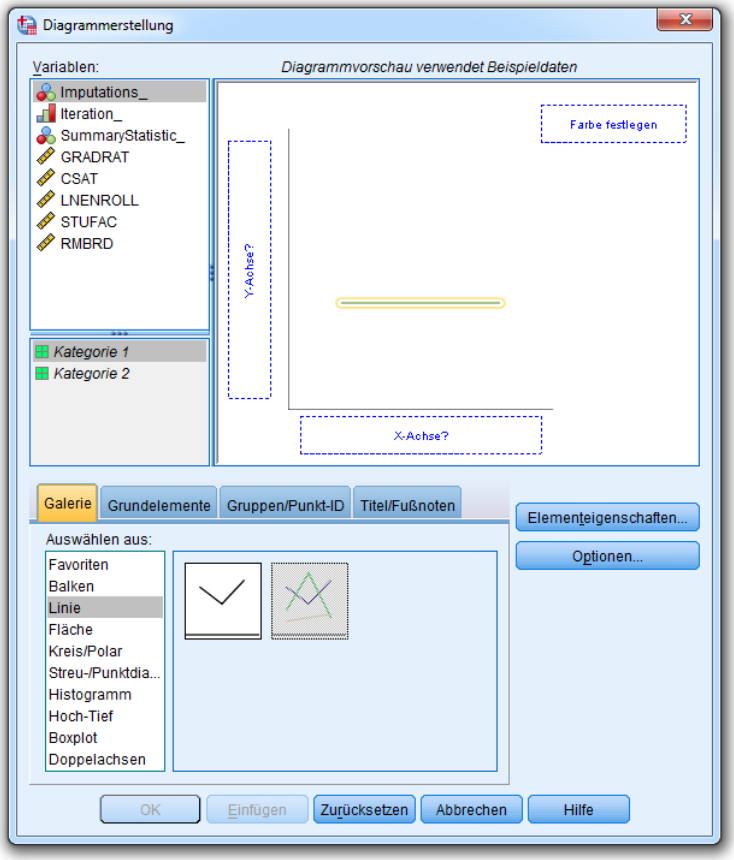
Zur Konvergenzbeurteilung wechseln wir zum Datenblatt mit dem Iterationsprotokoll (im Beispiel **CollegesMiProt** genannt) und fertigen für den Mittelwert und die Standardabweichung jeder metrischen Variablen ein Mehrliniendiagramm an, das den Iterationsverlauf für jede Imputationsstichprobe zeigt, z.B. bei GRADRAT:



Wir starten mit dem Menübefehl

Diagramme > Diagrammerstellung

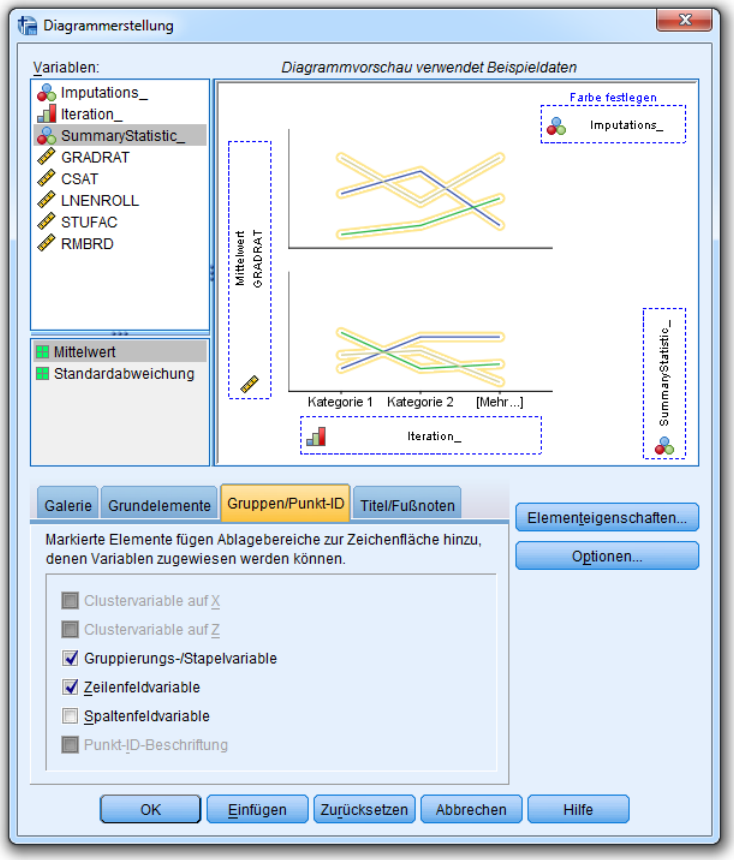
und wählen aus der **Galerie** das Mehrliniendiagramm:



Durch Ablageflächen werden drei Rollen symbolisiert, die so zu vergeben sind:

- **X-Ache**: **Iteration_**
- **Y-Achse**: GRADRAT (später dann die anderen metrischen Variablen)
- **Farbe**: **Imputations_**

Markieren Sie auf der Registerkarte **Gruppen/Punkt-ID** das Kontrollkästchen **Zeilenfeldvariable**, um eine weitere Ablagezone zu erhalten, die von der Variablen **SummaryStatistic_** eingenommen werden sollte. Mit dieser Dialogbox

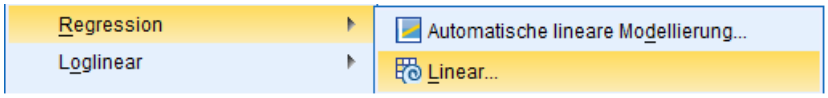


erhalten Sie das gewünschte Ergebnis für die Variable GRADRAT.

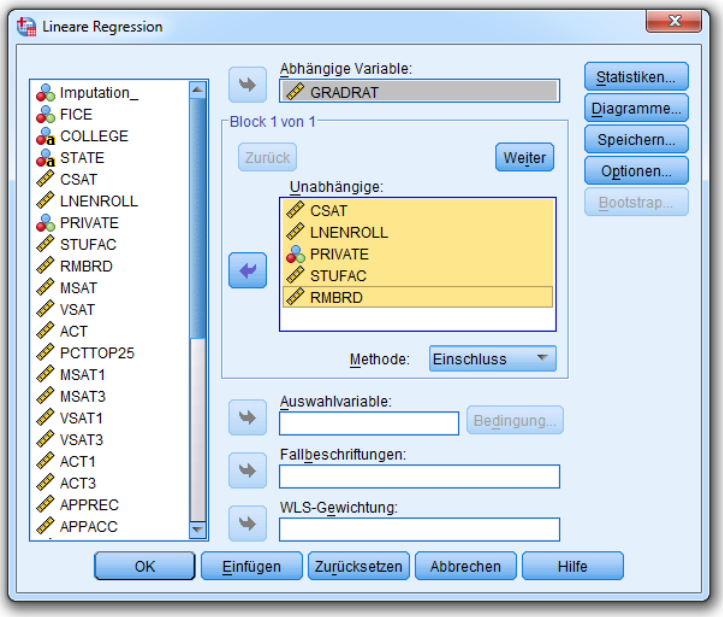
Im Diagramm zur Variablen GRADRAT (siehe oben) zeigt sich für alle Imputationsstichproben, dass die voreingestellte Anzahl von 10 Iterationen (vgl. Abschnitt 6.4.2) für eine stabilisierte Schätzung genügt. Für die restlichen metrischen Variablen erhält man ähnliche Ergebnisse.

6.4.3 Kombinierte Ergebnisse aus den Imputationsstichproben

Wir wechseln zum Imputations-Datenblatt (im Beispiel **CollegesMiSamples** genannt) und fordern über den Menübefehl



die lineare Regression von GRADRAT auf die Prädiktoren CSAT, LNENROLL, PRIVATE, STUFAC und RMBRD an:



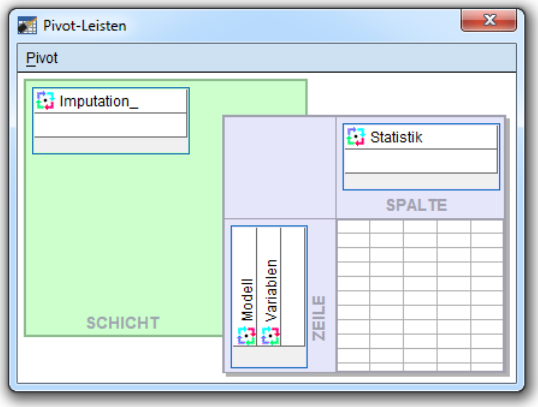
Die Konfidenzintervalle zu den Regressionskoeffizienten sind über den Schalter bzw. Subdialog **Statistiken** erhältlich.

Wir erhalten im Ausgabefenster eine längliche Koeffiziententabelle, die hintereinander Ergebnisse für die Originalstichprobe, für jede einzelne Imputationsstichprobe sowie für die Zusammenfassung der Imputationsstichproben enthält. Gehen Sie folgendermaßen vor, um nacheinander jeweils eine Tabelle ...

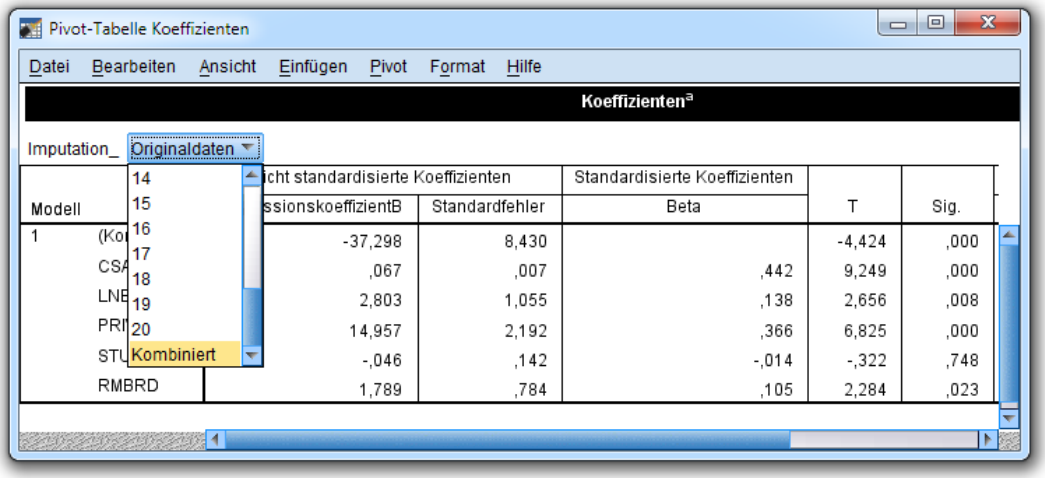
- mit den Ergebnissen für die Originaldaten (bei fallweiser Behandlung fehlender Werte)
- mit den kombinierten Imputationsergebnissen

zu erhalten:

- Koeffiziententabelle per Doppelklick im Pivot-Editor öffnen
- mit dem **Pivot**-Werkzeug aus der Zeilendimension **Imputation_** eine Schichtendimension machen:



- im Pivot-Editor die gewünschte Schicht wählen:



- Pivor-Editor schließen

Bei fallweiser Behandlung fehlender Werte (mit 372 verbliebenen Fällen) zeigt sich kein Effekt für den Regressor STUFAC (Betreuungsverhältnis), was Finanzminister begeistern wird:

Koeffizienten[a]

Originaldaten

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-37,298	8,430		-4,424	,000	-53,876	-20,719
	CSAT	,067	,007	,442	9,249	,000	,053	,081
	LNENROLL	2,803	1,055	,138	2,656	,008	,728	4,879
	PRIVATE	14,957	2,192	,366	6,825	,000	10,648	19,267
	STUFAC	-,046	,142	-,014	-,322	,748	-,325	,233
	RMBRD	1,789	,784	,105	2,284	,023	,249	3,330

a. Abhängige Variable: GRADRAT

In der Tabelle mit den kombinierten Ergebnissen zeigt sich für STUFAC ein signifikantes Testergebnis:

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B		Anteil fehlende Info.	Relative Zunahmevarianz	Relative Effizienz
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze			
1	(Konstante)	-29,289	4,962	-5,903	,000	-39,056	-19,522	,265	,350	,987
	CSAT	,064	,005	13,815	,000	,055	,073	,329	,473	,984
	LNENROLL	2,067	,643	3,215	,002	,800	3,334	,307	,429	,985
	PRIVATE	12,968	1,453	8,922	,000	10,094	15,842	,382	,596	,981
	STUFAC	-,201	,099	-2,030	,043	-,396	-,006	,292	,401	,986
	RMBRD	2,448	,542	4,513	,000	1,370	3,525	,469	,845	,977

a. Abhängige Variable: GRADRAT

Wir erfahren außerdem die **Anteile fehlender Informationen** in den einzelnen Parameterschätzungen (vgl. Abschnitt 6.3.3). Nicht nur bei den beiden am stärksten von fehlenden Werten betroffenen Variablen CSAT (40,2%) und RMBRD (39,9%) zeigen sich hohe Anteile. Auch beim Regressor STUFAC mit lediglich 2 fehlenden Werten (0,2%) tritt ein verblüffend hoher Anteil fehlender Informationen auf, den Allison (2002, S. 48f) überzeugend folgendermaßen erklärt:

- Der Anteilsschätzer verwendet nicht die Zahl fehlender Werte, sondern Varianzschätzungen innerhalb und zwischen den Imputationen.
- Bei der Schätzung eines Regressionskoeffizienten spielen auch korrelierte Regressoren eine wichtige Rolle. Dort vorhandene Lücken steigern folglich ebenfalls den Anteil fehlender Informationen.

6.4.4 Hilfsvariablen einbeziehen

Nach Überlegungen aus den Abschnitten 5.3.2 und 6.2.1 sollte es sich lohnen, in der Imputationsphase Hilfsvariablen einzubeziehen, die entweder Einfluss auf die Wahrscheinlichkeiten fehlender Werte bei Modellvariablen haben oder mit MD-belasteten Modellvariablen korrelieren. Verwendet man im Colleges-Beispiel (bei $M = 20$) die Variable MSAT als Imputationshelfer, resultiert für STUFAC ein (betragsmäßig) erheblich größerer Regressionskoeffizient mit hoch-signifikantem Testausgang:

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B		Anteil fehlende Info.	Relative Zunahmevarianz	Relative Effizienz
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze			
1	(Konstante)	-32,321	5,909	-5,470	,000	-43,996	-20,645	,366	,557	,982
	CSAT	,057	,006	10,297	,000	,046	,068	,456	,801	,978
	LNENROLL	3,119	,615	5,076	,000	1,911	4,327	,219	,275	,989
	PRIVATE	13,973	1,367	10,220	,000	11,282	16,663	,259	,340	,987
	STUFAC	-,310	,108	-2,865	,005	-,524	-,096	,384	,600	,981
	RMBRD	3,360	,601	5,593	,000	2,158	4,562	,581	1,310	,972

a. Abhängige Variable: GRADRAT

Bezieht man *alle* verfügbaren Hilfsvariablen ein (MSAT, VSAT, ACT, PCTTOP25, vgl. Abschnitt 3.1), was bei der multiplen Imputation im Unterschied zur FIML-Modellierung keine Schwierigkeiten macht, schrumpfen bei allen Regressoren mit Ausnahme von PRIVATE die Anteile fehlender Informationen und die Standardfehler:

Koeffizienten[a]

Kombiniert

Modell		Nicht standardisierte Koeffizienten		T	Sig.	95,0% Konfidenzintervalle für B		Anteil fehlende Info.	Relative Zunahmevarianz	Relative Effizienz
		Regressionskoeffizient B	Standardfehler			Untergrenze	Obergrenze			
1	(Konstante)	-34,696	5,495	-6,314	,000	-45,477	-23,914	,135	,154	,993
	CSAT	,058	,005	10,879	,000	,047	,068	,256	,335	,987
	LNENROLL	3,157	,611	5,168	,000	1,957	4,357	,188	,227	,991
	PRIVATE	13,930	1,392	10,007	,000	11,190	16,670	,260	,342	,987
	STUFAC	-,320	,106	-3,016	,003	-,530	-,111	,346	,510	,983
	RMBRD	3,748	,564	6,646	,000	2,625	4,871	,512	,998	,975

a. Abhängige Variable: GRADRAT

In der folgenden Tabelle werden die MI-Standardfehler zu den Regressionskoeffizienten mit den Ergebnissen der FIML-Analyse (vgl. Abschnitt 5.3) verglichen:

	FIML mit Hilfsvar. MSAT ($N = 1302$)	MI mit Hilfsvar. MSAT ($N = 1302$)	MI mit vier Hilfsvar. ($N = 1302$)
CSAT	,005	,006	,005
LNENROLL	0,627	0,615	0,611
PRIVATE	1,417	1,367	1,392
STUFAC	0,095	0,108	0,106
RMBRD	0,564	0,601	0,564

6.5 Unterstützung der multiple Imputation in Statistik-Programmen

Eine mehr oder weniger weit gehende Unterstützung für die multiple Imputation bieten neben SPSS u.a. die folgenden Programme, wobei die Liste (vgl. Acock 2005, S. 1025; Enders 2010, S. 329ff) keinen Anspruch auf Vollständigkeit erhebt:

- **Amos**, **LISREL**, **Mplus**
  Bei diesen Strukturgleichungsanalyseprogrammen ist zwar die FIML-Methode zur Behandlung fehlender Werte die erste Wahl, doch bieten sie auch einige Unterstützung bei der multiplen Imputation. Während sich Amos und LISREL auf die Erstellung von multiplen Imputationen beschränken, leistet Mplus auch eine Unterstützung bei der Kombination der Ergebnisse.
- **HLM**

- **Norm**
  Diese von Schafer (1997) entwickelte Windows-Freeware unterstützt die JM-Lösung für das Normalverteilungsmodell und ist über die folgende Webseite verfügbar:
  http://www.stat.psu.edu/~jls/misoftwa.html#win
- **R**
  Im freien Statistik-Entwicklungssystem R stehen für die multiple Imputation über Erweiterungspakete u.a. folgende Optionen zur Verfügung:
  - Das Paket **norm** ist eine Portierung der eben erwähnten Freeware Norm.
  - Das von Stef van Buuren (2011) entwickelte Paket **mice** (*Multiple Imputation using Chained Equations*) bietet eine FCS-Lösung, die im Funktionsumfang etwas über das SPSS-Angebot hinausgeht, aber weniger Bedienungskomfort bietet.
- **SAS** , **Stata**
  Diese beiden universellen Statistik-Pakete bieten schon seit einiger Zeit die JM-Lösung mit Normalverteilungsmodell und in den aktuellen Versionen (SAS 9.3, Stata 12) zusätzlich auch die FCS-Lösung.

7 Vergleich der behandelten Verfahren

7.1 FIML versus MI

Von allen im Manuskript besprochenen Verfahren schaffen es nur zwei auf die Empfehlungsliste:

- Direkte ML-Schätzung in Strukturgleichungsmodellen (FIML)
- Multiple Imputation (MI)

Beide Verfahren sollten bei einem saturierten Analysemodell und Verwendung derselben Hilfsvariablen zu gut übereinstimmenden Ergebnissen (Schätzungen und Standardfehlern) kommen, weil die MI Regressionsmodelle (also saturierte Modelle) zur Imputation verwendet (vgl. Enders 2010, S. 227f). Wenn das Analysemodell hingegen restringiert ist, also eine falsifizierbare empirische Behauptung enthält, ist beim Vergleich von MI und FIML mit folgendem Ergebnisbild zu rechnen (vgl. Enders 2010, S. 228):

- gut übereinstimmende Schätzer
- unwesentlich größere Standardfehler bei der MI

Bei einem restringierten (und dabei korrekten) Analysemodell sind die etwas größeren MI-Standarfehler dadurch zu erklären, dass in der Imputationsphase nicht das volle Modellwissen eingeht, was irrelevanten Zufallsassoziationen Einfluss verschafft. Man erhält stärkere Abweichungen zwischen den Parameterschätzungen aus den einzelnen Imputationsdatensätzen.

Verwendet die MI Hilfsvariablen, die im FIML-Analysemodell fehlen, sollte sich die MI als überlegen zeigen:

- bessere MAR-Erfüllung (geringere Verzerrungen der Parameterschätzungen)
- bessere Rekonstruktion fehlender Werte (kleinere Standardfehler)

Folgende Gründe können zur Entscheidung führen, für eine durch fehlende Werte belastete Analyse die FIML-Modellierung mit Amos zu verwenden:

- Latente Variablen oder nonrekursives Modell
- Kategorialer Moderator, per Mehrgruppenanalyse zu berücksichtigen
- Restringiertes Analysemodell (siehe oben)
- Flexible Mehrparametertests nach dem Likelihood-Quotienten - Prinzip
- Keine Abhängigkeit der Ergebnisse von Zufallszahlen

Bei der multiplen FCS-Imputation mit SPSS Statistics hat man u.a. folgende Vorteile:

- Einfache Beteiligung von Hilfsvariablen
  In der Imputationsphase der multiplen Imputation können ohne nennenswerten Aufwand (im Vergleich zu einem Strukturgleichungsmodell mit saturierten Korrelaten, vgl. Abschnitt 5.3.2) zahlreiche Hilfsvariablen einbezogen werden.
- Kategoriale Variablen als Kriterium, Regressor oder Hilfsvariable
  Hier ist die multiple FCS-Imputation flexibler als die FIML-Analyse in Amos, die kategoriale Variablen nur in bestimmten Fällen angemessen einbeziehen kann (exogene manifeste Variablen oder Gruppenvariablen, jeweils ohne fehlende Werte). In unserem Colleges-Beispiel haben wir allerdings mit der lückenhaften, dichotomen Variablen PRIVATE (exogen und manifest) die Normalverteilungsannahme ohne große Probleme strapaziert.

7.2 Übersichtstabelle zur Eignung der behandelten Verfahren

Die folgende Tabelle berücksichtigt Methoden zur Behandlung fehlender Werte, die unter bestimmten Voraussetzungen immerhin konsistente Schätzungen liefern:

	Methode	**Voraussetzung**	**Konsistente Schätzung**	**korrekte Standard-fehler**	**Bemerkungen**
Ausschluss fehlender Werte	Fallweise	MCAR	Ja	Ja	Informationsverlust
	Paarweise	MCAR	Ja	(Ja)	*N* unklar; indefinite Korrelationsmatrix mögl.
ML-Schätzung	**Direkte ML-Schätzung der interessierenden Parameter (FIML)**	**MAR, SEM-Voraussetzungen**	**Ja**	**Ja**	**Mit Amos möglich**
	ML-Schätzung der Verteilungsmomente (Mittelwerte, Varianzen, Kovarianzen) per EM-Algorithmus. Anschließend Verwendung von Standardmethoden (z.B. Regression)	MAR, multivariate NV, wobei Verletzungen unkritisch sind bei Var. ohne MD; hinreichende große Stichprobe	Ja	Nein	Mit SPSS umständlich zu realisieren; für deskriptive Analysen geeignet (z.B. explorative Faktorenanal., interne Konsistenz)
Imputation	Einfache Imputation per Regression mit Zufallskomponente	MAR	Ja	Nein	In SPSS MVA zumindest bei MAR mangelhaft (siehe Abschnitt 4.8)
	Einfache Imputation per EM-Schätzung mit Zufallskomponente	MAR (wie bei EM-Schätzung der Momente)	Ja	Nein	In SPSS MVA mangelhaft, weil **ohne** Zufallskomponente (siehe Abschnitt 5.2)
	Multiple Imputation	**MAR**	**Ja**	**Ja**	**Mit SPSS Statistics möglich**

Literatur

Acock, A.C. (2005). Working With Missing Values. *Journal of Marriage and Family*, 67, 1012–1028.

Allison, P.D. (2002). *Missing Data*. Thousand Oaks: Sage Publications.

Allison, P.D. (2009). Missing Data. In Millsap, R.E. & Maydeu-Olivares, A. (Hrsg.), *The SAGE Handbook of Quantitative Methods in Psychology* (S. 72-89). Thousand Oaks: Sage Publications.

Arbuckle, J.L. (2012). *IBM SPSS Amos 21 User's Guide*. Manual zu IBM SPSS Amos 21.

Baltes-Götz, B. (1994). *Einführung in die Analyse von Strukturgleichungsmodellen mit LISREL 7 und PRELIS unter SPSS*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=22734

Baltes-Götz, B. (2010). *Analyse von Strukturgleichungsmodellen mit Amos 18.0*. Online-Dokument: http://www.uni-trier.de/index.php?id=22640

Baltes-Götz, B. (2013). *Statistisches Praktikum mit IBM SPSS Statistics 21 für Windows*. Online-Dokument: http://www.uni-trier.de/index.php?id=22552

Barnard, J. & Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation, *Biometrika*, 86, 948–955.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K.A. & Stine, R.A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In: K.A. Bollen & J.S. Scott (S. 111-135), *Testing structural equation models*. Newbury Park, CA: Sage.

Cohen, J., Cohen, P., West, S.G. & Aiken, L. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah: Lawrence Erlbaum Associates.

Enders, C.K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

Graham, J.W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 10, 80-100.

Grynaviski, J. (2003). *Applied Bayesian Statistics*. Online-Dokument: http://home.uchicago.edu/~grynav/bayes/abs03.htm (abgerufen: 21.07.2008).

IBM Corporation (2012). *IBM SPSS Missing Value 21*. Manual zu IBM SPSS 21.

King, G., Honacker, J., Joseph, A. & Scheve, K. (2001). *Analyzing Incomplete political Science Data*. Online-Dokument: http://gking.harvard.edu/files/evil.pdf (abgerufen am 22.06.2013).

Kline, R.B. (2005). *Principles and Praxis of Structural Equation Modeling*. New York: Guilford Press.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. & Graham, J. W. (2002). Missing Data: Our View of the State of the Art, *Psychological Methods*, Vol. 7, No. 2, 147–177

Scheffer, J. (2002). Dealing with Missing Data. *Res. Lett. Inf. Math. Sci.*, 3, 153-160.

Tabachnik, B.G. & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, Vol. 16, 219–242.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. & Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, Vol. 76, No. 12, 1049–1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Computation*, Vol. 45, No. 3, 1049–1064.

von Hippel, P. T.(2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician*, Vol. 58, No. 2, 160-164. Online-Dokument:
http://www.sociology.ohio-state.edu/people/ptv/publications/MVA/published.pdf
(abgerufen am 22.06.2013)

Stichwortverzeichnis