https://doi.org/10.20378/irbo-51486

MISS - Multiple Imputation Seizes Surveys
Über den Umgang mit fehlenden Daten in Umfragen

Von Dipl. Math. Christine Licht

Zusammenfassung

Ein allgegenwärtiges und häufig auftretendes Problem bei der Aufbereitung und Analyse von Daten sind fehlende Werte. Bei Datenerhebungen werden fehlende Werte aus unterschiedlichen Gründen erzeugt - Personen werden nicht angetroffen, die Daten gehen nach der Befragung verloren, oder befragte Personen verweigern schlichtweg die Antwort auf bestimmte oder alle Fragen. Ein einfaches und standardmäßig verwendetes Verfahren zur Behandlung fehlender Daten besteht darin, die Analyse lediglich auf Subjekte zu restringieren, die vollständig beobachtete Variablen aufweisen. Die aus solch einer Analyse resultierenden Inferenzen sind meistens nicht statistisch valide, vor allem dann nicht, wenn der Datenausfall einem bestimmten Muster folgt und nicht rein zufällig ist.

Dieser Artikel stellt verschiedene Methoden zum Umgang mit Antwortverweigerung in Umfragen vor. Wir konzentrieren uns hierbei zunächst auf *multiple imputation*. Dabei vergleichen wir basierend auf einer Arbeit von Raghunathan [Ra04] anhand einer Simulationsstudie sechs verschiedene Verfahren zur Behandlung fehlender Daten bezüglich einer Gesundheitsumfrage miteinander. Anschließend erweitern und untersuchen wir den von Meng in [Me94] aufgezeigten Begriff der *congeniality* (Gleichartigkeit) an diesem medizinischen Beispiel.

Stichworte: *before deletion*-Analyse, *available case*-Analyse, *mean imputation, single imputation, multiple imputation, missing data*-Mechanismus, *(un)congeniality*

Christine Licht

1 Einleitung

Überall wird heutzutage eine kaum überschaubare Menge an Daten erhoben. Sei es in Meinungsumfragen auf der Straße, bei Gewinnspielen, über Kunden- und Paybackkarten, bei staatlichen Erhebungen wie dem Zensus 2011 in Deutschland oder bei Aufnahme als Patient im Krankenhaus und (anschließenden) medizinischen Befragungen. Ein allgegenwärtiges Problem bei der Aufbereitung bzw. Analyse von Daten sind fehlende Werte. Dabei können ganz verschiedene Ausfallmuster auftreten, wie beispielhaft in Abbildung 1.1 dargestellt ist.

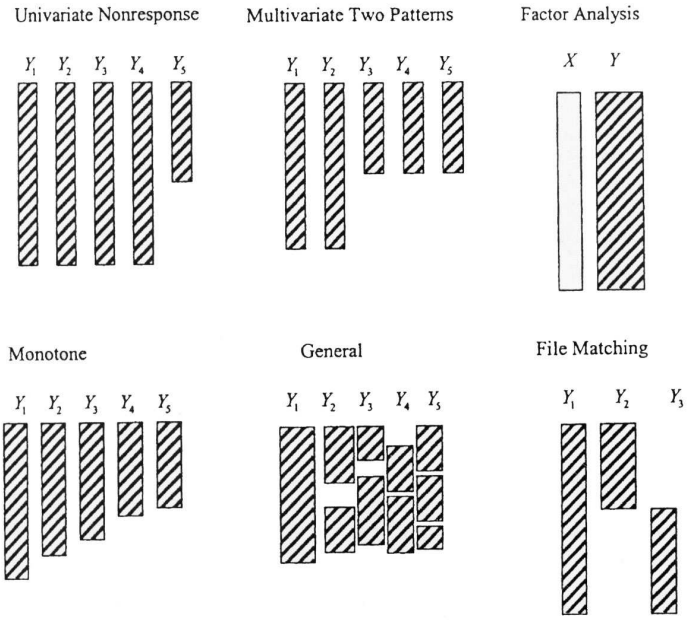

Abbildung 1.1: Ausfallmuster

Fehlende Daten weisen nicht nur verschiedene Ausfallmuster auf, sondern unterliegen auch einem gewissen Ausfallmechanismus. Dieser

wurde von Rubin [Ru76] entwickelt und später von Little [Li95] auf Längsschnittdaten ausgeweitet. Man unterscheidet hierbei die drei Fälle *missing at random* (MAR), *not missing at random* (NMAR) und *missing completely at random* (MCAR), welche in Kapitel 2.2 beschrieben werden. "What do we do with missing data?" Basierend auf dem gleichnamigen Artikel von Raghunathan [Ra04] und dem von ihm verwendeten Modell wird dieser Fragestellung nun nachgegangen. Das bei fehlenden Daten standardmäßig angewandte Verfahren besteht darin, nur die in den interessierenden Variablen vollständigen Beobachtungen (engl.: *available case*; Abkürzung: AC) oder die in allen Variablen vollständigen Daten (engl.: *complete case*; Abkürzung: CC) zu verwenden. Dies ist aber nur dann sinnvoll, wenn der dadurch entstehende Daten- bzw. Informationsverlust nicht zu groß ist und der Ausfallmechanismus MCAR war. In wenigen und sehr spezifischen Fällen kann AC auch unter einer schwächeren Annahme gültig sein. Andere Verfahren versuchen die fehlenden Werte anhand der beobachteten Daten zu ergänzen, wie z.B. *mean imputation* (engl. für: Mittelwertergänzung) und *single imputation* (engl. für: einfache Ergänzung; Abkürzung: SI). Allerdings treten bei diesen Verfahren häufig verzerrte Schätzer und/oder über-/unterschätzte Varianzen auf. Die Methoden sind deshalb zum Teil ungeeignet für gültige statistische Inferenzen. Eine mögliche Herangehensweise unvollständige Daten korrekt zu analysieren ist *multiple imputation* (engl. für: mehrfache Ergänzung; Abkürzung: MI). Hierbei ([Ru78], [Ru87]) werden die fehlenden Daten unter Ausnutzung aller zur Verfügung stehenden Informationen ersetzt. Um die Unsicherheit im Datenausfall und im Ergänzungsmodell widerzuspiegeln, wird mehrfach imputiert. Anschließend werden die zu schätzenden Parameter aus den vervollständigten Datensätzen ermittelt und ein MI-Schätzer gebildet. Wichtig dabei ist, dass die fehlenden Werte nicht vorhergesagt werden können, sondern das Ziel ist eine valide Datenanalyse, d.h. die Erzeugung unverzerrter Schätzer und korrekter Überdeckungen (engl.: *coverage*) durch die entsprechenden Konfidenzintervalle. Ein zum Teil deutlicher Mehraufwand an (Rechen-)Zeit lohnt sich, da auf vollständige Daten zahlreiche Standardanalysen angewendet werden können und umfangreiche Analysesoftware zur Verfügung steht. Zudem erlaubt MI

die Erzeugung von *public use files* und wird zudem im Datenschutz zur Gewinnung völlig synthetischer Daten eingesetzt.
Um die Auswirkung der verschiedenen Methoden zur Behandlung fehlender Werte zu demonstrieren, werden wir hierzu in Kapitel 3 sechs Verfahren und deren Wirkungsweise anhand einer Simulationsstudie vergleichen. Kapitel 4 befasst sich schließlich mit dem wenig untersuchten und in [Me94] von Meng aufgezeigten Problem der *uncongeniality* (engl. für: Ungleichartigkeit). Über den Ansatz von Meng hinausgehend, werden wir bei der *multiple imputation* vier Modelle betrachten: den datengenerierenden Prozess, das Analysemodell, den Ausfallmechanismus und das Imputationsmodell, wobei die verschiedenen Modelle idealerweise miteinander in gewisser Weise übereinstimmen sollten. Wir werden anhand einer Simulationsstudie zum gleichen Beispiel wie in der Vergleichsstudie der Fragestellung nachgehen, was passiert, wenn die Modelle nicht gleichartig sind. Schließlich geben wir in Kapitel 5 eine Zusammenfassung der wichtigsten Ergebnisse und einen kurzen Ausblick.

2 Simulationsstudie

Um die Wirkungsweise von *multiple imputation* im Vergleich mit anderen Verfahren sowie die Auswirkungen von *uncongeniality* genauer zu untersuchen, werden wir einer medizinischen Fragestellung nachgehen und angelehnt an Raghunathan [Ra04] das im folgenden beschriebene Modell verwenden.

2.1 Modellbeschreibung

Wir betrachten eine Gruppe von Studienteilnehmern mit einer binären Risikovariable D, einer binären Expositionsvariable E und einem stetigen Störfaktor x. Beispielsweise interessiert uns die Auswirkung von Strahlenbelastung durch Sendemasten auf Krebserkrankungen unter dem Einfluss der Belastung durch Röntgenstrahlen pro Jahr. Die beobachtete Exposition „Standort eines Sendemastes in maximal 500 Meter Entfernung vom Wohnort“ ist allerdings nicht bzw. nicht die alleinige Ursache für die beobachtete Wirkung. Der Krebs wird von einem so

genannten *confounder* (engl.: Störfaktor) hervorgerufen. D.h. die Belastung durch Röntgenstrahlung bestimmt das Auftreten des Risikofaktors Krebs mit.
Wir treffen nun folgende Modellannahmen:

- $x \sim N(0,1)$,
- $logit\,(p(E=1|x)) = 0.25 + 0.75 \cdot x,$

  d.h. $p(E=1|x) = p_E = (e^{-(0.25+0.75 \cdot x)})^{-1}$ und $E \sim Ber(p_E)$,
- $logit\,(p(D=1|x,E)) = -0.5 + 0.5 \cdot E + 0.5 \cdot x,$

  d.h. $p(D=1|x,E) = p_D = (1+e^{-(-0.5+0.5 \cdot E+0.5 \cdot x)})^{-1}$ und

  $D \sim Ber(p_D)$.

Wir ziehen nun eine standardnormalverteilte Stichprobe x der Größe 1000 und berechnen daraus mit obigem Modell die Variablen D und E. Anschließend schätzen wir die Parameter $ß_0$, $ß_1$ und $ß_2$ mittels einer logistischen Regression von E und x auf D und vergleichen diese mit den Originalparametern -0.5, 0.5 und 0.5. Unser Analysemodell lautet also:

$$logit\,(p(D=1|x,E)) = ß_0 + ß_1 \cdot E + ß_2 \cdot x.$$

Aus dieser Analyse resultieren für unsere Stichprobe der Größe 1000 die Schätzer $\hat{ß}_0 = -0.47$, $\hat{ß}_1 = -0.49$ und $\hat{ß}_2 = 0.55$ für den Achsenabschnitt und die Regressionskoeffizienten E bzw. x. Diese Schätzer liegen schon recht nah bei den wahren Werten -0.5, 0.5 und 0.5 der im datengenerierenden Prozess gegebenen logistischen Regression.

Nun werden wir absichtlich einige Werte von x löschen. Dabei existieren drei von Rubin [Ru76] formalisierte Ausfallmechanismen, die wir im Folgenden vorstellen werden.

2.2 Ausfallmechanismen

Zunächst werden wir einige Bezeichnungen einführen. Sei Y die ($M \times N$)-Datenmatrix, wobei M die Anzahl der Beobachtungen und N die Anzahl der Variablen sei. Sei weiterhin R die ($M \times N$)-Indikatormatrix von Y mit $r_{ij} = 1$, falls y_{ij} nicht beobachtet wurde und $r_{ij} = 0$, falls y_{ij} vorliegt für $i = 1, ..., M$ und $j = 1, \ldots, N$. Es sei angemerkt, dass man heute oftmals R als *response*-Indikator (engl. für Antwort) verwendet mit $r_{ij} = 1$, falls y_{ij} beobachtet wurde und $r_{ij} = 0$, falls y_{ij} nicht vorliegt ([LR02]). Wir werden hier aber der Notation von Raghunathan [Ra04] folgen und R als *missing*-Indikator wie eingangs beschrieben, verwenden. Y_{obs} bezeichne die beobachteten Werte in Y und Y_{mis} die fehlenden Werte in Y, so dass gilt $Y = (Y_{obs}, Y_{mis})$. Mit dem *missing data*-Mechanismus (engl. für: Ausfallmechanismus) bezeichnet man die Wahrscheinlichkeit, dass R den Wert 1 annimmt gegeben die Daten Y und weitere Parameter ξ, d.h.

missing data-Mechanismus: $p(R = 1 \mid Y, \xi)$.

2.2.1 Missing At Random (MAR)

Missing at random (MAR) liegt vor, wenn gilt

$$p(R \mid Y, \xi) = p(R \mid Y_{obs}, \xi).$$

Beispielsweise geben Personen ihr Gewicht nicht an, weil sie sich für zu dick halten und in der Familie übergewichtige Personen leben. Das Fehlen hängt also von anderen beobachteten Variablen ab, aber nicht vom Wert des Gewichts selbst. Um ein MAR in unserem konkreten Beispiel zu simulieren, verwenden wir das von Raghunathan in [Ra04] aufgestellte logistische Modell:

$$logit\,(p(x = missing)) = -1.11 - 1.09 \cdot D - 1.85 \cdot E + 2.31 \cdot D \cdot E.$$

Wir sehen, dass der Ausfall von (allen) anderen Variablen (hier D und E) abhängt, aber nicht vom Wert x selbst. Für jedes Subjekt generieren wir nun eine Bernoulli-verteilte Zufallsvariable mit der soeben berechneten Wahrscheinlichkeit. Nimmt diese Zufallsvariable den Wert 1 an, so wird der entsprechende Wert in x gelöscht. Das verwendete *logit*-Modell ist dabei so konstruiert, dass wir damit etwa 15% Datenausfall erzeugen.

2.2.2 Not Missing At Random (NMAR)

Not missing at random (NMAR) liegt vor, wenn gilt

$$p(R \mid Y, \xi) = p(R \mid Y_{mis}, \xi).$$

Beispielsweise geben Personen ihr Gewicht nicht an, weil sie 63 Kilogramm oder mehr wiegen, aber nicht, weil sie sich für zu dick halten oder in der Familie übergewichtige Personen leben.
Das Fehlen hängt also direkt vom Wert selbst ab, aber nicht von anderen Variablen. Lediglich in Abhängigkeit von x bestimmen wir eine Ausfallwahrscheinlichkeit p_{mis} über das Modell $p_{mis} = (1 + e^{-x+2.1})^{-1}$. Mit dieser Wahrscheinlichkeit generieren wir eine (binäre) Indikatorvariable *mis*. Nimmt diese den Wert 1 an, so wird der entsprechende Wert in x gelöscht. Dabei haben wir das für p_{mis} verwendete *logit*-Modell so konstruiert, dass wir, wie im MAR-Fall, etwa 15% Datenausfall erhalten.

2.2.3 Missing Completely At Random (MCAR)

Missing completely at random (MCAR) liegt vor, wenn gilt:

$$p(R \mid Y, \xi) = p(R \mid R, \xi).$$

Der Datenausfall ist demnach unabhängig von den Daten (beobachtet oder fehlend) bzw. von den erhobenen Variablen, sondern rein zufällig. Beispielsweise geben Personen ihr Gewicht nicht an - völlig unabhängig von ihrem tatsächlichen Gewicht oder ob sie sich zu dick fühlen oder ob

übergewichtige Personen im Haushalt leben. Um MCAR zu simulieren, legen wir eine „Ausfallwahrscheinlichkeit" fest und generieren damit eine Bernoulli-verteilte Zufallsvariable. Nimmt diese den Wert 1 an, so wird der entsprechende Wert in x gelöscht. Wählen wir als Ausfallwahrscheinlichkeit z.B. 0.15, erzeugen wir offensichtlich etwa 15% (zufälligen) Datenausfall in x.

2.3 Multiple Imputation

Es existieren Verfahren, die es ermöglichen die Verzerrung der Schätzer (engl.: *nonresponse bias*) in *available case*-Analysen zu beseitigen. Bezieht man allerdings nur Subjekte mit vollständig beobachteten Daten ein, so werden Teilinformationen von Subjekten mit unvollständigen Daten ignoriert. Wichtige Informationen gehen verloren. Imputation hingegen bezieht alle verfügbaren Informationen und Beziehungen zwischen den einzelnen Variablen ein. Für (ver)vollständig(t)e Daten kann auf eine Vielzahl von Analysesoftwares zurückgegriffen werden, die sich im Gegensatz zu *incomplete data*-Software schneller weiterentwickelt hat und an die aktuellen statistischen Methoden angepasst ist. Wird ein Datensatz für mehrere Wissenschaftler bzw. zur öffentlichen Nutzung bereitgestellt, kann der Imputer spezielles Wissen einbeziehen: beispielsweise über die Gründe des Datenausfalls oder vertrauliche Informationen, die nicht für eine breite Öffentlichkeit bestimmt sind, oder andere Variablen, die nicht von verschiedenen Personen genutzt werden dürfen. Zudem erhalten alle Nutzer dieselben vervollständigten Datensätze und müssen sich nicht mehr um das Problem der Behandlung fehlender Daten kümmern. Obwohl *single imputation*, wobei jeder fehlende Wert nur einmal ersetzt wird, all diese Vorteile aufweist, spiegelt dieses Verfahren nicht die Unsicherheit in den Daten wider. Diese Unsicherheit besteht aufgrund der Tatsache, dass die imputierten Werte plausible Ersetzungen, aber nicht die wahren Werte sind. *Single imputation* produziert im Allgemeinen zu kleine Standardabweichungen und zu schmale Konfidenzintervalle. *Multiple imputation* hingegen weist alle Vorteile der *single imputation* auf und beachtet zudem die Unsicherheit des Datenausfalls und in der Ergänzung. Die Idee besteht darin, m vervollständigte Datensätze zu erzeugen. Obwohl der MI-Schätzer $\hat{\vartheta}_{MI}$ erst

für $m \to \infty$ gegen den Erwartungswert $E(\Theta | Y_{obs})$ des unbekannten Parameters Θ gegeben die beobachteten Daten Y_{obs} konvergiert, genügen $m = 5$ Ergänzungen. Für größere m ist die Abweichung in den Resultaten vernachlässigbar gering, wie Rubin in [Ru87] zeigt. Die Varianz zwischen den m ergänzten Datensätzen spiegelt dabei die Unsicherheit in den Daten wider. Abbildung 2.1 veranschaulicht das Prinzip der *multiple imputation*.

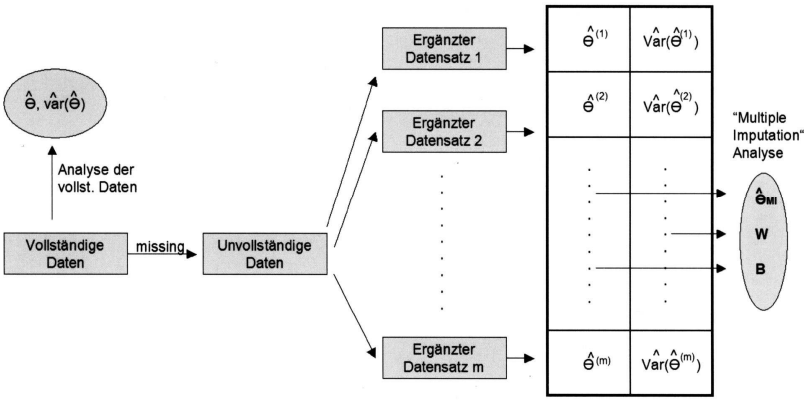

Abbildung 2.1: *Multiple imputation* - Prinzip

Haben wir m ergänzte Datensätze erzeugt, analysieren wir jeden Datensatz für sich und ermitteln die jeweiligen Punktschätzer und die Standardabweichungen. Kombinieren wir diese nach Rubins *combining rules* [Ru87], erhalten wir den so genannten MI-Schätzer $\hat{\vartheta}_{MI}$ und dessen Varianz sowie das assoziierte Konfidenzintervall.

Es bezeichne $\hat{\vartheta}^{(k)}$ den aus den jeweiligen vervollständigten Datensätzen ermittelten Schätzer und $\hat{Var}(\hat{\vartheta}^{(k)})$ dessen Varianz mit $k = 1, 2, \ldots, m$ und $m \geq 2$. Der MI-Schätzer ist der Durchschnitt

$$\hat{\vartheta}_{MI} = \frac{1}{m}\sum_{k=1}^{m}\hat{\vartheta}^{(k)}.$$

Die totale Varianz T des multipel imputierten Schätzers ist

$$T = W + \frac{m+1}{m}B,$$

wobei gilt:

$$W = \frac{1}{m}\sum_{k=1}^{m}\hat{Var}(\hat{\vartheta}^{(k)})$$

und

$$B = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\vartheta}^{(k)} - \hat{\vartheta}_{MI})^2 .$$

Die totale Varianz T setzt sich aus W und B zusammen. Dabei ist W die innere oder interne Varianz (engl.: *within-imputation variance*), d.h. wir behandeln die imputierten Werte so, als ob es die wahren Werte wären, und berechnen den Durchschnitt der einzelnen Varianzen. B hingegen ist die externe Varianz (engl.: *between-imputation variance*), d.h. die Varianz zwischen den imputierten Werten. Basierend auf Rubin [Ru87] können wir nun das entsprechende Konfidenzintervall

$$\hat{\vartheta}_{MI} \pm \sqrt{T}\cdot t_{1-\frac{\alpha}{2};\nu}$$

bilden, mit $\nu = (m-1)\cdot\left(1+\frac{W}{(1+m^{-1})\cdot B}\right)^2$ Freiheitsgraden.

Um in unserem konkreten Beispiel die fehlenden Werte mittels *multiple imputation* zu ergänzen, haben wir einen MI-Algorithmus für fehlende stetige Daten verwendet, da $x \sim N(0,1)$ und somit stetig ist. Dabei wird als Imputationsmodell ein lineares Regressionsmodell

$$x = \gamma_0 + \gamma_1\cdot D + \gamma_2\cdot E + \gamma_3\cdot D\cdot E + \varepsilon = U\gamma + \varepsilon, \quad \varepsilon \sim N(0,1)$$

unterstellt. Auf Basis der beobachteten Daten werden nun die OLS-Schätzer $\hat{\gamma}_{obs}$ und $\hat{\sigma}^2_{obs}$ bestimmt.
D.h. es ist

$$\hat{\gamma}_{obs} = (U^t_{obs} U_{obs})^{-1} U^t_{obs} x_{obs}$$

und

$$\sigma^2_{obs} = (x_{obs} - U_{obs}\hat{\gamma}_{obs})^t (x_{obs} - U_{obs}\hat{\gamma}_{obs}) / (n_{obs} - 4),$$

wobei $U = [1\, D\, E\, DE]$ die so genannte Designmatrix darstellt, x_{obs} die beobachteten Werte in x sind und entsprechend U_{obs} den zu x_{obs} gehörenden Teil von U und n_{obs} die Anzahl der beobachteten Werte in x bezeichnet. Die MI-Prozedur besteht nun aus den folgenden drei Schritten, die m-mal wiederholt werden:

1. Ziehe $\sigma^2 \mid U \sim \hat{\sigma}^2_{obs}(n_{obs} - 4)\chi^{-2}_{n_{obs}-4}$.

2. Ziehe einen Vektor von vier Variablen aus

   $\gamma \mid \sigma^2, U \sim N(\hat{\gamma}_{obs}, \sigma^2 (U^t_{obs} U_{obs})^{-1})$.

3. Ziehe $x_{mis} \mid \gamma, \sigma^2, U \sim N(U_{mis}\gamma, \sigma^2)$ unabhängig voneinander

   für jeden fehlenden Wert $i = 1, 2, \ldots, n_{mis}$.

Nach der Analyse der nach diesem Prinzip vervollständigten Datensätze und der Berechnung der entsprechenden Punktschätzer und Konfidenzintervalle haben wir diese mit den Ergebnissen aus den vollständigen Daten, d.h. vor dem Löschen (engl.: *before deletion*; Abkürzung: BD), verglichen.

2.4 Ergebnisse

Wir haben den in Kapitel 2.3 beschriebenen MI-Algorithmus in ***R*** implementiert und zunächst auf 2500 Datensätze, die nach einem der drei Mechanismen MAR, NMAR oder MCAR fehlende Werte enthalten,

angewendet. Wir haben für jeden der 2500 Datensätze fünf Imputationen erzeugt und anschließend die ergänzten Datensätze anhand der in Kapitel 2.1 beschriebenen logistischen Regression analysiert. Mit den oben angegebenen Kombinationsregeln haben wir danach die 2500 MI-Schätzer $\hat{ß}_{MI}$, deren Varianzen T und die entsprechenden Konfidenzintervalle berechnet. Im Verlauf unserer Untersuchungen haben wir festgestellt, dass wir schon mit 500 Durchläufen nahezu identische Ergebnisse erzielen, so dass wir uns aufgrund der wesentlich schnelleren Rechenzeit auf 500 Durchläufe bei einer Datensatzgröße von 1000 beschränken.

Die im folgenden gegebene Zusammenfassung unserer Ergebnisse, unterteilt nach der Art des *missing data* –Mechanismus, beinhaltet die Durchschnittswerte (der 500 Simulationen) der Regressionskoeffizienten für den uns interessierenden Parameter E vor dem Löschen (BD) und nach der *multiple imputation* sowie das jeweils zugehörige Histogramm von 500 simulierten Datensätzen und die entsprechenden Überdeckungen des wahren Werts $ß_1 = 0.5$ durch die assoziierten 95%-Konfidenzintervalle. Angemerkt sei, dass wir den Fall NMAR in zwei Varianten unterteilt haben. Zum einen betrachten wir den allgemeinen Fall des zufälligen, aber von der Variable x abhängigen Ausfalls und zum anderen als Spezialfall die Stutzung, d.h. die Störvariable x unterliegt nach dem Löschen einer gestutzten Normalverteilung. Erinnern wir uns noch einmal an das eingangs erwähnte Beispiel für unser medizinisches Modell. Deutschland nimmt beim Röntgen einen Spitzenplatz ein: etwa 1.3 Röntgenaufnahmen und eine Strahlenbelastung von etwa 2 mSv (Sv=Sievert) pro Einwohner und Jahr. Darauf lassen sich theoretisch 1.5% der jährlichen Krebsfälle zurückführen. Durch die Röntgenaufnahme der Wirbelsäule oder einer Mammographie der weiblichen Brust beträgt die Strahlenbelastung bis zu 5 mSv, bei einer Computertomographie des Brustkorbs sogar 10 mSv. Zudem ist der Mensch, abhängig vom Wohnort, einer generellen Strahlung von bis zu 5 mSv ausgesetzt.[1] Es ist somit durchaus sinnvoll den Spezialfall der Stutzung zu

[1] Dr. med. Carsten Körber, Nuklearmedizinische Praxis Fulda, www.medizin-netz.de

betrachten. Ab einer Strahlenbelastung durch Röntgenuntersuchungen von beispielsweise 20 mSv pro Jahr gelten Patienten als stark belastet. Bei Werten kleiner als 20 mSv, was für den Durchschnittsbürger realistisch ist, wird die Belastung dann durch die gestutzte Verteilung simuliert.
In unserer Simulation legen wir dazu vorher einen Entscheidungswert fest. Ist x größer als dieser Entscheidungswert, so wird x an dieser Stelle gelöscht, d.h. salopp gesprochen: Wir schneiden x einfach rechts vom Entscheidungswert ab. In Tabelle 2.1 sind unsere Ergebnisse für die Ausfallmechanismen MAR, NMAR allgemein, NMAR Stutzung und MCAR im Vergleich zum *before deletion*-Fall zusammengefasst. Wie bereits in Kapitel 2.2 erwähnt, sind alle Ausfallmechanismen so konstruiert, dass wir etwa 15% fehlende Werte erzeugen. Bei der Berechnung der entsprechenden Konfidenzintervalle haben wir ein Konfidenzniveau von 95% verwendet. Zudem beschränken wir im Folgenden die Darstellung unserer Ergebnisse, wie eingangs dieses Kapitels erwähnt, auf den Regressionskoeffizienten $\hat{\beta}_1$, da wir uns für Einfluss des Sendemastenstandorts E auf das Krebsrisiko D interessieren.

	BD	MAR	NMAR allg.	NMAR Stutzung	MCAR
$\widehat{\beta_1}$	0.4970	0.4949	0.5346	0.6089	0.504
coverage	0.96	0.948	0.936	0.868	0.948

Tabelle 2.1: Durchschnittswerte für $\hat{\beta}_1$ und *coverage* bei BD und MI

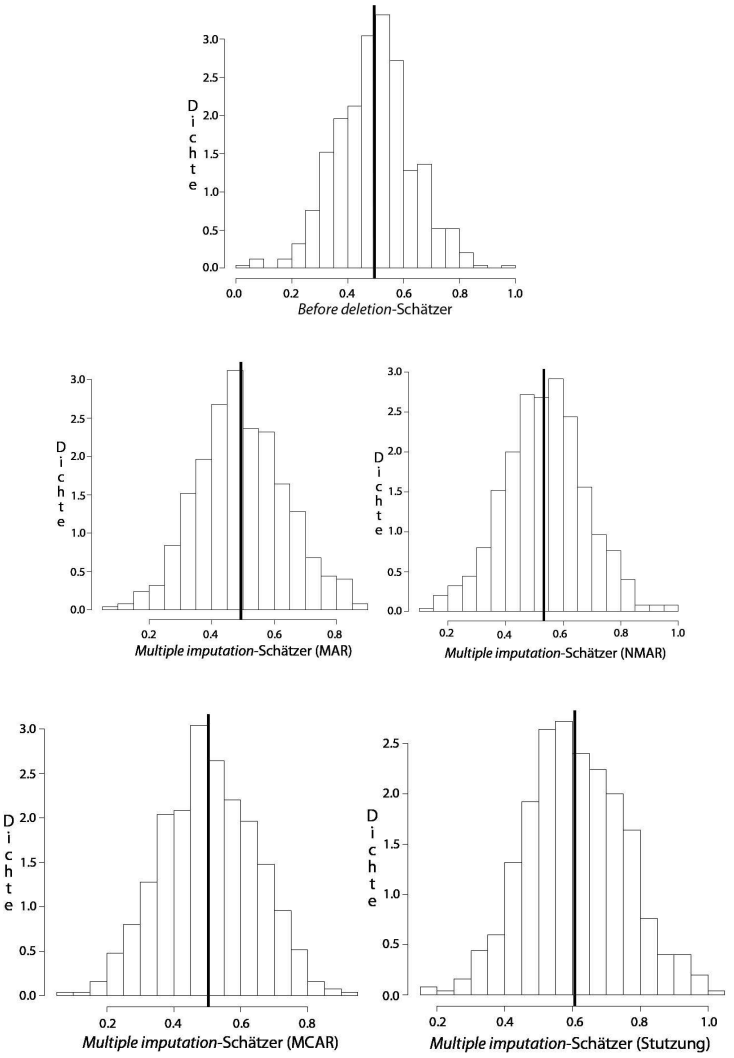

Abbildung 2.2: Histogramme der $\hat{\beta}_1$ für BD, MAR, NMAR, Stutzung und MCAR

Wir sehen, dass MI für die Ausfallmechanismen MAR und den weniger realistischen Fall MCAR sehr gute Ergebnisse liefert, d.h. wir konnten unverzerrte Schätzer und eine dementsprechend hohe *coverage* erzielen. Raghunathan, auf dessen Arbeit [Ra04] unsere Untersuchungen aufbauen, hat sich auf den MAR-Datenausfall beschränkt und dabei für MI vergleichbar gute Ergebnisse erzielt. Für die beiden NMAR-Fälle sind unsere Schätzer schlechter, wobei auffällt, dass wir bei der Stutzung sowohl einen stark verzerrten Schätzer als auch eine wesentlich schlechtere *coverage* als bei den anderen drei Fällen erzielen. Außerdem erkennen wir bereits einen ersten Fall von *uncongeniality*. Wir haben die Daten durch MAR oder NMAR oder MCAR gelöscht, im Imputationsmodell aber stets MAR unterstellt. Ausfall- und Imputationsmodell stimmen also nicht immer überein. Auf diese Problematik werden wir in Kapitel 4 noch genauer eingehen. Darauf aufbauend wollen wir im nächsten Kapitel die Ergebnisse von MI mit anderen Ergänzungsverfahren und der häufig standardmäßig angewandten *available case*-Analyse in Abhängigkeit von den verschiedenen Ausfallmechanismen vergleichen.

3 Vergleichsstudie

Um die unter Anwendung von MI erzielten hervorragenden Ergebnisse besser einordnen zu können, haben wir MI in einer weiteren Simulationsstudie mit den im Folgenden kurz beschriebenen Verfahren, unterteilt nach unseren vier Ausfallmechanismen, verglichen.

3.1 Verfahren

Der wohl einfachste Fall ist die *available case*-Analyse. Hierbei streichen wir jede Zeile unseres Datensatzes, in der x gelöscht wurde. Mit diesem reduzierten Datensatz führen wir dann eine *complete case*-Analyse durch, d.h. wir behandeln die übrig bleibenden Daten wie einen vollständigen Datensatz. Ein einfaches Ergänzungsverfahren stellt die *mean imputation* dar. Dabei ergänzen wir jeden fehlenden Wert von x durch den Mit-

telwert der beobachteten Daten. Des Weiteren haben wir zwei *single imputations* implementiert. Zum einen führen wir MI mit nur einer Imputation durch, d.h. m ist gleich 1. Zum anderen betrachten wir die vor dem Einsatz von MI verwendete *single imputation*. Dabei werden die fehlenden Werte aus $N(U_{mis}\hat{ß}_{obs};\hat{\sigma}^2_{obs})$ gezogen und nicht wie bei der klassischen MI (auch für $m=1$) aus $N(U_{mis}ß;\sigma^2)$. Im nächsten Kapitel stellen wir unsere Ergebnisse der Vergleichsstudie, unterteilt in die vier Fälle MAR, NMAR allgemein, NMAR Stutzung und MCAR, dar.

3.2 Ergebnisse

Wir haben bei allen Verfahren wiederum eine Ausfallrate von etwa 15% gewählt und 500 Durchläufe bei einer Datensatzgröße von 2000 simuliert. In den Tabellen 3.1 und 3.2 sind die Durchschnittswerte des Schätzers für $ß_1$ bzw. die entsprechende 95%-ige *coverage* dargestellt, unterteilt nach den verschiedenen Verfahren und Ausfallmechanismen. Die Abbildungen 3.1 und 3.2 im Anhang zeigen die entsprechenden Histogramme der $\hat{ß}_1$ für die beiden Ausfallmodelle MAR und NMAR Stutzung.

	MAR	NMAR allg.	NMAR Stutzung	MCAR
BD	0.4971	0.4996	0.4959	0.5033
AC	0.2048	0.5009	0.4951	0.5043
Mean Imputation	0.5790	0.5696	0.6296	0.5510
SI	0.4953	0.5242	0.5928	0.5034
MI (m=1)	0.4962	0.5238	0.5918	0.5031
MI (m=5)	0.4963	0.5236	0.5918	0.5033

Tabelle 3.1: Durchschnittswerte für $\hat{ß}_1$ im Vergleich

	MAR	NMAR allg.	NMAR Stutzung	MCAR
BD	0.964	0.952	0.958	0.934
AC	0.206	0.936	0.95	0.952
Mean Imputation	0.868	0.868	0.74	0.914
SI	0.948	0.93	0.848	0.936
MI (m=1)	0.95	0.93	0.852	0.934
MI (m=5)	0.956	0.934	0.858	0.932

Tabelle 3.2: *Coverage* von $ß_1$ im Vergleich

Betrachten wir zunächst die Ergebnisse, wenn wir mit MAR gelöscht haben. In Abbildung 3.1 sehen wir, dass der Schätzer für den Regressionskoeffizienten von E unter *available case* stark vom wahren Wert $ß_1 = 0.5$ abweicht. Dementsprechend haben wir eine sehr geringe *coverage* von nur etwa 21% (vgl. Tabelle 3.2). Auch Raghunathan hat in [Ra04] diesen Fall betrachtet und vergleichbare Ergebnisse erzielt, sich aber hierbei wie oben erwähnt auf einen MAR-Datenausfall und den Vergleich von AC und MI beschränkt. Führen wir eine *mean imputation* durch, erhalten wir zwar eine wesentlich bessere *coverage* als bei AC (vgl. Tabelle 3.2), welche aber aufgrund der unterschätzten Varianzen deutlich schlechter ist als bei den anderen betrachteten Ergänzungsverfahren. Die restlichen Ergänzungsverfahren (SI, MI mit $m = 1$ und $m = 5$) hingegen liefern unverzerrte Schätzer und eine hohe *coverage* von rund 95%. In unserem Beispiel sind diese drei Ergänzungsverfahren gleichwertig. Sie ergeben nahezu identische Schätzer und Überdeckungen, wobei MI mit $m = 5$ die höchste *coverage* aufweist. Liegt demnach MAR als Ausfallmechanismus vor, welches der in der Praxis am häufigsten auftretende Fall ist, so ist es sinnvoll, SI oder noch besser MI zu verwenden. In unserem sehr kleinen Modell mit nur zwei unabhängigen Variablen sind SI und MI gleichermaßen geeignet. In anderen Studien (z.B. [RR06]) zeigt sich der Vorteil von MI gegenüber SI allerdings deut-

licher. Nicht zu empfehlen ist hingegen die Durchführung einer *available case*-Analyse. Der Aufwand wäre zwar wesentlich geringer als bei MI, aber wie wir gesehen haben, liefert AC im Fall MAR unvalide Ergebnisse. Der Vorteil von Ergänzungsverfahren wie MI ist hierbei deutlich zu erkennen. Unterliegt der Datenausfall MCAR, was in der Praxis eher selten vorkommt, schneiden erwartungsgemäß alle Verfahren sehr gut ab. Die höchste *coverage* wird hierbei mit einer *availabl case*-Analyse erzielt, da nur eine zufällige Teilstichprobe aus unserem generierten Datensatz betrachtet wird und keine Werte ergänzt werden. Die Ergänzungsverfahren SI und MI mit $m = 1$ und $m = 5$ sind wieder als gleichwertig anzusehen. Sie liefern mit 93% eine nur geringfügig schlechtere Überdeckung als AC mit 95%, obwohl die fehlenden Werte ergänzt wurden und wir dabei im Imputationsmodell MAR unterstellt haben (vgl. auch Kapitel 4). Nur die *mean imputation* erzeugt etwas schlechtere Ergebnisse, ist mit einer *coverage* von etwa 91% aber immer noch vertretbar. Da *mean imputation* durch verzerrte Schätzer und die unterschätzte Varianz schlechte Ergebnisse für alle Ausfallmechanismen produziert und somit als sinnvolles Ergänzungsverfahren ausscheidet und sich SI, MI mit $m = 1$ und MI mit $m = 5$ in den Ergebnissen stark ähneln, werden wir im Folgenden die Resultate für die beiden NMAR-Fälle nur für AC und MI mit $m = 5$ diskutieren.

Sofort fällt auf, dass die *available case*-Analyse hervorragende Schätzer liefert. Besonders deutlich erkennt man dies im Spezialfall der Stutzung. Hier wird der wahre Wert $ß_1 = 0.5$ bei AC nahezu exakt getroffen, während der MI-Schätzer mit $\hat{ß}_{1_{MI}} = 0.5918$ um fast 0.1 vom wahren Wert abweicht. Das gleiche Bild ergibt sich bei den Überdeckungen. AC liefert beim Datenausfall durch NMAR mit 93.6% eine sehr hohe *coverage*. Diese ist genauso gut, als ob wir mit MCAR gelöscht hätten statt mit NMAR. Insbesondere im Spezialfall der Stutzung erkennt man, dass eine *available case*-Analyse mit 95% wesentlich bessere Überdeckungen liefert als eine *multiple imputation*, die aufgrund der verzerrten Schätzer nur eine *coverage* von 85.8% aufweist.

Was zunächst sehr verwunderlich erscheint, lässt sich wie folgt erklären: Betrachten wir die Likelihoodfunktion, die zur Analyse der vollständigen Daten verwendet wird:

$$L(ß; x, E, D) = \prod_i f_{D,E,x|ß}(d_i, e_i, x_i \mid ß)$$
$$= \prod_i f_{D|E,x,ß}(d_i \mid e_i, x_i, ß) \cdot f_{E|x,ß}(e_i \mid x_i, ß) \cdot f_{x|ß}(x_i \mid ß),$$

wobei $f_{D|E,x,ß}(d_i \mid e_i, x_i, ß)$ der relevante Teil der Likelihoodfunktion ist. Betrachten wir nun den Datenausfall mit NMAR, d.h. das Fehlen der Werte hängt nur von x selbst ab. Die für die *available case*-Analyse verwendete Likelihoodfunktion hat dann folgende Form:

$$L(ß; x, Z, E, D) = \prod_i f_{D|E,x,ß}(d_i \mid e_i, x_i, ß) \cdot f_{E|x,ß}(e_i \mid x_i, ß) \cdot$$
$$\cdot f_{Z|x,ß}(z_i \mid x_i, ß) \cdot f_{x|ß}(x_i \mid ß),$$

wobei Z eine Zufallsvariable ist, die nur von x abhängt, und $f_{Z|x,ß}(z_i \mid x_i, ß)$ die Verteilung von x nach dem Löschen darstellt. Wir sehen, dass sich der relevante Teil $f_{D|E,x,ß}(d_i \mid e_i, x_i, ß)$, also der Teil der Likelihoodfunktion in dem D zufällig ist, nicht verändert hat. Daher erhalten wir durch eine *available case*-Analyse unverzerrte Schätzer und dementsprechend hohe Überdeckungen, welche so gut sind, als hätten wir vollständige Daten analysiert. Es bleibt nun noch die Frage, warum ein Datenausfall mit MAR hingegen bei AC die erwartungsgemäß schlechten Schätzer liefert. Dazu betrachten wir auch für diesen Fall die entsprechende Likelihoodfunktion. Löschen wir mit MAR, d.h. der Datenausfall ist abhängig von D (und E), so hat die Likelihoodfunktion folgende Form:

$$L(ß; x, E, D, Z) = \prod_i f_{Z|D,E,x,ß}(z_i \mid d_i, e_i, x_i, ß) \cdot f_{D|E,x,ß}(d_i \mid e_i, x_i, ß) \cdot$$
$$\cdot f_{E|x,ß}(e_i \mid x_i, ß) \cdot f_{x|ß}(x_i \mid ß),$$

wobei Z eine Zufallsvariable ist, die von D, E und x abhängt und $f_{Z|D,E,x,ß}(z_i \mid d_i, e_i, x_i, ß)$ die Verteilung von x nach dem Löschen darstellt. Wir sehen nun, dass sich der relevante Teil dieser Likelihoodfunktion im Vergleich zur Likelihoodfunktion des datengenerierenden Pro-

zesses verändert hat. Der von D abhängige Teil ist nun $f_{Z|D,E,x,ß}(z_i \mid d_i, e_i, x_i, ß) \cdot f_{D|E,x,ß}(d_i \mid e_i, x_i, ß)$, d.h., dass wir hier eine andere Verteilungsfamilie vorliegen haben. Somit erhalten wir verzerrte Schätzer und dementsprechend niedrige Überdeckungen.
Betrachten wir demnach ein Analysemodell bei welchem die interessierende, also die abhängige Variable, **nicht** gleichzeitig die *missing*-Variable ist, d.h., dass die abhängige Variable vollständig beobachtet wurde, und der Datenausfall nach NMAR erfolgte, insbesondere durch Stutzung, so ist die *available case*-Analyse einem Ergänzungsverfahren vorzuziehen.

4 (Un)Congeniality

Der Begriff *congeniality* wurde von Xiao-Li Meng eingeführt und 1994 in „Multiple-Imputation Inferences with Uncongenial Sources of Input" [Me94] definiert. Unter *congeniality* versteht Meng die Übereinstimmung von Imputations- und Analysemodell und beschäftigt sich im Wesentlichen mit der Frage, was passiert, wenn der Imputer mehr Annahmen trifft als der Analyst und umgekehrt. Aufbauend auf der Problematik der Gleichartigkeit von Modellen wollen wir den Begriff der *congeniality* nun erweitern und uns mit verschiedenen Modellübereinstimmungen und den Auswirkungen bei Nichtübereinstimmung beschäftigen.
Gegeben sind die eingangs erwähnten vier Modelle - der datengenerierende Prozess, das Analysemodell, der Ausfallmechanismus und das Imputationsmodell. Aus diesen lassen sich fünf Modellpaare bilden, für die sich der Begriff der Gleichartigkeit sinnvoll definieren und untersuchen lässt. Dabei müssen wir abhängig vom jeweiligen Modellpaar verschiedene Arten der Übereinstimmung unterscheiden. In Abbildung 4.1 ist eine Übersicht für diese Paare und die entsprechende Art der Übereinstimmung dargestellt.

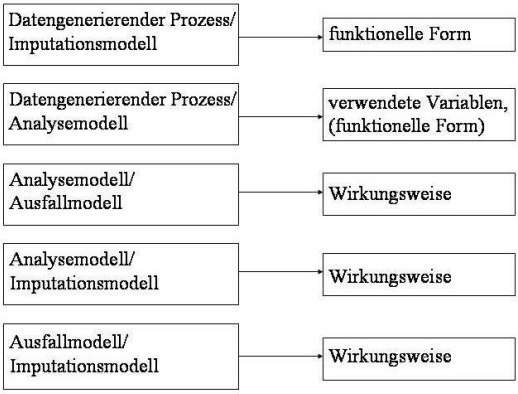

Abbildung 4.1: Gleichartigkeit für die verschiedenen Modellpaarungen

Im Folgenden wollen wir die einzelnen Formen der Gleichartigkeit näher erläutern und untersuchen.

4.1 Datengenerierender Prozess versus Imputationsmodell

Durch die Datenerhebung bzw. Generierung bei Simulationen sind uns die Daten fest vorgegeben. Aufgabe des Imputers ist es, das passende Imputationsmodell zu finden. Hierbei kommt es darauf an, wie die Variablen, die fehlende Werte enthalten, jeweils verteilt sind. Wie wir in Abbildung 4.2 sehen, unterscheidet man in stetige, binäre und sonstige Variablen.

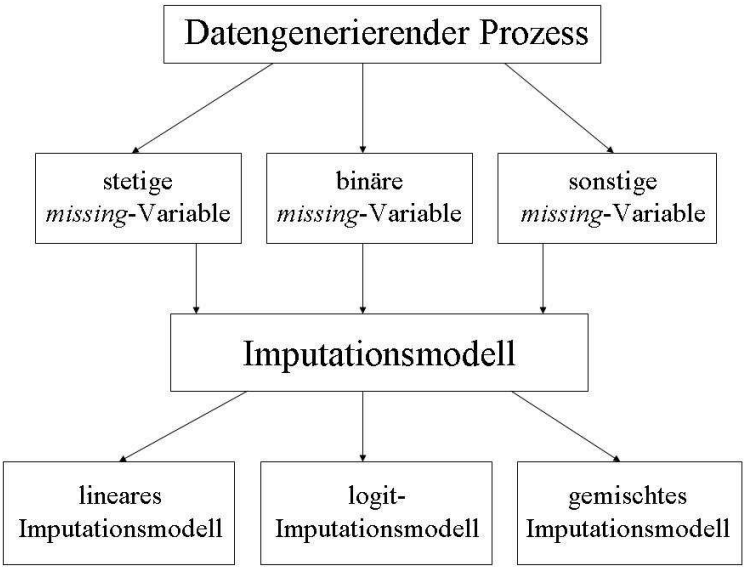

Abbildung 4.2: Wahl des passenden Imputationsmodells

In die letzte Gruppe fallen diskrete Daten, semi-stetige (großer Anteil an Nullen und ein stetiger, ordinalskalierter Rest) Daten und mehrfach kategoriale Variablen, die wie binäre Daten, aber sequentiell nach Kategorien unterteilt, behandelt werden. Für stetige Daten wird z.B. ein lineares Regressionsmodell verwendet, für binäre Daten eine logistische Regression und für sonstige Variablen ein Mix aus beidem. In unserem medizinischen Beispiel haben wir lediglich fehlende Daten in der stetigen Variable x. Somit ist unser Imputationsmodell $x = \gamma_0 + \gamma_1 \cdot D + + \gamma_2 \cdot E + \gamma_3 \cdot D \cdot E + \varepsilon$ eindeutig vorgegeben, und wir verwenden den in Kapitel 2.3 beschriebenen MI-Algorithmus für fehlende stetige Daten. Übereinstimmung in den Variablen oder der Wirkungsweise wie bei anderen Modellpaarungen als Definition heranzuziehen ist nicht sinnvoll, da die Wahl des Imputationsmodells von der Verteilung der *mis-*

sing-Variable abhängt und nicht von den zur Erzeugung verwendeten oder später zur Analyse herangezogenen anderen Variablen.
Gleichartigkeit von Datenmodell und Imputationsmodell in unserem Sinne ist also im Grunde genommen Voraussetzung und somit eine pathologische Definition. Beide Modelle stimmen demnach überein, wenn der Imputer das der Verteilung der *missing*-Variable(n) entsprechende Modell wählt, also wenn Gleichartigkeit in der funktionellen Form vorliegt. In unserem Beispiel ist die *missing*-Variable x standardnormalverteilt. Unabhängig davon, ob wir einen MAR-, NMAR- oder MCAR-Ausfall modellieren wollen, wählen wir ein stetiges Imputationsmodell. Auswirkungen von Nichtübereinstimmung zu untersuchen ist bei dieser Modellpaarung offensichtlich nicht sinnvoll.

4.2 Datengenerierender Prozess versus Analysemodell

Idealerweise sollten Datenmodell und Analysemodell völlig übereinstimmen, sie sollten sich also in den verwendeten Variablen und in der funktionellen Form entsprechen. Dabei ist die Übereinstimmung in der funktionellen Form, d.h. die richtige Wahl des zur Verteilung der interessierenden Variablen passenden Modells wie schon in Kapitel 4.1 Voraussetzung für eine valide Analyse. Wir können den Begriff der Gleichartigkeit also durch Gleichheit von Modellen ersetzen.
In unserem Simulationsbeispiel stimmen der datengenerierende Prozess $logit(p(D=1|x,E)) = -0.5 + 0.5 \cdot E + 0.5 \cdot x$ und das Analyse-modell $logit(p(D=1|x,E)) = ß_0 + ß_1 \cdot E + ß_2 \cdot x$ natürlich in jeglicher Hinsicht überein. Auswirkungen von Nichtübereinstimmungen sind leicht auszumachen. Bekanntlich ist es ungünstig, zu wenige Variablen in das Analysemodell aufzunehmen, da dies verzerrte Schätzer liefert. Nehmen wir hingegen zu viele Variablen auf, bekommen wir immerhin unverzerrte Schätzer, weshalb es besser ist zu viele als zu wenige Variablen in das Modell aufzunehmen. Diese Schätzer sind erwartungstreu, aber nicht mehr effizient, d.h. wir erhalten überschätzte Varianzen. Diese Fälle der Nichtübereinstimmung brauchen wir demnach nicht auf die Wirkungsweise von *multiple imputation* zu untersuchen, denn wenn die Analyse mit vollständigen Daten schon verzerrte Schätzer oder überschätzte Varianzen liefert, so kann MI dies natürlich nicht kompensie-

ren. Vertauschen wir aber die Rollen von D und E im datengenerierenden Prozess, d.h.:

- $x \sim N(0,1)$
- $logit\,(p(D=1|\,x)) = 0.25 + 0.75 \cdot x$
- $\text{logit}\,(p(E=1|\,x,D)) = -0.5 + 0.5 \cdot D + 0.5 \cdot x,$

behalten das Analysemodell $logit\,(p(D=1|\,x,E)) = ß_0 + ß_1 \cdot E + ß_2 \cdot x$ hingegen bei, so liegt Übereinstimmung in den verwendeten Variablen D, E und x und natürlich in der funktionalen Form vor. Die Analyse der vollständigen Daten liefert allerdings eine extrem unterschätzte Varianz für $ß_1$. Es wurde offensichtlich eine wichtige Voraussetzung der Regressionsanalyse verletzt, denn die erklärende Variable E ist nicht unabhängig von der erklärten Variablen D. Von Übereinstimmung können wir also nur dann sprechen, wenn die entsprechenden Voraussetzungen an das Analysemodell erfüllt sind.

4.3 Analysemodell versus Ausfallmechanismus

Da in unserem gewählten Beispiel der Datenausfall in der unabhängigen Variablen x vorliegt und nicht in der zu analysierenden abhängigen Variablen D, können wir Gleichartigkeit für diese beiden Modelle nicht durch Übereinstimmung in den Variablen und/oder der funktionellen Form definieren. In Kapitel 2.2 haben wir die drei bzw. vier Ausfallmechanismen MAR, MCAR und NMAR mit dem Spezialfall der Stutzung kennengelernt. Während unser Analysemodell $logit\,(p(D=1|x,E)) = ß_0 + ß_1 \cdot E + ß_2 \cdot x$ alle Variablen enthält, ist im Ausfallmodell die Verwendung der Variablen vom jeweiligen Mechanismus abhängig. Beispielsweise hängt der Ausfallmechanismus MCAR von keiner Variablen ab, sondern ist rein zufällig. Über die funktionelle Form zu gehen ist bei diesen beiden Modellen nicht sinnvoll. In der Analyse verwenden wir ein *logit*-Modell für binäre Variablen, während der Ausfallmechanismus auf verschiedene Arten eine Indikatorvariable *mis* für den Datenausfall erzeugt. Beide Modelle sind also hinsichtlich der verwende-

ten Variablen und der funktionellen Form nicht sinnvoll vergleichbar, da die zu analysierende Variable D nicht identisch mit der *missing*-Variable x ist.
Wir können Gleichartigkeit aber durch die Übereinstimmung in der Wirkungsweise definieren. Analyse- und Ausfallmodell sind gleichartig, wenn der Analyst das zu seinen Annahmen bzw. (Zusatz)Informationen passende Verfahren zum Umgang mit fehlenden Daten wählt. Weiß der Analyst, dass die vorliegenden Daten mit NMAR ausgefallen oder sogar gestutzt sind und die abhängige Variable nicht identisch mit der ergänzten Variable ist, so sollte er die *available case*-Analyse den Ergänzungsverfahren vorziehen. Hat er keine Informationen über den Datenausfall, sollten multipel imputierte Daten verwendet werden, wobei der Imputer ein MAR-Imputationsmodell wählen sollte, wie wir in Kapitel 4.5 sehen werden. Analyse- und Ausfallmodell sind dann in unserem Sinne gleichartig. Weiß der Analyst nichts über den Datenausfall und wählt trotzdem die *available case*-Analyse oder liegt Stutzung vor und er wählt trotz der Kenntnis darüber die *multiple imputation*, so sind beide Modelle ungleichartig. Ergebnisse dieser Nichtübereinstimmung haben wir in den Tabellen 3.1 und 3.2 bereits dargestellt.

4.4 Analysemodell versus Imputationsmodell

Meng hat in [Me94] den Begriff der *congeniality* des Imputationsmodells zu einem Analysemodell eingeführt. Er definiert Gleichartigkeit als Übereinstimmung der verfügbaren Informationen von Analyst und Imputer. Anhand zweier Beispiele beschreibt er, was passiert, wenn der Imputer mehr Informationen (z.B. über den Ausfallmechanismus) hat bzw. mehr Annahmen trifft als der Analyst und umgekehrt. Dabei vergleicht er die Güte einer *multiple imputation* mit der *available case*-Analyse. Im Prinzip betrachtet Meng genau die Kombination der drei Modelle „Ausfall-Analyse-Imputation“ und erweitert das Ausfallmodell zu einem allgemeineren Informationsmodell.
In unserem konkreten Beispiel ist diese Definition allerdings etwas problematisch. Da wir die Variable D in Abhängigkeit von den unabhängigen Variablen E und x untersuchen, der Datenausfall aber in x vorliegt, differieren Analyse- und Imputationsmodell zwangsläufig in

jeglicher Hinsicht. Meng betrachtet bei seinen Untersuchungen jedoch solche Modelle bei denen die zu analysierende Variable auch gleichzeitig die zu ergänzende Variable ist. D.h., dass er zunächst in der Funktionsweise gleichartige Modelle voraussetzt und die Auswirkung von Nichtübereinstimmung unter Einbeziehung des Ausfallmodells bzw. des allgemeineren Informationsmodells untersucht. Gleichartigkeit bezieht sich bei Meng demnach auf die Übereinstimmung in den verwendeten Informationsmodellen. Dabei können sich unterschiedliche Informationsmodelle auch auf die Struktur des verwendeten Analyse- bzw. Imputationsmodells auswirken. Beide Modelle stimmen dann zwar - wie vorausgesetzt - in ihrer Funktionsweise überein, aber nicht in den verwendeten Variablen. Sie sind somit nicht *congenial*.

Da wir allerdings in unserem Beispiel im Analysemodell eine andere Variable untersuchen als im Imputationsmodell können wir Mengs Definition hier nicht anwenden. Wir definieren Gleichartigkeit für die vorliegenden Modellstrukturen nun wie folgt: Analyse- und Imputationsmodell sind gleichartig, wenn sich der Analyst unter dem Wissen, dass der Datenausfall NMAR war eine *available case*-Analyse durchführt. Entscheidet sich der Analyst ohne weitere Informationen über den Datenausfall für die *available case*-Analyse, so sind beide Modelle ungleichartig. Führt der Analyst bei unbekanntem Datenausfall hingegen eine MI durch, so stimmt das Imputationsmodell in unserem Sinne mit dem Analysemodell überein, welches wiederum, wie in Kapitel 4.3 beschrieben, auch ungleichartig zum Ausfallmodell sein kann. Das heißt aber nicht, dass deshalb auch Ausfall- und Imputationsmodell ungleichartig sind. Wie wir sehen, haben auch wir hier eine Kombination von verschiedenen „Gleichartigkeiten" der drei Modelle Analyse-Ausfall-Imputation. Der Datenausfall spielt also stets eine Rolle.
In der Realität sind Analyst und Imputer nur selten identisch, weshalb sich Ihre Modelle zwangsläufig unterscheiden. Durch die Bereitstellung von *public use files* mittels Imputation gibt es sogar eine Vielzahl von verschiedenen Anaylsemodellen zu einem Imputationsmodell. Zudem kann beispielsweise der Imputer aus Datenschutzgründen nicht alle zur Ergänzung verwendeten Informationen weitergeben. Zielen beide Mo-

delle auf die gleiche Variable ab, so kann man wie bei Mengs Untersuchungen zur *uncongeniality* auf das Informationsmodell und insbesondere auf das Ausfallmodell als „Gleichartigkeitskriterium“ zurückgreifen. Unterscheiden sich aber die Zielvariablen wie in unserem konkreten Beispiel, können wir den Datenausfall nur bedingt als Informationsmodell nutzen. Unser Analysemodell ist fest vorgegeben, d.h., dass sich die verschiedenen Ausfallmechanismen nicht in der funktionalen Form des Analysemodells widerspiegeln können. Der Datenausfall fließt daher nach unserer Definition lediglich bei der Wahl des Analyseverfahrens mit ein, d.h. ob eine *available case*-Analyse anzuwenden ist oder die Daten zunächst durch MI ergänzt werden sollten. In Kapitel 3 haben wir bereits die *available case*-Analyse und die *multiple imputation* für unterschiedliche Ausfallmechanismen miteinander verglichen und somit die Auswirkungen von Ungleichartigkeit beider Modelle nach unserer Definition aufgezeigt.

4.5 Ausfallmechanismus versus Imputationsmodell

Für diese beiden Modelle ist es sehr einfach, Gleichartigkeit in unserem Sinne zu definieren. Wir sagen, das Imputationsmodell ist gleichartig zum Ausfallmodell, wenn das Imputationsmodell genau den Ausfallmechanismus modelliert, mit dem die Daten ausgefallen sind. Beide Modelle sind demnach gleichartig, wenn sie in ihrer Wirkungsweise - bezogen auf den Ausfallmechanismus - gleich sind. In Kapitel 2 und 3 haben wir stets als Imputationsmodell $x = \gamma_0 + \gamma_1 \cdot D + \gamma_2 \cdot E + \gamma_3 \cdot D \cdot E + \varepsilon$ verwendet, welches einen MAR-Ausfall beschreibt, da die zu ergänzende Variable x von allen anderen (hier D, E und $D \cdot E$), aber nicht von sich selbst abhängig ist. Die zusätzliche Verwendung von $D \cdot E$ erklärt sich dadurch, dass es besser ist, zu viele als zu wenige Variablen ins Modell aufzunehmen. Als Faustregel gilt, dass alle möglichen Variablen im Imputationsmodell auftreten sollten. In unserem Beispiel haben wir genau wie im MAR-Ausfallmodell D, E und $D \cdot E$ als Modellvariablen. Weitere Kombinationen von D und E sind aber nicht sinnvoll, da beides binäre Variablen sind. In Kapitel 2 haben wir bereits implizit gesehen, was passiert, wenn wir einen anderen Ausfallmechanismus als MAR benutzen, das Imputationsmodell aber beibehalten. In diesem

Abschnitt wollen wir nun die Ergebnisse für alle möglichen Kombinationen von Ausfall- und Imputationsmodell zusammenfassen, welche wir anhand unseres Eingangsbeispiels in mehreren Studien simuliert haben. Sie sind am Ende des Kapitels noch einmal in Tabelle 4.1 als Übersicht dargestellt.
In unseren Studien haben wir wiederum 500 Durchläufe mit einer Datensatzgröße von 1000 verwendet. Für die vier Ausfallmechanismen MAR, NMAR allgemein, NMAR Stutzung und MCAR haben wir jeweils die fehlenden Werte in x mit den verschiedenen Imputationsmodellen ergänzt. Fassen wir diese noch einmal kurz zusammen:

Missing At Random (MAR)
Ausfallmechanismus:

$$p_{mis} = \frac{\exp(-1.11 - 1.09D - 1.85E + 2.31D \cdot E)}{1 + \exp(-1.11 - 1.09D - 1.85E + 2.31D \cdot E)}$$

Ist $mis \sim \text{Bernoulli}(p_{mis})$ gleich 1, so wird der entsprechende Wert in x gelöscht.

Imputationsmodell:

$$x = \gamma_0 + \gamma_1 \cdot D + \gamma_2 \cdot E + \gamma_3 \cdot D \cdot E + \varepsilon$$

Not Missing At Random (NMAR) – allgemein
Ausfallmechanismus:

$$p_{mis} = \frac{1}{1 + \exp(-x + 2.1)}$$

Ist $mis \sim \text{Bernoulli}(p_{mis})$ gleich 1, so wird der entsprechende Wert in x gelöscht.

Imputationsmodell:

$$p_{imp} = \frac{1}{1 + \exp(-t + 2.1)},$$

wobei t eine Realisation einer standardnormalverteilten Zufallsvariablen sei. Ist $imp \sim \text{Bernoulli}(p_{imp})$ gleich 1, so wird der fehlende Wert in x durch t ersetzt; sonst wird die Prozedur von vorn wiederholt.

Not Missing At Random (NMAR) - Stutzung

Ausfallmechanismus:

Ist der x-Wert größer als der festgelegte Entscheidungswert, so wird der entsprechende Wert in x gelöscht.

Imputationsmodell:

Wir ziehen für jeden fehlenden x-Wert solange aus einer Standardnormalverteilung bis diese einen Wert größer als der Entscheidungswert annimmt und ergänzen mit diesem den fehlenden x-Wert.

Missing Completely At Random (MCAR)

Ausfallmechanismus:

Ist $mis \sim \text{Bernoulli}(p_{mis})$ gleich 1, so wird der entsprechende Wert in x gelöscht, wobei p_{mis} die festgelegte Ausfallwahrscheinlichkeit bezeichnet.

Imputationsmodell:

Jeder fehlende Wert von x wird durch eine Realisation einer standardnormalverteilten Zufallsvariablen ersetzt.

In Tabelle 4.1 sind nun unsere Ergebnisse getrennt nach dem jeweiligen Ausfallmechanismus zusammengefasst.

Ausfall MAR				
Imputation	**MAR**	NMAR allg.	NMAR Stutzung	MCAR
$\widehat{\beta}_1$	0.5034	0.6629	0.6991	0.6146
coverage	0.95	0.78	0.682	0.868
Ausfall NMAR allgemein				
Imputation	MAR	**NMAR allg.**	NMAR Stutzung	MCAR
$\widehat{\beta}_1$	0.5277	0.5749	0.5632	0.598
coverage	0.952	0.922	0.922	0.88
Ausfall NMAR Stutzung				
Imputation	MAR	NMAR allg.	**NMAR Stutzung**	MCAR
$\widehat{\beta}_1$	0.6016	0.5807	0.5163	0.6279
coverage	0.884	0.912	0.948	0.858
Ausfall MCAR				
Imputation	MAR	NMAR allg.	NMAR Stutzung	**MCAR**
$\widehat{\beta}_1$	0.4984	0.6032	0.6250	0.5902
coverage	0.94	0.894	0.826	0.92

Tabelle 4.1: Ausfallmechanismus versus Imputationsmodell

Die Ergebnisse zeigen, dass man bei unbekanntem Ausfallmechanismus mit einem MAR-Imputationsmodell ergänzen sollte, da dies die besten Schätzer liefert und die höchste *coverage* erzielt. Lediglich bei gestutzten Daten schneidet das MAR-Modell schlechter ab. Hier wäre zumindest ein allgemeines NMAR-Ergänzungsmodell wegen der höheren *coverage* geeigneter. Noch besser wäre das Stutzungsmodell selbst. Generell gilt natürlich, dass man am besten gleichartige Modelle verwendet, wobei in der Praxis das Ausfallmodell meist unbekannt ist. Zudem sehen wir, dass beim MCAR-Ausfall eine Imputation mit MAR, also mit einem ungleichartigen Modell, bessere Ergebnisse liefert. Das liegt daran, dass durch eine MCAR-Imputation die bereits völlig zufällig ausgefallenen Werte ebenso zufällig ergänzt werden und sich diese Unsicherheit in der Güte des Schätzers widerspiegelt. Die Auswirkungen von (Un)Gleichartigkeit hierbei führen uns zu dem Schluss, dass - wie auch in der Praxis verwendet - ohne weiteres Wissen und weitere

Informationen über den Datenausfall ein MAR-Imputationsmodell unter Verwendung aller verfügbaren Variablen benutzt werden sollte.

5 Zusammenfassung und Ausblick

Anhand eines einfachen Beispiels untersuchten wir in mehreren Simulationsstudien die Auswirkungen der Anwendung von *multiple imputation* als geeignetes Verfahren zum Umgang mit fehlenden Daten. Dabei betrachteten wir zunächst, welche Ergebnisse wir mit MI im Vergleich zur Analyse mit vollständigen Daten erzielen können, wobei wir die vier verschiedenen Mechanismen des Datenausfalls simulierten. Danach widmeten wir uns dem Vergleich von MI mit anderen Verfahren zum Umgang mit fehlenden Daten. Unser Hauptaugenmerk lag dabei auf dem Vergleich von MI mit der in der Praxis noch häufig angewendeten *available case*-Analyse vor dem Hintergrund der vier verschiedenen Datenausfallmodelle. Unterliegt der Datenausfall einem NMAR-Modell oder sogar dem Spezialfall der Stutzung und ist die abhängige Variable nicht identisch mit der ergänzten Variable, wie in unserem Beispiel, so konnten wir feststellen, dass eine *available case*-Analyse in diesem Fall besser geeignet ist als ein Ergänzungsverfahren. Ist der Datenausfall hingegen unbekannt, so empfiehlt sich in jedem Fall eine MI, da insbesondere im realistischen Fall MAR eine *available case*-Analyse erwartungsgemäß äußerst verzerrte Schätzer und somit eine sehr geringe *coverage* hervorbringt, MI aber durchweg (mit vertretbaren Abstrichen im Stutzungsfall) hervorragende Ergebnisse liefert. Schließlich wendeten wir uns dem von Meng aufgezeigten Problem der *congeniality* (Gleichartigkeit) zu. Dabei erweiterten wir den Begriff auf die verschiedenen sinnvollen Modellkombinationen und untersuchten die Auswirkungen von Nichtübereinstimmung anhand unseres eingangs gewählten Beispiels mittels mehrerer Simulationen. Dabei konnten wir insbesondere feststellen, dass es gut ist, ein MAR-Imputationsmodell mit allen möglichen Variablenkombinationen zu wählen, wenn der Ausfallmechanismus - wie bei den meisten realen Datensätzen - unbekannt ist. Natürlich ist es am besten, außer im MCAR-Fall vielleicht, wenn die

gewählten Modelle übereinstimmen bzw. nach unseren Definitionen gleichartig sind. Allerdings konnten wir gerade beim MAR-Imputationsmodell sehen, dass auch Nichtübereinstimmung von Modellen zu sehr guten Ergebnissen führen kann.
Obwohl unser Beispiel nur sehr wenige Variablen beinhaltet, hat es dennoch Realitätsbezug, wie in Kapitel 2.4 beschrieben, und war für unsere Simulationsstudien sehr gut geeignet. Auf dem Gebiet der *multiple imputation* gibt es natürlich noch zahlreiche offene Fragestellungen. Denkbar sind beispielsweise weitere Modelle aus dem medizinischen Bereich - die Medizinstatistik ist ein junges Gebiet, welches vor dem Hintergrund einer immer älter werdenden Gesellschaft stetig an Bedeutung gewinnt. Das Gesundheitswesen hält ein breites Feld an Forschungsthemen im statistischen Bereich, insbesondere im Umgang mit Fragebögen und statistischen Erhebungen, bereit. Gesundheit ist und wird immer ein aktuelles Thema sein. Außerdem gibt es natürlich zahlreiche andere Einsatzmöglichkeiten für Ergänzungsverfahren im Allgemeinen und MI im Besonderen, wie zum Beispiel die aus der Ökonomie bekannten Minzergleichungen. Offen ist auch die Frage, wie MI bei komplexen Stichprobendesigns eingesetzt werden kann, und welche Auswirkungen Ungleichartigkeit sowohl in Mengs als auch in unserem Sinne bei solchen Modellen hat, und ob sich vielleicht eine allgemeingültige Definition für alle möglichen Modellkombinationen finden lässt. Gewiss ist jedenfalls, dass MI in der modernen und vor allem validen Survey-Statistik - zu Recht - immer stärker Einzug hält. Abschließend bleibt uns - wie bereits im Titel erwähnt - zu sagen: Multiple Imputation Seizes Surveys!

Anhang

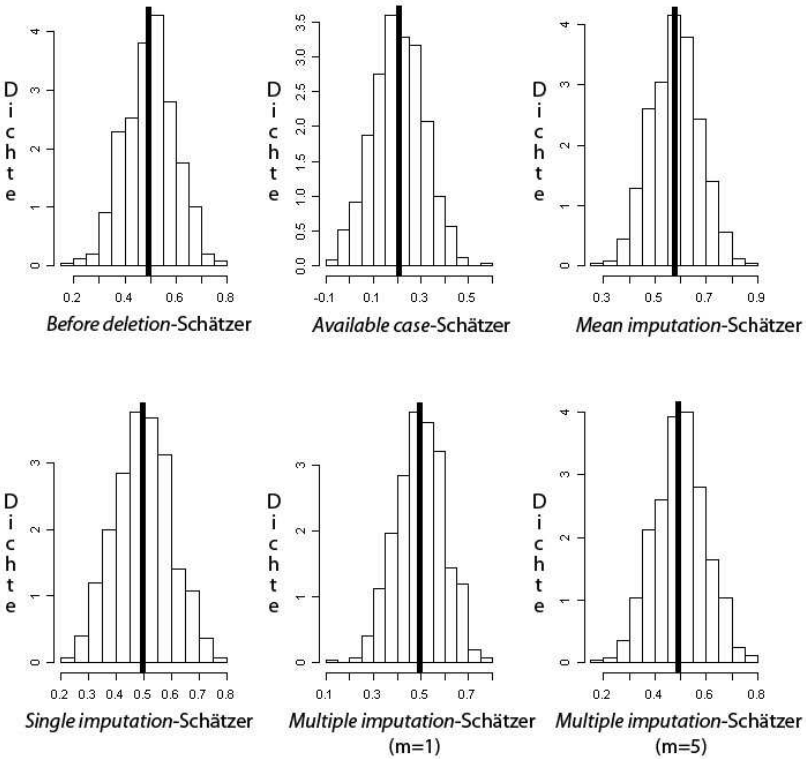

Abbildung 3.1: Histogramme der $\hat{\beta}_1$ für die 6 Verfahren im Fall MAR

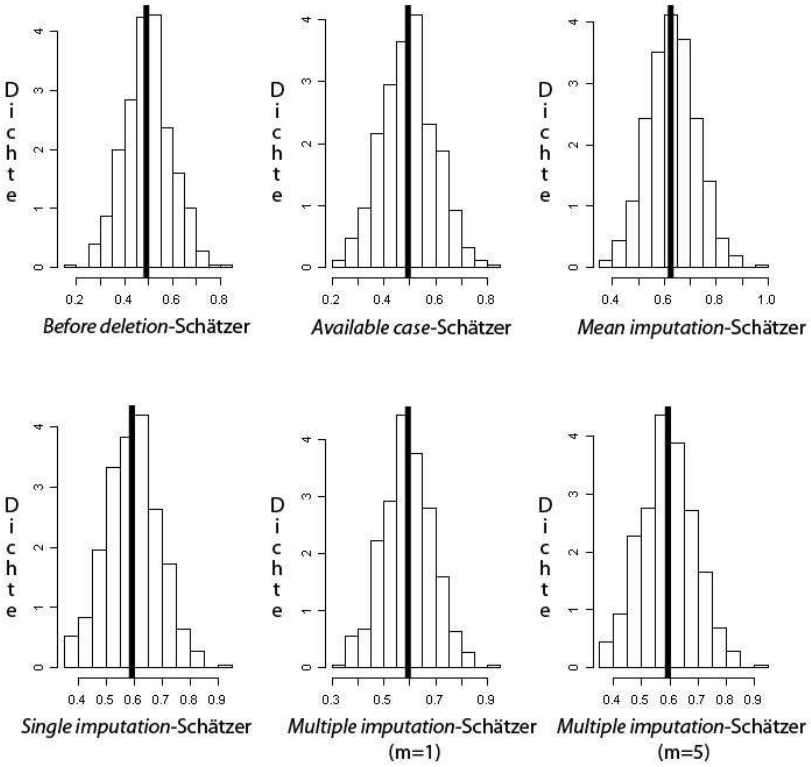

Abbildung 3.2: Histogramme der $\hat{ß}_1$ für die 6 Verfahren im Fall NMAR Stutzung

Literatur

[BT92] Box, G.E.P., Tiao, G.C. (1992). Bayesian Inference in Statistical Analysis. *John Wiley and Sons, New York*.

[LR02] Little, R.J.A., Rubin, D.B. (2002). Statistical Analysis with Missing Data. *John Willey and Sons, New York*, 2. Auflage.

[Li95] Little, R.J.A. (1995). Modeling the Drop-Out Mechanism in Longitudinal Studies. *Journal of the American Statistical Association*, 90, 1112 - 1121.

[Me94] Meng, X.L. (1994). Multiple-Imputation Inference with Uncon genial Sources of Input (with discussion). *Statistical Science*, 6, 538 - 573.

[RR06] Rässler, S., Riphan, R.T. (2006). Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv,* 90, 217 - 232.

[Ra04] Raghunathan, T.E. (2004). What do we do with missing data? Some options for analysis of inclomplete data. *Annu. Rev. Public Health*, 25, 99 - 117.

[Ru76] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, 63, 581 - 592.

[Ru78] Rubin, D.B. (1978) Multiple Imputation in Sample Surveys – a Phenomenological Bayesian Approach to Nonresponse. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20 - 40.

[Ru87] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. *John Wiley and Sons, New York.*

[Sch02] Schafer, J.R. (2002). Missing Data in Longitudinal Studies. W*orkshop Nürnberg, 6. September 2002.*

Bildnachweis

Abbildung 1:

Little, R.J.A., Rubin, D.B. (2002). Statistical Analysis with Missing Data. *John Willey and Sons, New York*, 2. Auflage, Figure 1.1, p. 5.

Abbildung 2:

Rässler, S. (2008). 30 Jahre Multiple Imputation. *Workshop Bamberg, 11. März 2008*, S. 36.