BACHELORARBEIT

Statistical Matching

- Multiple Imputation und Datenfusion am Beispiel von Daten zu Religiosität und Gesundheit

- Multiple Imputation and Data Fusion using the example of religiousness and health data

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
BACHELOR OF SCIENCE (B.SC.)

IN BETRIEBSWIRTSCHAFTSLEHRE

AM LADISLAUS VON BORTKIEWICZ CHAIR OF STATISTICS
DER WIRTSCHAFTSWISSENSCHAFTLICHEN FAKULTÄT
DER HUMBOLDT UNIVERSITÄT ZU BERLIN

Autor:
Sarah ASMAH
Matrikel-Nr.: 517859

Prüfer:
Prof. Dr. Ostap OKHRIN
Prof. Dr. Wolfgang HÄRDLE

Betreuer:
Dr. Sigbert KLINKE

8. November 2010

Inhaltsverzeichnis

Kapitel 1

Einführung

> Errors using inadequate data are much less than those using no data at all.
>
> Charles Babbage (ca. 1850)

Statistik hat in der Soziologie eine weitreichende Geschichte. Angefangen bei Volkszählungen, über Bevölkerungsstatistiken zu Geburten, Eheschließungen und Todesfällen, bis hin zu detaillierten Meinungsumfragen, steht heute eine Vielzahl an statistischen Methoden und Daten zur Verfügung, um soziologische Entwicklungen und Zusammenhänge zu erforschen. Häufig jedoch sind die Fragestellungen, die mithilfe einer Datenanalyse erörtert werden sollen, sehr komplex und zusätzlich ist die betreffende Grundgesamtheit oft sehr groß. Das hat zur Folge, dass eine eigens durchgeführte Datenerhebung, welche alle interessierenden Daten abfragt, sehr aufwändig ist. Um Zeit und Kosten zu sparen, kann auf eine Methode zurückgegriffen werden, die nur noch die Erhebung eines Teildatensatzes erfordert, welcher dann mit einem bestehenden Datensatz fusioniert wird, der die nicht abgefragten Daten enthält. Eine Technik dieser Datenfusion ist das Statistical Matching.

Ziel des Statistical Matchings ist die systematische Zusammenführung zweier Datensätze. Es ist ein nützliches Werkzeug, um Analysen über zwei (oder mehrere) Variablen zu machen, welche nicht im gleichen sondern in zwei unterschiedlichen Datensätzen vorliegen. Zur Veranschaulichung ist hier der Fall einer Untersuchung des Zusammenhangs zwischen den Variablen V_1 und V_2 betrachtet. Es liegt allerdings keine Umfrage vor, welche beide Variablen erfasst hat. Anstatt dessen liegt ein Datensatz D_1 mit der Variablen V_1 vor, und ein Datensatz D_2, in welchem unter anderem Daten der Variable V_2 erhoben wurden. Darüber hinaus finden sich gemeinsame Variablen, welche in beiden Datensätzen vorkommen, die aber nicht im Fokus der Analyse stehen. Es sei nicht davon auszugehen, dass in D_1 und D_2 die gleichen Personen befragt

wurden. Wie können also Zusammenhänge zwischen V_1 und V_2 analysiert werden, ohne dass auch nur eine einzige Person Aussagen zu beiden Variablen getroffen hat? Die Lösung liegt darin, dass die Datensätze anhand der in beiden Datensätzen vorkommenden Variablen fusioniert werden, so dass ein Datensatz entsteht, welcher sowohl V_1 als auch V_2 erfasst.

In dieser Arbeit werden sowohl die Theorie des Statistical Matchings als auch die vorbereitenden Schritte, die zum Matching notwendig sind, erörtert. An einem praktischen Beispiel soll die Anwendung jedes einzelnen Schrittes verdeutlicht werden.

Einleitend sollen im folgenden Kapitel zunächst das praktische Anwendungsbeispiel und seine Datenquelle erläutert werden. Nachfolgend wird in Kapitel 3 auf die Vorbereitung der Datensätze bei Vorliegen von Antwortausfall eingegangen. Eine ausführliche Beschreibung des Mechanismus von Statistical Matching findet sich in Kapitel 4. Zuletzt wird das praktische Anwendungsbeispiel durch einen Analyseansatz abgeschlossen (Kapitel 5) und sowohl ein Fazit der gewonnenen Erkenntnisse gezogen als auch ein Ausblick auf weiterführende Untersuchungen gegeben (Kapitel 6).

Kapitel 2

Vorstellung des Anwendungsbeispiels

> I don't think there's any doubt that people derive enormous comfort from religion, and they should continue to do that. What they shouldn't expect is that religious activity is going to promote their health.
>
> Richard Sloan et al. (2000)

In Zusammenarbeit des Lehrstuhls für Praktische Theologie und Religionspädagogik und des Ladislaus von Bortkiewicz Lehrstuhls für Statistik der Humboldt Universität zu Berlin sollte die Frage nach einem Zusammenhang zwischen „Religiosität" und „Gesundheit" erörtert werden. Eine solche Fragestellung wäre beispielsweise für eine Krankenkasse von Interesse, die sich auf die Versicherung kirchlicher Mitarbeiter spezialisiert hat. Als Datengrundlage für die oben genannte Untersuchung wurde ein Datensatz benötigt, der die beiden relevanten Themen „Religiosität" und „Gesundheit" erfasst hatte. Dies stellte sich als Problem heraus, da ein ebensolcher Datensatz nicht vorlag. Aus Versichertendaten einer Krankenversicherung hätte man prüfen können, ob sich das Krankenbild bei Versicherten in einem kirchlichen Amt anders verhält als bei den übrigen Versicherten. Derartige Informationen unterliegen jedoch dem Datenschutz und konnten somit nicht herangezogen werden. Darüber hinaus wurde in den Versichertendaten nicht Religiosität, sondern nur der ausgeübte Beruf (z. B. Priester) erhoben. Die Ausübung eines kirchlichen Amtes ist vielleicht ein Indikator, aber kein Maß für Religiosität.

Da also kein geeigneter Datensatz vorlag, musste ein ebensolcher mithilfe von Statistical Matching erzeugt werden.

Als Datenquellen wurden zwei Datensätze der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) herangezogen (GESIS, 2010). ALLBUS ist ein Datengenerierungsprogramm, welches regelmäßig Daten über Einstellungen, Verhaltensweisen und Sozialstruktur der Bevölkerung in der Bundesrepublik Deutschland

erhebt. Bei jedem Durchlauf des Programms (etwa alle zwei Jahre) wird eine repräsentative Stichprobe aus der Bevölkerung (etwa 3000 Personen) gezogen und in persönlichen Interviews befragt. Als Grundgesamtheit dieser Ziehung dient die erwachsene Wohnbevölkerung in West- und Ostdeutschland. Kurz nach der Wiedervereinigung (1992) wurde die Anzahl der Interviews auf 3500 erhöht, wovon 2400 Interviews in in West- und 1100 Interviews in Ostdeutschland geführt werden. Dies bedeutet, dass Bewohner der neuen Bundesländer in der ALLBUS-Stichprobe überrepräsentiert sind. In Anbetracht des Untersuchungsziel, kann dies in anschließenden Analyseschritten von Bedeutung sein, wenn angenommen wird, dass Religion in der ehemaligen DDR eine andere Rolle spielte als in Westdeutschland. Bei der Stichprobenziehung wird ein mehrstufiges Verfahren mit den Auswahlstufen „Gemeinden – Personen“ angewendet.

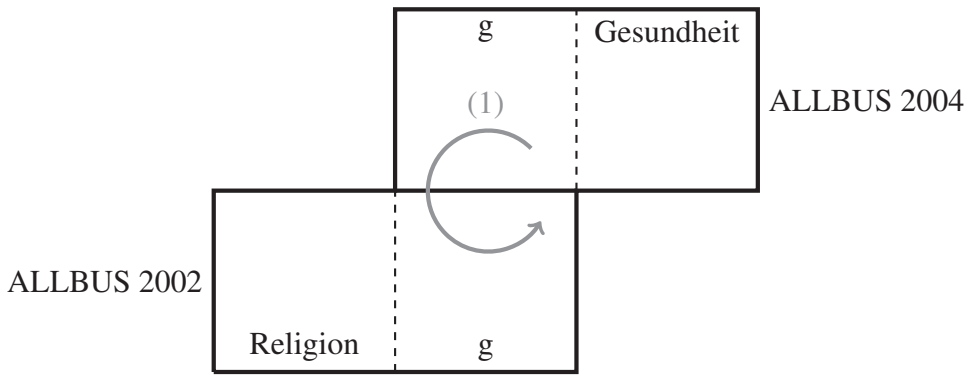
Nicht nur bei der Stichprobenziehung, sondern auch inhaltlich decken die ALLBUS Umfragen ein breites Feld ab. Neben konstanten Fragen, welche in jedem Durchlauf des Programms gestellt werden (beispielsweise Angaben zu „Alter“, Geschlecht“ oder „Haushaltsnettoeinkommen“), gibt es in jedem Fragenkatalog einen oder mehrere thematische Schwerpunkte, mit welchen sich etwa zwei Drittel der abgefragten Variablen befassen. Beispiele für ebensolche Schwerpunkte sind etwa „Politische Partizipation und politische Kultur“ (ALLBUS 2008[1]) oder „Soziale Ungleichheit“ (ALLBUS 2004[2]). Um auf einen Zusammenhang zwischen „Religiosität“ und „Gesundheit“ zu prüfen, wurden in dieser Arbeit die ALLBUS Datensätze mit den inhaltlichen Schwerpunkten „Religion und Weltanschauung“ (ALLBUS 2002[3]) und „Gesundheit und Digital Divide“ (ALLBUS 2004[4]) genutzt. Der ALLBUS-Datensatz aus 2002 enthält 2820 Fälle zu 722 Variablen, während der ALLBUS 2004 je 2946 Fälle zu 898 Variablen aufweist. 343 der Variablen liegen in beiden betrachteten Datensätzen vor. Allerdings kommen diejenigen Variablen, welche für diese Untersuchung interessant sind — also die Variablen zu den Themen „Religiosität“ und „Gesundheit“ — jeweils nur in einem der beiden Datensätze vor. Somit musste der erste Schritt vor Beginn einer Analyse die Zusammenführung beider Datensätze sein. Dies wäre eine relativ einfache Aufgabe gewesen, wenn davon auszugehen wäre, dass beide Datensätze aus der gleichen Stichprobe resultierten. Da aber bei jeder ALLBUS-Umfrage eine neue Stichprobe aus der Bevölkerung gezogen wird, beziehen sich die Antworten aus dem Datensatz aus 2002 real auf andere statistische Einheiten als die Daten aus der 2004er Umfrage. Daher mussten die beiden Datensätze mithilfe von Statistical Matching zusammengeführt werden, um einen gemeinsamen Datensatz zu erhalten, welcher alle für diese Untersu-

[1] vgl. http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/studienprofile/2008/
[2] vgl. http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/studienprofile/2004/
[3] CD: Kapitel 2 – Ausgangsdatensätze/ALLBUS2002
[4] CD: Kapitel 2 – Ausgangsdatensätze/ALLBUS2004

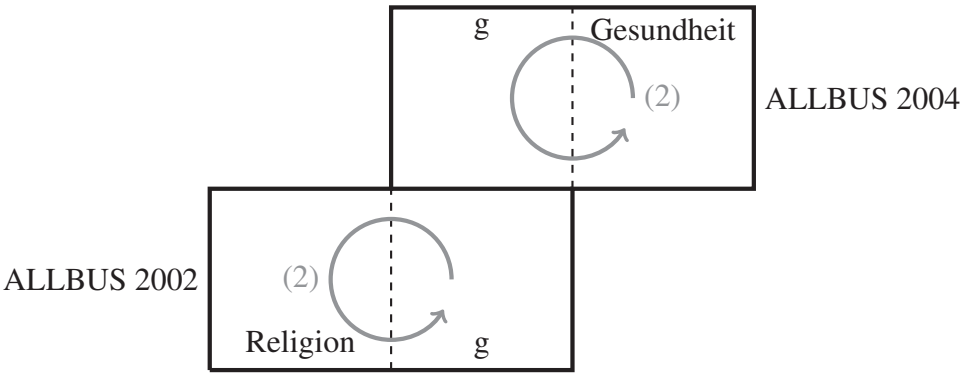
chung relevanten Variablen erfasste.

Die Vorgehensweise in dieser Arbeit geschah in den folgenden drei Schritten, welche schematisch in Abbildung 2.1 dargestellt sind: Da in den betrachteten Datensätzen Antwortausfall vorlag, wurden zunächst die fehlenden Werte der gemeinsamen Variablen geschätzt und imputiert, um einen vollständigen Datensatz der gemeinsamen Variablen zu erhalten (vgl. Abb. 2.1a). Danach wurden auch die spezifischen Variablen zu „Religiosität“ und „Gesundheit“ im jeweiligen Datensatz mithilfe der ergänzten gemeinsamen Variablen vervollständigt (vgl. Abb. 2.1b). Im letzten Schritt sollten die spezifischen Variablen durch das Matching-Verfahren im jeweils anderen Datensatz ergänzt werden (vgl. Abb. 2.1c).

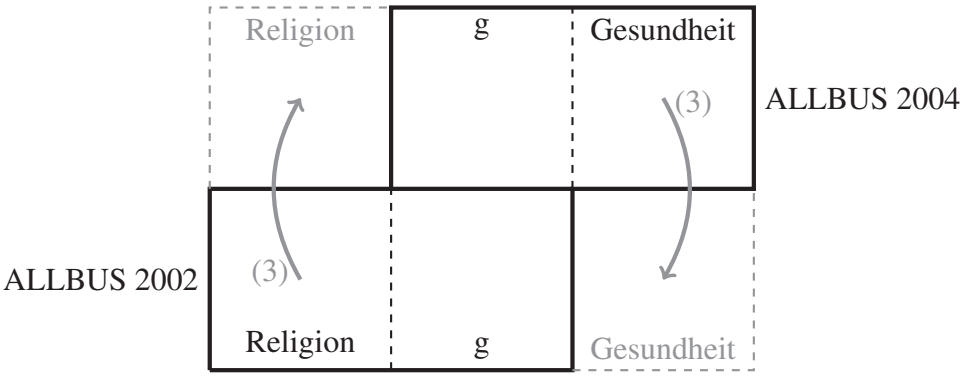
Im Fokus dieser Arbeit steht dabei nicht die Analyse des Zusammenhangs zwischen „Religiosität“ und „Gesundheit“ selbst. Vielmehr soll die Datenvorbereitung, welche eine solche Analyse überhaupt erst möglich macht, im Vordergrund stehen.



(a) Schritt 1: Imputation innerhalb der gemeinsamen Variablen



(b) Schritt 2: Imputation der spezifischen mithilfe der gemeinsamen Variablen



(c) Schritt 3: Datenfusion mithilfe von Statistical Matching

Abbildung 2.1: Schematische Darstellung der Vorgehensweise im Anwendungsbeispiel

Kapitel 3

Imputation der fehlenden Werte

> With or without missing data, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest – not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data.
>
> Schafer & Graham (2002)

Da die Fusion zweier Datensätze anhand ihrer gemeinsamen Variablen erfolgt, ist es wichtig, dass diese gut selektiert und möglichst vollständig sind. Daher sollte im ersten Schritt überlegt werden, welche dieser Variablen als Einflussvariablen in Frage kommen. Hierbei sind einerseits inhaltliche Zusammenhänge zwischen den gemeinsamen und den zu imputierenden Variablen zu prüfen. Zum anderen sind jedoch auch die Variableneigenschaften (wie z.B. fehlende Werte oder doppelte Informationen) entscheidend. Dieses Kapitel beschäftigt sich mit dem Problem des Antwortausfalls (= *Nonresponse*) und den Möglichkeiten, diesen durch die Imputation geschätzter Werte zu eliminieren.

Beim Umgang mit Datensätzen ist man häufig vor das Problem fehlender Werte gestellt und muss entscheiden, wie man damit umgeht. Das Fehlen von Information kann verschiedene Gründe haben. Neben den Fällen, in welchen Fragen einfach übersehen oder ihre Beantwortung vergessen wurden, kann auch Datenverlust durch technischen Ausfall ein Grund für fehlende Information sein. So können vollständig erhobene Daten beispielsweise im Nachhinein zu einem unvollständigen Datensatz führen, da bei der Datenübertragung zufällig Informationen verloren gegangen sind. Ein anderes Beispiel wäre eine pharmazeutische Langzeitstudie, bei der Patienten beispielsweise durch Todesfall aus der Studie ausfallen. Ob dieses Ausfallen zufällig auftritt oder nicht, ist in diesem Fall nicht so einfach zu bestimmen.

Auch mangelnde Antwortbereitschaft der Befragten kann ein Grund für fehlende

Werte sein. In den betrachteten ALLBUS-Daten schienen die Testpersonen je nach Inhalt mehr oder weniger bereit zu sein, eine Frage zu beantworten. So wies die Variable „Geschlecht“ 0 % fehlende Werte auf, während circa jeder sechste der Befragten (17,6 %) keine Angabe zum Haushaltsnettoeinkommen machen wollte. Es ist zu vermuten, dass diese Differenz dadurch begründet ist, dass eine Angabe über den Verdienst eine viel privatere ist als eine über das Geschlecht. Daher waren wahrscheinlich weniger Personen bereit, die Frage nach dem Einkommen zu beantworten. Zusätzlich ist anzunehmen, dass die Befragten mit einem höheren Einkommen häufiger eine Antwort zum Verdienst verweigern. Somit unterläge dem Antwortausfall eine Struktur, die von anderen Ausprägungen (beobachtet und/oder fehlend) abhängt.

Dieses Beispiel macht deutlich, dass die Gründe für fehlende Werte nicht ignoriert werden können, da dies eine Verzerrung der Analyse zur Folge hätte. Vielmehr muss die Struktur der fehlenden Werte in der Imputation berücksichtigt werden. Also sollte bei einem vorliegenden unvollständigen Datensatz zunächst geprüft werden, ob und welche Struktur hinter dem Fehlen von Werten steht.

Zum besseren Verständnis wird die Indikatorvariable M definiert, welche das Fehlen bzw. Nichtfehlen eines Wertes ausdrückt. M nehme bei Vorliegen einer fehlenden Ausprägung den Wert 1 an, die Ausprägung $M = 0$ stehe für keinen fehlenden Wert. Um die Struktur der fehlenden Werte zu bestimmen, soll die Verteilung von M nun näher definiert werden. Dabei geht es zunächst nicht darum, diese Verteilung genau zu benennen, sondern darum festzustellen, dass M eine Verteilung hat und wovon sie abhängt.

Desweiteren wird ein unvollständiger Datensatz Y_{com} definiert, welcher sich aus beobachteten Y_{obs} und nicht beobachteten Werten Y_{mis} zusammensetzt:

$$Y_{com} = (Y_{obs}, Y_{mis})$$

Die Arbeiten von Rubin (1987) und Little und Rubin (1987) stellen erstmals die folgende Klassifikation des Antwortausfallphänomens vor, welche in der Statistik weitgehend angenommen wurde:

MCAR Man spricht bei den Antwortausfällen von *Missing Completely At Random (MCAR)*, wenn die Wahrscheinlichkeit für das Fehlen der Werte im gesamten Datensatz ($P(M|Y_{com})$) weder von den beobachteten (Y_{obs}) noch von den nicht beobachteten (Y_{mis}) Werten abhängt und somit unabhängig vom Datensatz ist:

$$P(M|Y_{com}) = P(M)$$

Dies ist beispielsweise der Fall, wenn Daten zufällig verloren gegangen sind.

Ein realistisches Beispiel hierfür wäre in einer medizinischen Untersuchung der Verlust von Blutproben, weil im Labor willkürlich Proben herunter gefallen sind.

Ein solches Fehlen der Werte wird als *ignorierbar* betrachtet, da es von rein zufälliger Struktur ist.

MAR Die Daten sind *Missing At Random (MAR)*, wenn das Fehlen der Werte von den beobachteten, nicht aber von den nicht beobachteten Daten abhängt:

$$P(M|Y_{com}) = P(M|Y_{obs})$$

Dies bedeutet, dass die Wahrscheinlichkeit für das Fehlen eines Wertes in einer Variablen von einer anderen beobachteten Ausprägung einer Variablen abhängig ist. Dies ist beispielsweise der Fall, wenn festzustellen ist, dass die Antwortbereitschaft zur Variablen „Einkommen“ mit zunehmendem Alter abnimmt. Hier ist die Wahrscheinlichkeit für das Auftreten eines fehlenden Wertes in der Variablen „Einkommen“ abhängig von der beobachteten Variable „Alter“. Auch dieser Anwortausfall wird als ignorierbar bezeichnet, da er durch andere beobachtbare Werte erklärt werden kann.

An dieser Stelle sei auch erwähnt, dass es unter MAR einen Zusammenhang zwischen dem Fehlen der Werte ($P(M|Y_{com})$) und der nicht beobachtbaren Werte (Y_{mis}) geben kann, welcher jedoch durch ihren jeweiligen Zusammenhang zu den beobachtbaren Werten (Y_{obs}) zu erklären ist. Nach Miteinbeziehung von Y_{obs} verschwindet somit der Zusammenhang zwischen $P(M|Y_{com})$ und Y_{mis}.

Es lässt sich darüber streiten, ob der Ausdruck „Missing At Random“ hier treffend ist, da die Werte tatsächlich nicht zufällig fehlen. Die Terminologie nach Rubin hat sich jedoch durchgesetzt und ist weit verbreitet.

MNAR Hängt das Fehlen der Werte von den nichtbeobachteten Werten ab, so beschreibt man den Antwortausfall als *Missing Not At Random (MNAR)*:

$$P(M|Y_{com}) = P(M|Y_{mis})$$

Das Fehlen der Werte hängt also von den fehlenden Werten selbst ab. Beispielhaft ist hier der Fall, dass die Angabe über das Einkommen häufiger verweigert wird, je höher das Einkommen ist. Eine zusätzliche Abhängigkeit von den beobachteten Werten (Y_{obs}) ist hierbei nicht ausgeschlossen:

$$P(M|Y_{com}) = P(M|Y_{mis}, Y_{obs})$$

In diesem Fall ist die Struktur der fehlenden Werte nicht ignorierbar, da die Einflüsse auf die Wahrscheinlichkeit des Fehlens von nichtbeobachtbaren Werten abhängen.

In allen drei Fällen kann es selbstverständlich weitere externe Ursachen für den Antwortausfall geben. Ist beispielsweise ein Befragter aus gesundheitlichen Gründen weniger bereit, Fragen zu beantworten, oder hat eine Person einer Paper-und-Pencil-Befragung eine so unleserliche Schrift, dass die Antworten nicht interpretiert werden können, so stellen diese auch Gründe für die fehlenden Werte dar. Das bedeutet, dass auch im MCAR-Fall der Antwortausfall nicht zufällig sein muss, sondern nur von für den Untersuchungszusammenhang irrelevanten externen Variablen abhängt.

Schon die Definitionen von MCAR, MAR und MNAR lassen vermuten, dass die Verteilung der Missingness für das weitere Vorgehen relevant ist. Denn sollte es eine Fehlstruktur geben, so muss diese im Imputationsverfahren berücksichtigt werden. Um einen Datensatz auf MCAR zu testen, formulierten Kim und Curry (1977) einen χ^2-Test, welcher überprüft, ob das Fehlen von Werten zweier Variablen unabhänig voneinander ist. Little (1988) dagegen prüft mit einem Hypothesentest die Unabhängigkeit zwischen dem Fehlen der Werte und den vorhandenen Ausprägungen anderer Werte. Beide dieser Tests prüfen nur darauf, ob der Datensatz MCAR ist. Bei einer Ablehnung der Nullhypothese können keine Aussagen darüber gemacht werden, ob MAR oder MNAR vorliegt.

Je nach Einstufung des Antwortausfalls gibt es verschiedene Herangehensweisen im Umgang mit unvollständigen Datensätzen. Die wahrscheinlich einfachste Methode, um einen vollständigen Datensatz zu generieren, ist die sogenannte *Complete Case Analysis*. Hierbei werden alle statistischen Einheiten, welche in einer oder mehreren Variablen unvollständige Daten aufweisen, von der weiteren Analyse ausgeschlossen. Somit verbleibt ein zwar kleineres, doch komplettes Datenset. Die Vorteile der Complete Case Analysis liegen vor allem in ihrer leichten Anwendbarkeit. Das einfache Ausschließen von Fällen bedarf keinerlei Rechnung, sondern erstellt nur eine Teilmenge des Originaldatensatzes. Dieses Eliminierungsverfahren[5] ist in gängige statistische Software (z. B. SPSS) implementiert, da für viele computergestützte Analysen vollständige Datensätze vorausgesetzt werden.

Trotz der Einfachheit der Anwendung ist von dieser Methode in vielen Fällen abzuraten, da sie eine Menge Informationsverlust und weitere Nachteile in sich birgt. Insbesondere bei großen Mengen an Antwortausfall, ist die Frage nach der Brauchbarkeit des verbleibenden Teildatensatzes zu stellen. Ist der Datensatz noch repräsentativ,

[5]Obwohl die Complete Case Analysis nur eines der Eliminierungsverfahren darstellt, soll sie hier das einzig diskutierte bleiben. Somit werden in dieser Arbeit beide Begriffe synonym verwendet.

wenn beispielsweise 70 % der Fälle ausgeschlossen werden mussten? Zudem ergibt sich ein weiteres Problem, wenn die fehlenden Werte nicht MCAR sind. Gibt es also Zusammenhänge zwischen dem Antwortausfall und den Ausprägungen der Merkmale, so wird der Datensatz durch die Streichung dieser Fälle verzerrt und ist somit nicht mehr repräsentativ. Um dieser Verzerrung vorzubeugen, kann man bei einem Datensatz, bei welchem die fehlenden Werte nicht MCAR sind, die nach dem Eliminierungsverfahren verbleibenden Fälle gewichten, so dass sie der tatsächlichen Verteilung vor Streichung wieder mehr entsprechen. Das genaue Verfahren dieser Methode kann bei Little und Rubin (1987) nachgelesen werden. Ist das Fehlen der Werte MCAR, so hat der fallweise Ausschluss zwar keinen Einfluss auf die Verzerrung, doch wird durch das Ausschließen von Beobachtungen jegliche darauf basierende Schätzung weniger präzise, da sich die Varianz der Schätzfunktionen vergrößert.

Um die Aussagekräftigkeit im Datensatz zu erhalten ist eine gängigere Herangehensweise als die Streichung der unvollständigen Fälle die sogenannte *Imputation*, bei welcher die fehlenden Antworten durch plausible geschätzte Werte ersetzt werden. Die Imputation hat gegenüber der zuvor beschriebenen Complete Case Analysis den Vorteil, dass Einheiten mit fehlenden Werten nicht kategorisch aus der Analyse ausgeschlossen werden. Somit geht weniger Information verloren und es kann eine höhrere Präzision erreicht werden. Auch durch Anwendung von Imputation wird ein vollständiger Datensatz erzeugt, womit auch hier die Anwendbarkeit von Standardmethoden und Standardsoftware gegeben ist. All diese Vorteile greifen aber offensichtlich nur, wenn die Imputation zufriedenstellende Ergebnisse erzielt hat, d. h. wenn die imputierten Werte möglichst realistisch sind. Auch wenn dies schwer bis gar nicht zu prüfen ist, so ist doch anzunehmen, dass die Implementierung dieses Verfahrens präzisere Ergebnisse erzielt als die Complete Case Analysis (Schafer & Graham, 2002).

Bei der Imputation unterscheidet man zwischen der *singulären* und der *multiplen Imputation*. Diese beiden Methodengruppen unterscheiden sich grundlegend darin, dass bei der singulären Imputation für jeden fehlenden Wert ein Wert geschätzt und an entsprechender Stelle imputiert wird. Im Gegensatz hierzu wird bei der multiplen Imputation für jeden Antwortausfall gleich eine Reihe plausibler Werte geschätzt, welche dann zu einem Imputationswert kombiniert werden. Im Folgenden soll näher auf die verschiedenen Techniken singulärer und multipler Imputation eingegangen werden.

3.1 Singuläre Imputation

Wie bereits zuvor erwähnt, wird bei der singulären Imputation für jeden fehlenden Wert genau ein Wert geschätzt. Je nach Imputationsmethode werden hierzu unterschiedliche Schätzverfahren verwendet. Eine Methode ist beispielsweise, fehlende Werte durch den Stichprobenmittelwert — berechnet durch die beobachteten Ausprägungen der Variablen — zu ersetzen. Es wird also hier angenommen, dass die ausgefallenen Antworten in etwa den Wert annehmen, der im Mittel bei den vorliegenden Fällen beobachtet wurde. Auch wenn dies eine gängige Methode für einfache Analysen ist, so ist es doch offensichtlich, dass dies Methode Probleme bereitet, wenn eine nicht-symmetrische Verteilung vorliegt. Zudem wird durch das Einsetzen des gleichen Wertes in jedem fehlenden Fall die Varianz verkleinert. Dieser mangelnden Varianz beugt die Methode des personenspezifischen Mittelwertes vor. Hier werden die Antworten aus den anderen Variablen der gleichen statistischen Einheit zur Hilfe genommen, um einen Mittelwert zu erzeugen, der den fehlenden Wert ersetzt. So wird zwar nicht für jede statistische Einheit innerhalb einer Variablen der gleiche Wert eingesetzt, doch umgekehrt verringert sich die Varianz innerhalb einer statistischen Einheit.

Es ist leicht zu erkennen, dass die bisher beschriebenen Schätzverfahren mithilfe von Lageparametern eine einfache Methode darstellen, jedoch aufgrund ihrer mangelnden Präzision nur unter bestimmten Voraussetzungen angewendet werden können und sollten. Ist beispielsweise die Stichprobe sehr groß und der Anteil fehlender Werte entsprechend klein, so können Lageparameter hilfreich sein, um die fehlenden Werte zu ersetzen. Trifft dies nicht zu, so sollte man sich anspruchsvollerer Schätzverfahren bedienen.

Eine alternative Herangehensweise stellen die sogenannten *Cold-Deck-* und *Hot-Deck-Verfahren* dar. Bei diesen Methoden werden nicht Lageparameter, sondern tatsächliche Antwortausprägungen aus den vorliegenden Daten für die fehlenden Werte eingesetzt. Bei der Cold-Deck-Methode werden hierzu Daten aus alten Erhebungen herangezogen, bei den Hot-Deck-Techniken wird der aktuelle Datensatz verwendet. Üblich ist hierbei die Unterteilung der Fälle in Gruppen gemäß der Ausprägungen eines vollständig erhobenen Merkmals. Innerhalb dieser Gruppen wird jeweils für die fehlenden Werte einer Variablen der gleiche Wert eingesetzt, welcher in anderen Fällen tatsächlich beobachtet wurde und für diese Gruppe wahrscheinlich ist.

Um die Antwortausfälle möglichst präzise zu schätzen, sollte so viel zur Verfügung stehende Information wie möglich benutzt werden. Dies ist der Ansatz der Regressionsverfahren. Bei diesen Verfahren wird versucht, einen Zusammenhang zwischen der Variablen, in der der entsprechende Nonresponse auftaucht, und den übrigen beobachteten Variablen festzustellen. Angenommen der Datensatz bestünde nur aus drei

Variablen: X_1, X_2 und Y. X_1 und X_2 wurden vollständig bei n Personen beobachtet, bei Y liegt genau ein Antwortausfall vor. Um nun diesen fehlenden Wert zu ersetzen, wird mithilfe der $n-1$ vollständigen Beobachtungsreihen eine Regressionsfunktion der folgenden Gestalt aufgestellt:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Dabei sind $\hat{\beta}_0$, $\hat{\beta}_1$ und $\hat{\beta}_2$ die geschätzten Regressionsparameter. Mit dieser Gleichung, welche den Zusammenhang zwischen Y und den übrigen Variablen beschreibt, kann der fehlende Wert in Y leicht und plausibel mithilfe der Ausprägungen in X_1 und X_2 geschätzt werden.

Das Regressionsverfahren nutzt also die Korrelation der Variablen untereinander aus und errechnet so für jeden fehlenden Wert eigens eine Schätzung. Dabei wird jedoch kein Fehler mit einberechnet, was bedeutet, dass die geschätzten Werte von der Regressionsgleichung perfekt beschrieben werden und somit der Standardfehler unterschätzt wird. Außerdem kann es durch die rein mathematische Berechnung der einzusetzenden Werte dazu kommen, dass die Schätzer außerhalb des Wertebereiches der Variablen liegen. Dieses Problem löst das Verfahren des *Predictive Mean Matching*s. Die Vorgehensweise entspricht der des Regressionsverfahrens, außer dass der Regressionswert nicht direkt imputiert wird. Stattdessen wird ein Wert aus dem Wertebereich der Variablen gesucht, der dem Regressionswert am ehesten entspricht, und dieser dann eingesetzt.

Ein Nachteil der betrachteten singulären Imputationsmethoden liegt darin, dass keines der Verfahren berücksichtigt, dass es eine Struktur hinter den Antwortausfällen geben kann. Im zuvor erwähnten Beispiel einer Einkommensumfrage kann es vorkommen, dass Personen mit einem höheren Gehalt die Antwort zu ihrem Einkommen häufiger verweigern. Das bedeutet, dass die fehlenden Werte in dieser Variablen tatsächlich durchschnittlich höher sind als die beobachteten Werte. Wendet man in diesem Fall eines der gerade vorgestellten singulären Imputationsverfahren an, so werden die geschätzten Werte kleiner ausfallen als die realen Werte. Denn alle Schätzverfahren (sowohl Lageparameter, Cold- oder Hot-Deck-Techniken als auch Regressionsverfahren) beziehen sich auf die beobachteten Ausprägungen und setzen somit voraus, dass die Daten mindestens MAR sind.

Hinzu kommt das Problem, dass durch das Schätzen eines einzigen Wertes pro Zelle, der Unsicherheitsfaktor verloren geht. Die geschätzten Werte werden im neu generierten Datenset behandelt als seien sie tatsächlich beobachtet worden. Die durch die Schätzung entstandene Unsicherheit wird nicht berücksichtigt. Das hat zur Folge, dass der Standardfehler der Parameter im neu generierten Datensatz meist unterschätzt

wird (Schafer & Graham, 2002).

Um diese Unsicherheit des Schätzens zu berücksichtigen, kann auf multiple Imputationsverfahren zurückgegriffen werden.

3.2 Multiple Imputation

Der grundlegendste Unterschied zwischen der singulären und der multiplen Imputation ist die Anzahl der Werte, die für einen Antwortausfall geschätzt werden. Da bei den singulären Verfahren nur jeweils ein Schätzwert berechnet und in den Datensatz eingesetzt wird, wird der Unsicherheit nicht genügend Rechnung getragen. Um diesem Problem entgegenzuwirken, wird bei der multiplen Imputation dieser Vorgang des Schätzens und Einsetzens mehrfach wiederholt. So entstehen nicht einer, sondern gleich mehrere komplette Datensätze, mit welchen die betrachtete Analyse und Parameterberechnung durchgeführt wird.

Die Vorgehensweise der multiplen Imputation erfolgt in drei Schritten: Imputation, Analyse und Integration (Zahn, 2009).

Im ersten Schritt werden für jeden fehlenden Wert $m > 1$ Werte geschätzt und eingesetzt. Hierbei können wie bei der singulären Imputation unterschiedliche Schätzverfahren zur Hilfe genommen werden. Nach dem Imputationsschritt liegen somit m vollständige Datensätze vor. Mit diesen Datensätzen kann nun wie gewohnt die Analyse vollständiger Datensätze mit den üblichen Methoden und Standardsoftware durchgeführt werden. Dies muss allerdings m-mal getrennt voneinander geschehen, so dass schließlich m unterschiedliche Analyseergebnissets vorliegen, welche erst im nächsten Schritt wieder vereint werden. Die m Datensätze spiegeln die Unsicherheit wider, welche durch das Schätzen entstanden ist. Somit werden im letzten Schritt der Integration durch Kombinieren der Ergebnisse (Rubin, 1987) allgemeine Schätzer und Standardfehler erzeugt, welche die Unsicherheit der Daten reflektieren. (Schafer & Graham, 2002)

Zum Schätzen der zu imputierenden Werte wird im Rahmen der multiplen Imputation häufig auf den sogenannten *Expectation-Maximization-Algorithmus* (EM-Algorithmus) zurückgegriffen. Dieser stellt eine Verwendung von Maximum-Likelihood-Methoden bei unvollständiger Information vor. Der EM-Algorithmus basiert darauf, dass der Gesamtdatensatz einer bestimmten Verteilung folgt, deren Parameter es zu schätzen gilt. Da die Daten nicht vollständig vorliegen, können die Parameter nicht perfekt berechnet werden. Das Prinzip des EM-Algorithmus besteht darin, die Parameter aus den beobachteten Daten zu schätzen, um anhand dieser Parameter die fehlenden Werte zu ergänzen. Aus diesem neuen vervollständigten Datensatz werden die Parame-

ter neu geschätzt und aus den neuen Parametern wiederum neue Schätzungen für die fehlenden Werte errechnet. Bei iterativer Wiederholung dieser Schritte konvergieren die geschätzten Parameter gegen ihre wahren Werte.

Im Detail werden die folgenden beiden Schritte iterativ wiederholt:

E-Schritt Ziel der Berechnung ist es, den Parameter θ so zu schätzen, dass er für den kompletten Datensatz y_{com} (also beobachtete und fehlende Daten) am wahrscheinlichsten ist. Dazu wird zunächst eine Likelihoodfunktion $L(\theta, y_{com})$ aufgestellt, die ebendieses misst.

Der Parameter θ soll unter der Voraussetzung am wahrscheinlichsten sein, dass y_{obs} beobachtet wurde und der in der vorangegangenen Iteration geschätzte Parameter θ_i als bis dahin bester Näherungswert an das echte θ angenommen wird. Deshalb wird der Erwartungswert der Likelihoodfunktion unter den Bedingungen y_{obs} und θ_i berechnet. Dieser bedingte Erwartungswert soll gemäß einschlägiger Fachliteratur (Little & Rubin, 1987) mit Q bezeichnet werden:

$$Q(\theta) = Q(\theta|\theta_i) = E(L(\theta, y_{com})|y_{obs}, \theta_i)$$

M-Schritt Im Maximization-Schritt wird $Q(\theta)$ nach θ maximiert.

$$\theta_i = \arg\max_{\theta} Q(\theta|\theta_{i-1})$$

Das θ_i für welches $Q(\theta_i)$ am größten ist wird für die nächste Iteration verwendet.

Im ersten Durchlauf muss ein Startwert θ_0 „manuell“ geschätzt werden. Je nachdem welche Parameter für welche Art von Verteilung geschätzt werden sollen, ist hier ein plausibler Wert frei zu wählen. Der EM-Algorithmus konvergiert in jedem Fall - unabhängig davon, welcher Startwert gewählt wird. Allerdings kann es vorkommen, dass die Konvergenz in lokalen Maxima hängen bleibt. Es empfiehlt sich daher, unterschiedliche Startwerte auszuprobieren. Ein weiterer Nachteil besteht darin, dass die Berechnung keine Standardfehler berücksichtigt, d.h. es wird davon ausgegangen, dass die zu imputierenden Werte der unterliegenden Verteilung perfekt folgen, was nicht realistisch ist. Trotz dieser Nachteile führt der EM-Algorithmus meist zu sehr guten Ergebnissen. Deshalb ist diese Methode zur Schätzung von Verteilungsparametern bei unvollständiger Information weit verbreitet und auch in gängiger Software (z.B. NORM, SPSS, R) implementiert.

Im nächsten Schritt der Datenanreicherung (Data Augmentation) werden zunächst die fehlenden Werte anhand des im EM-Algorithmus erhaltenen Parmeters θ_{EM} geschätzt. Da dieser Parameter jedoch auf den beobachteten Werte basiert, stellt er nur

einen der möglichen Parameter der gesamten betrachteten Population dar. Um dieser Unsicherheit Rechnung zu tragen, wird dem Parameter eine plausible Verteilung unterstellt. Diese geschätzte Verteilung repräsentiert, wie die Verteilung des Parameters tatsächlich aussieht unter Einbeziehung des anhand der beobachteten Werte geschätzten Parameters θ_{EM}. Mithilfe dieser Verteilung wird ein neuer Parameter $\hat{\theta}_1$ geschätzt, welcher sich in einem aus der unterstellten Verteilung aufgespannten Konfidenzintervall um θ_{EM} befindet. Dieses neue $\hat{\theta}_1$ wird genutzt, um die fehlenden Werte im Datensatz zu bestimmen. Aus dem neuen vollständigen Datensatz wird wiederum ein neuer Parameter $\hat{\theta}_2$ geschätzt und der Prozess beginnt von vorn. Nach k Iterationen konvergiert der Prozess und es liegt ein neuer vollständiger Datensatz vor.

Da bei der multiplen Imputation jedoch nicht ein sondern m ergänzte Datensätze vorliegen sollen, deren Ergebnisse anschließend zu kombinieren sind, wird der gesamte Prozess m-mal wiederholt.

Die Anzahl der Iterationen m kann dabei vom Anwender selbst bestimmt werden. Auch wenn theoretisch nur eine unendliche Anzahl an Iterationen die wahren Werte am genauesten schätzt, so zeigt sich doch, dass relativ wenige Iterationen nötig sind, um gute Ergebnisse zu erzielen. Rubin (1987) stellte hierzu eine Formel auf, welche bei einer Informationsausfallrate von λ die Effizienz eines Schätzers mit m Iterationen im Vergleich zu $m = \infty$ berechnet:

$$(1 + \lambda/m)^{-\frac{1}{2}}$$

Liegt beispielsweise eine Informationsausfallrate von 40 % vor, so ist der Schätzer nach 10 Iterationen schon zu $(1 + 0,4/10)^{-1/2} = 96\,\%$ effizient (Rubin, 1987).

Im Anschluss an die Imputation liegen also m Datensätze mit m Parametersets vor. Wie diese Ergebnisse zu Aussagen über den Gesamtdatensatz kombiniert werden können, wird von Rubin (1987) näher erläutert. Um seinen Ansatz zu erklären, sei angenommen, θ sei ein Modellparameter des Gesamtdatensatzes, welchen es in m Iterationen zu schätzen galt. $\hat{\theta}_i$ sei die i-te Schätzung des Parameters θ. Nach Rubin kann der Parameter des Gesamtdatensatzes $\tilde{\theta}$ durch einfache Mittelwertberechnung geschätzt werden:

$$\tilde{\theta} = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m}$$

Bei der Kombination der Analyseergebnisse ist allerdings immer eine große Varianz zu erwarten, da sie sowohl die bei der Kombination entstandene Varianz als auch die Standardfehler, verursacht durch die nicht perfekte Schätzung von $\tilde{\theta}$ ($\tilde{\theta} = \hat{\theta}_i + e_i$), enthält.

3.3 Praktische Anwendung: Multiple Imputation

Die Datensätze ALLBUS 2002 und ALLBUS 2004 zeigten 343 gemeinsame Variablen auf. Eine Variable mit mehr als 50 % fehlender Werte wurde als „nicht aussagekräftig“ eingestuft, wonach 151 gemeinsame Variablen verblieben, welche als Einflussvariablen in Frage kamen. Bevor aus diesen gemeinsamen Variablen die Einflussvariablen bestimmt werden konnten, mussten diese zunächst auf Multikollinearität geprüft werden. So sollte ausgeschlossen werden, dass fälschlicherweise angenommen wird, dass mehrere Variablen einen Einfluss ausüben, obwohl sie die gleiche (oder ähnliche) Information beinhalteten. Daher wurde die Korrelationsmatrix dieser Variablen betrachtet, um eventuell vorliegende doppelte Information auszuschließen. Es war festzustellen, dass in den 151 Einflussvariablen viel doppelte Information enthalten war. So gab es neben der Variablen „Alter: Befragte<r>“ zusätzlich die Variablen „Geburtsjahr: Befragte<r>“ und „Alter: Befragte<r>, kategorisiert“, obwohl die Miteinbeziehung dieser Variablen keine zusätzliche Information lieferte. In der Korrelationsmatrix war zu erkennen, dass all diese Variablen eine perfekte Korrelation zu „Alter: Befragte<r>“ aufwiesen und somit problemlos als Einflussvariablen gestrichen werden konnten.

Im betrachteten Anwendungsbeispiel wurde in jedem Fall, in dem zwei Variablen eine signifikante Korrelation miteinander aufwiesen, einer der beiden Variablen als Einflussvariable eliminiert. Um zu entscheiden, welche der zwei (oder mehr) stark korrelierenden Variablen als Einflussvariable bestehen sollte und welche nicht, wurden vor allem ihre jeweiligen Skalenniveaus betrachtet. Beschreiben zwei Variablen inhaltlich das Gleiche, wurden sie aber auf verschiedenen Skalenniveaus gemessen, so ist stets die Variable mit dem höheren Skalenniveau zu wählen, da sie mehr Information enthält. So wurde beispielsweise die ordinale Variable „Alter: Befragte<r> (kategorisiert)“ aus dem Datensatz der Einflussvariablen entfernt, da die metrische Variable „Alter: Befragte<r>“ die gleiche Information enthielt und zudem von einem höheren Skalenniveau war. Sollte die gewählte Einflussvariable mehr fehlende Werte aufweisen als die Variable mit dem niedrigeren Skalenniveau, so können diese Werte mithilfe der Ausprägung der kategorisierten Variable (z.B. durch Berechnung der Klassenmitte) geschätzt werden. Zudem konnte bei der Entscheidung darüber, welche Einflussvariable eliminiert werden sollte, die jeweilige Korrelation mit denjenigen Variablen berücksichtigt werden, welche nachfolgend durch das Matching ergänzt werden sollten. Diese Variablen werden als *Fusionsvariablen* bezeichnet. Diejenige Variable, welche den größten Einfluss auf die Fusionsvariablen hat, sollte als Einflussvariable selektiert werden. Nach Ausschluss dieser doppelten Information verblieben 91 mögliche Einflussvariablen.

Als nächstes mussten die Fusionsvariablen bestimmt werden. Da es das Ziel einer

an diese Untersuchung anschließenden Analyse sein sollte, Aussagen zum Zusammenhang zwischen Religion und Gesundheit zu treffen, wurde im späteren Fusionsverfahren sowohl ein Matching der Gesundheitsvariablen auf den 2002er Datensatz als auch ein Matching der Religionsvariablen auf den Datensatz aus 2004 durchgeführt. Somit wurden diejenigen Variablen, die sich mit diesen beiden Themen befassten, als Fusionsvariablen deklariert. Um auch hier Multikollinearität auszuschließen, mussten auch diese Variablen zunächst auf Korrelation geprüft werden. So konnten wiederum einige Variablen ausgeschlossen werden. Nach diesem Prozess lagen 55 Fusionsvariablen zum Thema „Gesundheit“ und 84 Variablen zum Thema „Religion“ vor.

Um zu prüfen, welche dieser gemeinsamen Variablen den größten Einfluss auf die Fusionsvariablen hatte, wurde anschließend die Korrelationen der gemeinsamen mit den themenspezifischen Variablen betrachtet. So konnten die gemeinsamen Variablen in eine Rangfolge anhand ihres Einflusses auf die Fusionsvariablen gebracht und die 28 einflussreichsten gemeinsamen Variablen selektiert werden. Aus zwei Gründen wurde genau diese Menge an Einflussvariablen gewählt. Zunächst war zu überlegen, wie viele Einflussvariablen Sinn machen, um 55 bzw. 84 andere Variablen zu prognostizieren. Es sollten nicht zu wenige sein, damit die Vorhersagen möglichst genau sind. Zu viele Einflussvariablen dagegen würden das Modell unhandlich machen. Zusätzlich wurden die Variablen inhaltlich nach ihrem Einfluss auf die Fusionsvariablen betrachtet. Nach der 28. Variablen folgten Variablen wie „Wichtigkeit der Inflationsbekämpfung“ oder „Typ der Wohnung, in der Befragter wohnt“, welche nach individueller Einschätzung weniger bedeutend für das Modell waren.

Nach diesem Eliminierungsprozess verblieb ein Datensatz G[6] mit 28 Einflussvariablen, von denen 6 metrisch-, 8 ordinal- und 15 nominalskaliert waren.

Aufgrund der Auswahlkriterien für die Einflussvariablen enthielten diese höchstens 50 % fehlende Werte. Doch offensichtlich ist auch eine Variable mit Antwortausfällen von beispielsweise 40 % schlecht dazu geeignet, Prognosen über andere Variablen zu treffen. Deshalb sollten anschließend die fehlenden Werte in G imputiert werden.

Im ersten Schritt war dazu zu prüfen, ob der Antwortausfall im Datensatz G strukturiert war. Dazu wurde die Statistik-Software SPSS benutzt, welche zur Prüfung den Test nach Little (1988) verwendet. Nach dessen Durchführung wurde der betrachtete Datensatz G nicht für MCAR befunden. Dies rechtfertigte eine multiple Imputation.

Um die multiple Imputation im Datensatz G durchzuführen, wurde die Software *NORM* genutzt, welche von Schafer (1997) speziell für diese Anwendung entwickelt wurde. Zuvor wurde sowohl SPSS als auch R als Imputationswerkzeug ausprobiert. Beide Softwares kamen zu keinem Ergebnis, da sie mit der Größe der Datensätze über-

[6] CD: Kapitel 3 – Multiple Imputation/20022004_28gVariablen

fordert schienen. In NORM dagegen verlief der Imputationsprozess einwandfrei.

In NORM werden der EM-Algorithmus und der Schritt der Datenergänzung getrennt voneinander durchgeführt. In beiden Schritten gibt es zusätzliche Einstellungsoptionen, welche vom Anwender spezifisch angepasst werden können.

Bevor der Datensatz jedoch in NORM eingelesen wurde, musste zunächst überlegt werden, welche Transformationen an den Variablen notwendig waren. Die Werte von nominalskalierten Variablen sind willkürlich, d.h. sie unterliegen keiner Intervallskala oder Rangfolge und wurden somit frei gewählt. Dies stellt ein Problem dar, da die Ausprägungen nicht numerisch vergleichbar sind. Daher mussten zunächst alle nominalen Variablen in Dummies transformiert werden. Dafür wurden für jede nominalskalierte Variable mit k Kategorien $k-1$ Dummy-Variablen erzeugt, welche aussagten, ob die jeweilige Kategorie zustraf (Dummy = 1) oder nicht zutraf (Dummy = 0). Bei einer solchen Transformation werden nur maximal $k-1$ Dummies benötigt, da die letzte Kategorie (*Referenzkategorie*) durch die Ausprägungen in den anderen Dummies erklärt wird. Im Praxisbeispiel wurde jeweils diejenige Ausprägung als Referenzkategorie ausgewählt, welche die größte Häufigkeit aufwies. Im Anschluss an die Zerlegung wurde überprüft, ob es Kategorien in den nominalskalierten Variablen gab, welche von niemandem geantwortet wurden. Die Dummies dieser Kategorien müssen zusätzlich aus dem Datensatz entfernt werden, da sie nur eine Ausprägung (=0) vorweisen und somit keinerlei Information lieferten. Im Datensatz G traf dies nur für die Kategorie „Noch Schüler“ der Variablen „Gegenw. Ehepartner: Allgemeiner Schulabschluss“ zu.

Nun könnte man argumentieren, dass auch die ordinalen Variablen nicht untransformiert in die Imputation eingehen dürfen, da sie zwar eine Rangfolge haben, jedoch die Abstände zwischen den Kategorien nicht einheitlich sind. Eine Standardisierung der ordinalskalierten Variablen würde somit zu genaueren Ergebnissen führen. Doch da die Ergebnisunterschiede minimal und somit zu vernachlässigen sind, wurde an dieser Stelle auf eine Transformation verzichtet (Rohrschneider, 2007). Bei den metrischen Variablen ist offensichtlich keine Transformation notwendig, da die Rangfolge und Abstände zwischen Ausprägungen eindeutig definiert sind. Somit bedurften nur die nominalskalierten Variablen einer Transformation und es lagen nach ihrer Zerlegung in Dummies 96 Variablen (6 metrische, 8 ordinale und 82 Dummy-Variablen)[7] vor. Diese wurden anschließend in NORM eingelesen.

In den Variablen-Einstellungen von NORM wurde zunächst festgelegt, dass die geschätzten Werte für fehlende Daten jeweils auf den nächstgelegenen beobachteten Wert gerundet werden sollten. So sollte sichergestellt werden, dass alle imputierten Werte im Wertebereich der Variablen liegen würden und keine unsinnigen Aussagen (wie z.B.

[7] CD: Kapitel 3 – Multiple Imputation/20022004_28gVariablen_Dummies

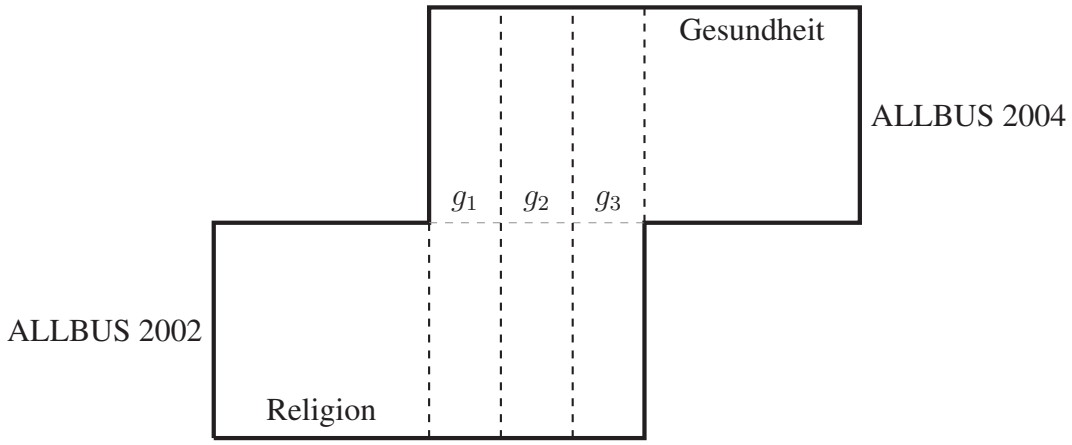


Abbildung 3.1: Schematische Darstellung der imputierten Datensätze der gemeinsamen Variablen

ein negativer Wert für die Variable „Befr.: Nettoeinkommen“) getroffen würden. Allerdings sollte hier bedacht werden, dass diese Vorgehensweise nicht nur Verzerrungen zur Folge haben könnte, sondern auch die Varianz im ergänzten Datensatz verringert.

Beim EM-Algorithmus in NORM können neben der Angabe der maximalen Anzahl an Iterationen auch die Startwerte bestimmt werden. Im betrachteten Beispieldatensatz wurden die aus den beobachteten Daten errechneten Parameter als Startwerte benutzt. Nach Durchführung des EM-Algorithmus wird in einer Textdatei[8] eine Zusammenfassung der Ergebnisse gespeichert. Hierin ist abzulesen, nach der wievielten Iteration der Algorithmus konvergiert ist und wie die Parameterschätzungen nach dieser Iteration aussehen. Zusätzlich kann vom Anwender entweder die Kovarianz- oder die Korrelationsmatrix der Variablen nach Ergänzen der fehlenden Werte ausgegeben werden.

Im nächsten Schritt der Data Augmentation wurden die durch den EM-Algorithmus erhaltenen Parameter als Startwerte bestimmt. Zusätzlich kann vom Anwender die Anzahl der Iterationen und der Imputationen bestimmt werden. Die Ergebnisse der Datenanreicherung werden wiederum in einer Textdatei[9] gespeichert. Für den Datensatz G wurden 600 Iterationen festgelegt. Nach jeweils 200 Iterationen sollte eine Imputation erfolgen, so dass nach dem Durchlauf 3 vollständige Datensätze[10] vorlagen (vgl. Abb. 3.1).

Alle nachfolgenden Analyseschritte wurden nun jeweils mit allen drei Datensätzen durchgeführt, um im Anschluss die drei Ergebnissets miteinander zu kombinieren. Dies ist der Grund dafür, dass immer eine ungerade Anzahl an Imputationen durch-

[8] CD: Kapitel 3 – Multiple Imputation/Imputation G/em
[9] CD: Kapitel 3 – Multiple Imputation/Imputation G/da
[10] CD: Kapitel 3 – Multiple Imputation/Imputation G/gImputation*

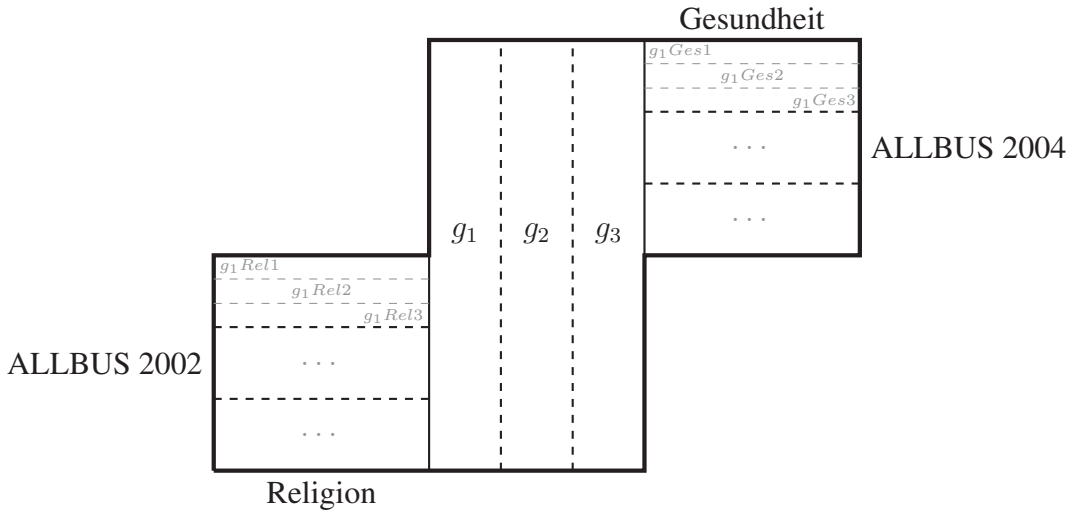


Abbildung 3.2: Schematische Darstellung der imputierten Datensätze der spezifischen Variablen

geführt werden sollte. So wird sichergestellt, dass die Kombination der Ergebnisse zu einer eindeutigen Aussage führt und es keine „Gleichstandssituation“ geben kann.

Da im anschließenden Matching-Schritt sowohl die Gesundheits- als auch die Religionsvariablen integriert werden sollten, mussten auch die zuvor definierten Fusionsvariablen von fehlenden Werten befreit werden. Deshalb wurde als nächstes eine Multiple Imputation dieser Variablen durchgeführt. Diese multiple Imputation erfolgte analog zur Imputation der gemeinsamen Variablen. Die Transformation der nominalskalierten Variablen in Dummies erhöhte die Anzahl der Fusionsvariablen zum Thema „Religion“ auf 94. Unter den Gesundheitsvariablen lagen keine nominalskalierten Variablen vor. Somit gingen weiterhin 55 Variablen in das Imputationsmodell ein.

Die bereits imputierten Datensätze der gemeinsamen Variablen sollten für die Schätzung der fehlenden Werte genutzt werden. Im Falle der Gesundheitsvariablen bedeutete dies, dass aus den imputierten G_i-Datensätzen jeweils die 2946 Fälle der Befragten aus dem 2004er Datensatz extrahiert und diese zusammen mit den 55 Gesundheitsvariablen in das Imputationsmodell gegeben wurden. Für jeden der drei imputierten Datensätze der gemeinsamen Variablen wurden weitere drei imputierte Gesundheits-Datensätze[11] konstruiert. Somit lagen nun neun Datensätze der 96 gemeinsamen und der 55 Gesundheitsvariablen vor.

Analog wurden die fehlenden Werte in den Religionsvariablen imputiert und so weitere neun Datensätze[12] erzeugt, welche die gemeinsamen und die Religionsvariablen vollständig an 2820 Fällen erfassten. Diese 18 Datensätze waren nun soweit vor-

[11] CD: Kapitel 3 – Multiple Imputation/Imputation Gesundheit/gImputation*_Ges*
[12] CD: Kapitel 3 – Multiple Imputation/Imputation Religion/gImputation*_Rel*

bereitet, dass mit ihnen das Statistical Matching durchzuführen war (vgl. Abb. 3.2).

Kapitel 4

Statistical Matching – Fusion der Datensätze

> To the uninitiated data fusion might sound like major market research fraud. How do you react to the proposition that two different sample surveys covering different subject areas with different respondents can be statistically joined together? The respondents in the resulting data set will have all the answers to all the questions in both the original surveys. Your first reaction will probably be cynical disbelief.
>
> Baker, Harris & O'Brien (1989)

4.1 Das Prinzip der Datenfusion

Moderne Datenerhebungs- und -verarbeitungstechniken ermöglichen die Verfügbarkeit von immer mehr Daten. Zu nahezu jedem Themenbereich wurde bereits eine Umfrage gemacht und es liegen entsprechend große Datenmengen vor. Doch meist ist die Fragestellung vor der statistischen Analyse eine sehr komplexe, die mehrere Themenbereiche umfasst und häufig zusätzlich große Populationen betrifft. Möchte beispielsweise ein Telekommunikationsunternehmen Charakteristiken und Bedürfnisse ihrer Kunden analysieren, so wäre eine umfassende Umfrage am gesamten Kundenstamm notwendig. Da dies mit hohem Zeit- und Kostenaufwand verbunden wäre, wird auf eine eigene Datenerhebung meist verzichtet. Stattdessen könnte in diesem Fall auf eine bereits durchgeführte Umfrage zurückgegriffen werden, welche die interessierten Variablen bereits an einer anderen großen Population abgefragt hat. Da diese Daten sich jedoch auf eine andere Grundgesamtheit beziehen, müssen diese mit den vorliegenden Kundendaten fusioniert werden. Dabei werden die nicht erhobenen und somit fehlenden Daten im Kundenstamm mithilfe des Umfragedatensatz anhand

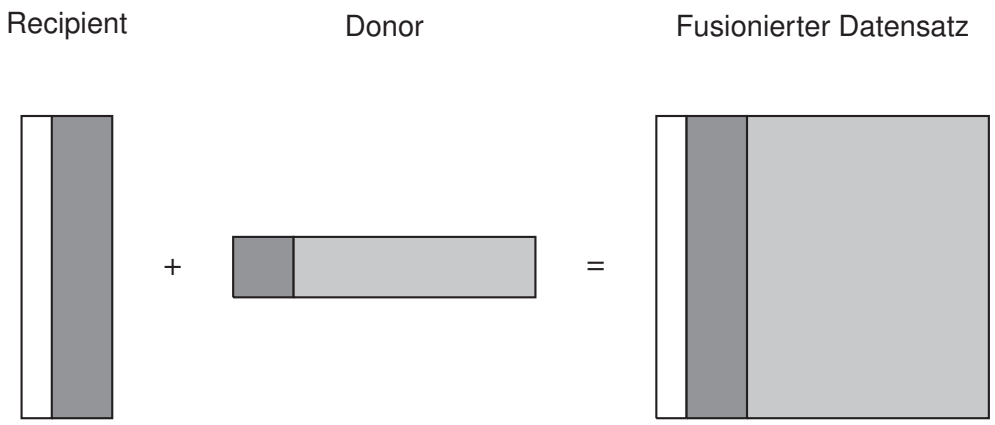


Abbildung 4.1: Graphische Darstellung der Datenfusion nach Van Der Puttan, Kok & Gupta (2002)

gemeinsamer Ausprägungen (beispielsweise des Geschlechts) geschätzt. Eine denkbare Vorgehensweise wäre hier, für jeden weiblichen Kunden die Charakteristiken mit einem aus den Antworten aller befragten Frauen errechneten Mittelwert zu ersetzen. Dies klingt auf den ersten Blick sehr willkürlich, da die Annahme, dass die Ausprägungen aller Frauen, die nicht geantwortet haben, durch aus den vorliegenden Werten aller Frauen gemittelte Ausprägungen geschätzt werden können, mehr als gewagt ist. Man könnte einwenden, dass eine solche Datenimputation so unrealistisch ist, dass auf diese Methode gleich verzichtet werden kann, da sie die Daten eher verzerrt als ergänzt. Allerdings wurde hier nur eine Variable (Geschlecht) als Einflussvariable gewählt. Nimmt man mehrere Variablen (z.B. Alter, Wohnort, Familienstand) mit auf, so werden die Gruppierungen immer spezifischer und Vorhersagen über die fehlenden Ausprägungen werden realistischer.

Selbstverständlich ist die Datenfusion in ihrer realen Umsetzung nicht so rudimentär wie in diesem Einführungsbeispiel beschrieben. In diesem Kapitel soll zunächst das Prinzip der Datenfusion erläutert werden. Im Anschluss wird der spezielle Fall von Statistical Matching betrachtet. Obwohl die Begriffe „Datenfusion“ und „Statistical Matching“ im deutschen Sprachgebrauch oft synonym verwendet werden, sei jedoch an dieser Stelle darauf hingewiesen, dass es durchaus andere Arten der Datenfusion gibt. Auf diese soll allerdings aufgrund der eingeschränkten Rahmenbedingungen dieser Arbeit nicht weiter eingegangen werden.

Nachdem die Theorie des Statistical Matchings beschrieben wurde, soll am Ende dieses Kapitels am „Religion-Gesundheits-Beispiel“ eine praktische Durchführung des Matchings beschrieben werden.

Bei jeder Fusion zweier Datensätze fungiert einer der Datensätze als *Recipient* (Empfänger) und der andere als *Donor* (Spender). Dabei ist der Recipient derjenige

Datensatz, welcher um Variablen ergänzt werden soll. Die nötige Information hierzu stammt aus dem Donor-Datensatz. Die beiden Datensätze bestehen jeweils aus gemeinsamen Variablen, die in beiden Datensätzen vorliegen (X), und einzigartigen Variablen, die nur in einem der beiden Datensätze vorzufinden sind. Diejenigen Variablen, die nur im Recipient-Set vorkommen, sollen mit Y beschrieben werden; Variablen, die im Spender- nicht aber im Empfänger-Datensatz vorhanden sind, seien mit Z bezeichnet. Bei der Datenfusion wird mithilfe des Donors ein Modell aufgestellt, welches beim Input von X den Output von Z beschreibt. Dieses wird dann am Recipient angewendet, um diesen um Z zu ergänzen. Dieser Prozess wird in Abb. 4.1 grafisch dargestellt.

4.2 Statistical Matching

4.2.1 Theorie

Um zwei Datensätze zu fusionieren, wird meist auf den Algorithmus des *Statistical Matching*s zurückgegriffen. Das Prinzip dieser Methode besteht darin, zu jeder statistischen Einheit des Empfänger-Datensatzes einen sogenannten statistischen Zwilling im Spender-Set zu finden, welcher ihr in ausgewählten gemeinsamen Variablen möglichst ähnlich ist. Hieraus folgt die Bedingung, dass sich die statistischen Einheiten in den zu matchenden Datensätzen auf das Gleiche beziehen. Ist für den einen Datensatz beispielsweise eine statistische Einheit ein Haushalt, und der andere Datensatz befragte einzelne Individuen, so werden im Matching falsche Werte prognostiziert. In diesem Fall müsste man die Datensätze so angleichen, dass sie sich auf dem gleichen Messlevel befinden. Zudem ist sicherzustellen, dass die zu fusionierenden Datensätze vergleichbare Populationen abbilden. Handelt es sich etwa beim Donor um eine bundesweite Umfrage, während der Recipient nur Daten aus Berlin erfasst, so müssen die Datensätze zunächst angepasst werden. In diesem Beispiel könnte man den Donor so anpassen, dass er nur noch die in Berlin erfassten Daten enthält.

Im nächsten Schritt müssen diejenigen Variablen identifiziert werden, die für das Matching verwendet werden sollen. Welche Variablen hierfür selektiert werden sollten, hängt von den Variablen selbst und ihrem möglichen Einfluss auf die zu ergänzenden Fusionsvariablen Z ab. Hierbei kommen allerdings nur gemeinsame Variablen in Frage, die zusätzlich auch in der gleichen Einheit gemessen wurden, damit die Vergleichbarkeit gewährleistet ist. Also sollten diejenigen gemeinsamen Variablen selektiert werden, welche mit den Fusionsvariablen am stärksten korrelieren. Handelt es sich bei den betrachteten statistischen Einheiten um Personen, so werden meist Variablen wie „Geschlecht“, „Alter“, „Familienstand“ und Ähnliche als Matchingvariablen

genutzt.

Um die „wahre“ Verteilung im konstruierten Datensatz beizubehalten, sollte möglichst bedingte Unabhängigkeit von Y und Z zu X bestehen. Dies wurde von Sims (1972 a und b) erstmals erkannt. Ein ausführlicher Beweis ist in Rässler (2002) zu finden.

Es liegt nun also ein Recipient-Datensatz der Größe n_R vor, welcher sich aus den Variablen X und Y zusammensetzt. Der Datensatz des Donors enthält n_D Einheiten und besteht aus den Variablen X und Z. Dies bedeutet, dass für jedes i ($i = 1, \ldots, n_R$) die Observationen (x_i, y_i) und für jedes j ($j = 1, \ldots, n_D$) die Observationen (x_j, z_j) vorliegen.[13] Ziel des Matchings ist es, jede Beobachtung (x_i, y_i) anhand des Donors um einen Wert z_i zu ergänzen. Dafür wird eine Einheit j aus der Kontrollgruppe (= die Menge der statistischen Einheiten aus dem Donor) mit $(x_j, z_j) \in \{(x_1, z_1), \ldots, (x_{n_D}, z_{n_D})\}$ gesucht, welche der Einheit i aus der Behandlungsgruppe (= die Menge der statistischen Einheiten des Recipients) in den gemeinsamen Variablen X am ähnlichsten ist. Um die Ähnlichkeit zwischen zwei Einheiten zu messen, kann auf bekannte Distanzmaße (z.B. euklidische oder City-Block-Distanz) zurückgegriffen werden oder eine Ähnlichkeit anhand der *Propensity Scores* bestimmt werden. Da die Distanzmaße bereits aus anderen statistischen Zusammenhängen bekannt sein sollten, soll zunächst die Ähnlichkeitsmessung anhand eines Distanzmaßes näher betrachtet werden.

Unabhängig von der Wahl des Distanzmaßes gilt: Je geringer die Distanz $d(x_i, x_j)$, desto ähnlicher sind sich die Einheiten i und j. Bei Formulierung des Distanzmaßes $d(x_i, x_j)$ kann miteinbezogen werden, wie stark der Einfluss der einzelnen Variablen auf die Ähnlichkeit der Einheiten sein soll. Somit können die Variablen verschieden stark gewichtet werden. So könnte beispielsweise bei zu ergänzenden Daten zum Thema „Ehe“ der Einflussvariable „Familienstand“ ein höherer Einfluss zugewiesen werden als der Variable „Wohnort“.

In der Praxis wird häufig die euklidische Distanz verwendet, deshalb soll ihre Anwendung kurz genauer betrachtet werden. Um die unterschiedlichen Skalenniveaus der Variablen zu berücksichtigen, werden zunächst alle metrischen und ordinalen Variablen standardisiert. Die Dummy-Variablen bleiben unverändert und für die nominalskalierten Variablen werden für jede Kategorie Dummy-Variablen erzeugt. So ist die Vergleichbarkeit zwischen Variablen verschiedener Skalenniveaus gewährleistet. Sei $g(X)$ das der Variablen X zugesprochene Gewicht anhand ihres Einflusses auf die Ähnlichkeit. Dann ist die Distanz zwischen zwei Einheiten i und j auf Basis von p

[13] Dies berücksichtigt keine fehlenden Werte. Wir gehen in diesem Fall von vollständigen Datensätzen aus.

gemeinsamen Variablen gegeben durch:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} g(x_k)(x_k i - x_k j)^2} \tag{4.1}$$

Anhand Gleichung 4.1 können für alle Paare (i, j) die Distanzen berechnet und miteinander verglichen werden. Die Einheit $j*$, für welche die Distanzfunktion $d(x_i, x_j)$ minimal ist, ist statistischer Zwilling von i.

Im Unterschied zur Ähnlichkeitsmessung per Distanzmaß wird bei der Berechnung mithilfe von Propensity Scores nicht die Distanz zwischen den jeweiligen Ausprägungen der Einflussvariablen X bestimmt. Anstatt dessen wird eine neue Variable erzeugt (der Prognosewert oder Propensity Score), in welcher die statistischen Zwillinge möglichst wenig voneinander abweichen. Das Propensity Score Matching findet meist Anwendung, wenn die Behandlungsgruppe und die Kontrollgruppe stark voneinander abweichen, d.h. wenn in der Kontrollgruppe wenig vergleichbare Einheiten zur Behandlungsgruppe vorliegen. Dies lässt vermuten, dass die Gruppenzugehörigkeit einen Einfluss auf die beobachteten Ausprägungen hat. Deshalb wird zunächst in einem zusammengefügten Datensatz (Recipient und Donor) eine neue Variable S erzeugt, welche angibt, ob der Fall aus dem Recipient- oder dem Donor-Datensatz stammt. Sei $S = 0$ für alle Einheiten $j = 1, \ldots, n_D$ der Kontrollgruppe und $S = 1$ für alle Einheiten $i = 1, \ldots, n_R$ der Behandlungsgruppe. Diese Variable S soll nun als abhängige Variable in ein logistisches[14] Regressionsmodell eingehen, in welchem die ausgewählten gemeinsamen Variablen X als unabhängige Variablen wirken. Es wird also eine Regressionsfunktion formuliert, die bei M Einflussvariablen von folgender Gestalt ist:

$$P(S = 1|X = X_k) = \frac{e^z}{1 + e^z} \tag{4.2}$$

$$\text{mit } z = \beta_0 + \sum_{m=1}^{M} \beta_m X_m$$

Bei der Schätzung der Parameter $\beta_0, \ldots, \beta_k$ wurden alle Fälle aus dem Recipient und dem Donor miteinbezogen.

Gleichung 4.2 errechnet für jeden Fall k aus der Kontroll- und der Untersuchungsgruppe die Wahrscheinlichkeit dafür, dass der Fall aus der Behandlungsgruppe stammt, gegeben der beobachteten Ausprägungen X_k. Auf Basis dieser Propensity Scores wird nun die Ähnlichkeit zwischen zwei Einheiten i und j bestimmt. Diejenige Einheit j^* ist statistischer Zwilling zu i, wenn die Distanz zwischen ihren Propensity Scores mi-

[14] Hier kann auch ein linearen Regressionsmodell oder eine Diskriminanzanalyse verwendet werden, wenn davon auszugehen ist, dass diese zu besseren Prognosewerten führen.

nimal ist.

$$d(P_i, P_j) = |P_i - P_j|$$
$$j^* = \arg\min_j d(P_i, P_j)$$

Unabhängig davon, ob die beobachteten Ausprägungen oder die Propensity Scores auf Distanz geprüft werden, können beim Auffinden statistischer Zwillinge sogenannte kritische Variablen bestimmt werden, welche eine 100 %-ige Übereinstimmung in dieser Variable fordern. Das bedeutet, dass zwei Einheiten nicht als statistische Zwillinge erklärt werden können, wenn sie in dieser Variable nicht übereinstimmen. Diese Zuweisung von kritischen Variablen macht oft Sinn, um unsinnige Folgerungen zu vermeiden. Es kann beispielsweise die Variable „Geschlecht“ als kritisch ausgezeichnet werden, um unmögliche Ausprägungen in Variablen wie „letzte Menstruation“ bei männlichen Befragten zu verhindern.

Darüber hinaus sollte man sich vor der Durchführung des Matchings Gedanken über das zu verwendende Suchverfahren machen. Bei Auswahl eines geeigneten Suchverfahrens kann festgelegt werden, in welcher Reihenfolge der Datensatz durchsucht werden soll, ob eine Einheit aus dem Donor mehrfach als statistischer Zwilling erklärt werden kann oder ob ein Zwillingspaar wieder getrennt werden soll, wenn im Nachhinein ein besserer Zwilling für den Donor gefunden werden kann. Häufig macht es Sinn, die Anzahl, wie häufig eine Einheit aus dem Donor-Set als statistischer Zwilling gewählt werden darf, zu beschränken. Da es vorkommen kann, dass eine Einheit öfter als die Ähnlichste erklärt wird als andere, kann das mehrfache Nutzen dieser Einheit als Zwilling den ergänzten Recipient-Datensatz verzerren und seine Varianz verringern. Zudem ist bei mehrfacher Nutzung des gleichen Donors die Annahme der Unabhängigkeit zwischen den einzelnen betrachteten Einheiten nicht mehr gegeben. Um diesem Problem entgegenzuwirken, kann man die Distanzmessung so formulieren, dass das vorherige Nutzen einer Einheit als statistischer Zwilling als positiver Summand in die Distanzfunktion miteingeht. Ein solches Matching mit beschränkter Anzahl der Nutzung als statistischer Zwilling nennt man *bedingtes* Matching. Ein Nachteil der Streichung bereits genutzter Zwillinge aus dem Spender ist, dass im Vergleich eventuell schlechtere Zwillinge gefunden werden als wenn immer alle Fälle zur Verfügung stünden. Beispielhaft soll hier das Suchverfahren der *random order, nearest available pair-matching method*(Smith, 1997) beschrieben werden. Bei dieser Methode werden die Fälle sowohl im Recipient- als auch im Donordatensatz zufällig angeordnet. Dann wird für den ersten Fall des Recipients r_1 ein Fall ${d_j}^*$ aus dem Donor-Set gesucht, welcher einen statistischen Zwilling zu r_1 darstellt. Kommen mehrere d_j als Zwilling

in Frage, so wird der zuerst gefundene Fall – also der Fall mit dem kleineren Index – als Zwilling deklariert. Für die weitere Zwillingssuche wird der Fall d_j^* aus dem Donordatensatz gestrichen. Dieser Vorgang wird für alle r_i, $i = 1, \ldots, n_R$ wiederholt. Um der Unsicherheit bei der Findung statistischer Zwillinge Rechnung zu tragen, kann der Prozess mehrmals mit anderen zufälligen Anordnungen wiederholt werden. Diese Methode wird als *multiples Matching* bezeichnet.

Da ein gefundenes Zwillingspaar (i, j^*) sich nicht nur im Vergleich zu allen anderen (i, j) am ähnlichsten sein sollte, sondern auch objektiv betrachtet eine Ähnlichkeit zwischen i und j^* bestehen sollte, wird häufig zusätzlich vorausgesetzt, dass die Abweichung von x_i zu x_j unterhalb eines bestimmten Schwellenwerts c liegt. Wie dieser Schwellenwert zu wählen ist, ist nicht allgemein zu definieren. Bei einem zu großen Schwellenwert kann es sein, dass ein Paar (i, j^*) zum Zwillingspaar erklärt wird, obwohl sie wenig Ähnlichkeiten aufweisen. Auf der anderen Seite ist es denkbar, dass bei einem zu niedrigen Schwellenwert für viele Einheiten kein statistischer Zwilling gefunden werden kann. Daher empfiehlt es sich, den Prozess zunächst mit einem sehr großen Schwellenwert durchzuführen und ihn nach und nach zu verkleinern, bis das Matching zu zufriedenstellenden Ergebnissen führt. Diese Variation des Schwellenwertes kann auch umgekehrt durchgeführt werden, beginnend mit einem sehr kleinen Wert für c.

Nachdem alle Einheiten des Donor-Datensatzes auf Ähnlichkeit zu der betrachteten Einheit des Recipient-Sets geprüft wurden, wird nicht ein, sondern die k ähnlichsten statistischen Zwillinge bestimmt. Aus diesem Set von k Einheiten werden die Ausprägungen der zu ergänzenden Fusionsvariablen im Recipient-Set z.B. durch die zufällige Auswahl einer der k Ausprägungen der Zwillinge geschätzt. Auch die Schätzung durch den Mittelwert der k Beobachtungen ist denkbar, allerdings hat dies offensichtlich einen Varianzverlust im neu erzeugten Datensatz zur Folge. Durch die Verwendung multipler Zwillinge kann die Effizienz erhöht werden, da berücksichtigt wird, dass die imputierten Werte nicht tatsächlich beobachtet, sondern anhand des Donors geschätzt wurden. Allerdings kann es vorkommen, dass unter den k Zwillingen Fälle gefunden werden, die der Recipient-Einheit nur noch wenig ähnlich sind.

4.2.2 Evaluation

Durch den Matchingprozess wurde zwar ein vollständiger Datensatz erzeugt, welcher alle interessierenden Variablen enthält, doch liegt keine Information über die Qualität der Imputation vor. Die Evaluation der Ergebnisse ist daher unabdingbar und als Teil des Matchingprozesses selbst zu betrachten. Hierbei unterscheidet man zwischen der *internen* und der *externen* Evaluation. Die interne Evaluation beurteilt die Qualität

des gematchten Datensatzes nur anhand der vorliegenden Daten ohne Miteinbeziehung der nachfolgenden Analyse. Das Matching wird also nur bis zum Schritt der Datenvervollständigung geprüft. Bei der externen Evaluation hingegen werden nachfolgende Analyseschritte in die Beurteilung miteinbezogen. Hier wird also geprüft, ob das Matching im Hinblick auf die Analyse und deren Ergebnisse zufriedenstellende Ergebnisse liefert.

Interne Evaluation

Ziel der internen Evaluation ist es, zu prüfen, ob die im Recipient imputierten Werte der Fusionsvariablen die Informationen aus dem Donor gut widerspiegeln. Das bedeutet, dass die Fusionsvariablen im Recipient einer ähnlichen Verteilung folgen sollten wie ebendiese im Donor. Dies ist auf verschiedene Weisen zu prüfen.

Um zunächst zu prüfen, ob der Donor ein repräsentativer Datensatz für den Recipient darstellt, wird ein Mittelwertvergleich der gemeinsamen Variablen durchgeführt. Dabei wird für jede gemeinsame Variable jeweils der Mittelwert aus den Recipient- und den Donordaten errechnet, um sie dann miteinander zu vergleichen. Sollten signifikante Unterschiede festgestellt werden, so ist eine Ausprägungsgruppe (z.B. „verheiratet“) in einem der Datensätze überrepräsentiert. Ob dies Auswirkungen auf die Analyse hat, hängt von der Untersuchung selbst ab.

Auch die Ausprägungen der Fusionsvariablen sollten sich möglichst im Recipient und im Donor um einen ähnlichen Mittelwert streuen. Um dies zu prüfen, wird der zuvor beschriebene Prozess mit den Fusionsvariablen wiederholt. Bei der Analyse der Mittelwertsabweichungen sollte miteinbezogen werden, wie repräsentativ der Donor ist. Wenn beim Mittelwertvergleich der gemeinsamen Variablen festgestellt wurde, dass eine bestimmte Gruppe überrepräsentiert ist, kann es sein, dass dies die Erklärung für eine Mittelwertabweichung in den Fusionsvariablen ist. Die betrachtete Variable könnte mit der überrepräsentierten Variable stark korreliert sein und hätte damit vermutlich auch überrepräsentierte Ausprägungen.

Darüber hinaus soll sichergestellt werden, dass die im Donor beobachteten Relationen von Fusionsvariablen zu den gemeinsamen Variablen und zu sich untereinander im vervollständigten Datensatz beibehalten wurden. Dazu wird zunächst sowohl im Donor als auch im gematchten Datensatz die Korrelation jeder gemeinsamen Variable zu den Fusionsvariablen berechnet. Wie beim Mittelwertvergleich werden die erhaltenen Daten auf signifikante Unterschiede geprüft. Dieser Prozess wird analog für die Korrelation zwischen den Fusionsvariablen durchgeführt, um festzustellen, ob die Relationen innerhalb der ergänzten Variablen in etwa beibehalten wurden.

Externe Evaluation

Da der Fokus des praktischen Teils dieser Arbeit auf der Datenbearbeitung liegt und keine vollständige Analyse durchgeführt wurde, konnte die externe Evaluation im betrachteten Praxisbeispiel nicht angewendet werden. Daher soll auch in der Theorie auf eine ausführliche Beschreibung verzichtet werden. Dennoch sei die Vorgehensweise an dieser Stelle kurz erläutert.

Wie zuvor erwähnt, prüft die externe Evaluation den Wert des Matchings für die weitere Analyse. Um diesen festzustellen, muss geprüft werden, ob die Analyse nach dem Matching bessere Ergebnisse liefert als wenn auf den Matchingprozess verzichtet und nur die vorliegenden gemeinsamen Variablen benutzt worden wären. Die naheliegendste Methode, die Qualität des Matchings zu festzustellen, ist wohl die Prüfung, ob die ergänzten Fusionsvariablen in der Analyse überhaupt eine Rolle spielen. Liefern die Fusionsvariablen keine zusätzliche Information und sind sie für die Analyse uninteressant, so wäre das Matching extern betrachtet ein Misserfolg. Hier sei festzuhalten, dass es durchaus sein kann, dass die interne Evaluation zu einem anderen, positiven Ergebnis kommt.

Um die Qualität des Matchings zu testen, kann darüber hinaus geprüft werden, ob die Hinzunahme der Fusionsvariablen eine Modellierung genauer macht. Hierzu wird ein statistisches Modell betrachtet, wie z.B. ein Regressionsmodell, welches einmal nur mit den ursprünglich vorgelegenen Daten und einmal mit den fusionierten Daten aufgestellt wird. Wird die Modellierung genauer (z. B. durch signifikante Verringerung der Standardfehler), so hatte die Hinzunahme der Fusionsvariablen einen Wert für die Analyse.

4.2.3 Anwendungsbereiche

Die oben beschriebene Problemstellung von der Analyser zweier (oder mehrerer) Variablen, die an keiner statistischen Einheit gleichzeitig erhoben worden sind, stellt nur einen von vielen Anwendungsbereichen des Statistical Matchings dar. Werden die nicht erfassten Variablen im Recipient-Set als fehlende Werte betrachtet, so lässt sich leicht ein Zusammenhang zur in Kapitel 3 behandelten Imputation fehlender Werte herstellen. Folgendes Beispiel soll die Vorgehensweise des Ersetzens fehlender Werte mithilfe von Statistical Matching erläutern: In einer Umfrage hat die befragte Person i keine Angabe zum Einkommen (Z) gemacht, d.h. dass in der Variablen Z an der Stelle i ein fehlender Wert vorliegt. Um diesen Wert zu ersetzen, wird ein statistischer Zwilling j gesucht, bei welchem eine Ausprägung der Variablen „Einkommen“ (Z) beobachtet wurde. Diese Person j soll zudem in ausgesuchten Variablen, welche so-

wohl bei i als auch bei j vorliegen, der Person i möglichst ähnliche Ausprägungen vorweisen. Diejenigen Variablen, welche einen Einfluss auf Z haben, sollten für das Matching ausgewählt werden. In diesem Beispiel würden sich Variablen wie „Alter“, „Allgemeiner Schulabschluss“ und „Beruf“ als Einflussvariablen eignen. Ist ein statistischer Zwilling j gefunden, so wird sein Wert von Z für den fehlenden Wert eingesetzt ($z_i = z_j$).

Ein weiterer Anwendungsbereich ist das Schätzen der Wirkung einer Untersuchungsvariablen. Angenommen man möchte anhand einer umfangreichen Umfrage in einer großen Grundgesamtheit den Einfluss der Dummy-Variablen X auf die Variable Y untersuchen. Desweiteren sei die Ausprägung x_1 nur sehr selten aufgetreten, während die meisten Befragten x_0 geantwortet haben. Zur Veranschaulichung beschreibe X die Anzahl der Geschwister, mit x_1: „5 und mehr Geschwister“ und x_0: „0-4 Geschwister“. Es sei nun zu prüfen, ob eine große Anzahl an Geschwistern einen signifikanten Einfluss auf die Variable Y: „Altruistisch veranlagt“ habe. Da die Ausprägung x_1 nur selten beobachtet wurde, ist die Vergleichsgruppe X_0[15] so groß, dass ein Vergleich zur relativ kleinen Untersuchungsgruppe erschwert wird. Die Ergebnisse der Analyse würden vor allem von der Vergleichsgruppe abhängen. Zudem ist es in solch einem Vergleich wahrscheinlich, dass eine signifikante Wirkung von X auf Y festgestellt wird, obwohl diese Wirkung eigentlich von anderen Variablen Z abhängt, welche sowohl mit X als auch mit Y korreliert sind. Soll der Einfluss der anderen Variablen Z kontrolliert werden, so würde man dies mithilfe einer multivariaten Analyse oder eines Strukturgleichungsmodells tun. Da eine relativ große Vergleichsgruppe auch meist große Heterogenität bedeutet, sind in den multivariaten „Standardmethoden“ große Verzerrungen zu erwarten. Denn in einer heterogenen Vergleichsgruppe kann es wiederum andere Wirkungszusammenhänge geben, welche nicht im Fokus der Analyse stehen. Um die Vergleichbarkeit innerhalb der Gruppen zu gewährleisten, wird hier auf das Statistical Matching zurückgegriffen. Dazu wird zu jeder Person i aus X_1 eine Person j aus X_0 gesucht, welche sich in den mit X und Y korrelierenden Variablen Z gar nicht oder kaum unterscheiden und als ihr statistischer Zwilling deklariert. So kann der Einfluss der anderen Variablen Z mithilfe der Methode des Statistical Matchings eliminiert werden. Ein ausführliches Anwendungsbeispiel wurde hierzu von Bacher (2002) vorgestellt.

Obwohl es viele weitere Problemstellungen gibt, bei welchen die Anwendung von Statistical Matching vermutlich zu guten Ergebnissen führen würde, wird es in der Praxis bisher meist nur zur Imputation fehlender Werte oder der Datenfusion genutzt.

[15] nachfolgend soll X_i die Gruppe der Personen beschreiben, welche in der Variablen X die Ausprägung x_i vorweisen.

Dies kann dadurch erklärt werden, dass die Methoden des Matchings leider nur selten oder rudimentär in Standard-Statistik-Software implementiert ist.

4.3 Praktische Anwendung: Statistical Matching

Um das Statistical Matching durchführen zu können, mussten zunächst die imputierten Datensätze der Gesundheits-Variablen wieder mit den Fällen aus 2002 zusammengeführt werden. Dazu wurde jeder der neun Datensätze um die 2820 Fälle aus ALLBUS 2002 ergänzt, in welchen nur Ausprägungen für die Einflussvariablen, nicht aber für die Gesundheits-Variablen vorlagen. Dies erfolgte analog mit den neun imputierten Datensätzen der Religions-Variablen. Zusätzlich musste in jedem Datensatz eine Variable ergänzt werden, welche bestimmte, ob der jeweilige Fall aus der Behandlungs- oder der Kontrollgruppe stammte. Diese Variable wurde mit „Treatment“ bezeichnet und nahm für alle Fälle der Behandlungsgruppe die Ausprägung 1, für alle Fälle der Kontrollgruppe die Ausprägung 0 an. Das Matching sollte in beide Richtungen durchgeführt werden, damit nach dem Matching für jede statistische Einheit aus beiden betrachteten ALLBUS-Datensätzen eine vollständige Datenreihe vorlag und keinerlei Information verloren ging. Somit sollten sowohl die Gesundheitsvariablen in ALLBUS 2002 als auch die Religionsvariablen für alle Fälle aus ALLBUS 2004 ergänzt werden. Demnach wurden in denjenigen Datensätzen, welche die Fusionsvariablen „Gesundheit“ enthielten, die Fälle aus ALLBUS 2002 als Behandlungsgruppe betrachtet, da sie diejenigen Fälle waren, die ergänzt („behandelt“) werden sollten. Analog stellten in allen Datensätze, welche die Religionsvariablen enthielten, die Einheiten aus 2004 die Behandlungsgruppe dar. Es lagen nun also neun Datensätze mit je 152 Variablen (96 Einflussvariablen, 55 Fusionsvariablen + eine Treatment-Variable) und neun Datensätze mit je 191 Variablen (96 Einflussvariablen, 94 Fusionsvariablen + eine Treatment-Variable) vor. Alle 18 Datensätze enthielten alle 5766 Fälle (2820 aus ALLBUS 2002 + 2946 aus ALLBUS 2004). Durch die zuvor vorgenommene Imputation wiesen die Datensätze keinerlei fehlende Werte mehr auf bis auf die fehlenden Ausprägungen der Behandlungsgruppe in den Fusionsvariablen. Diese fehlende Information sollte nun mithilfe von Statistical Matching ergänzt werden.

Für den Prozess des Matchings wurde die Statistik-Software R genutzt. Das in dieser Software installierbare Paket „Matching“ enthält einen Algorithmus, welcher zu jeder Einheit aus der Behandlungsgruppe einen oder mehrere statistische Zwillinge aus der Kontrollgruppe sucht. Im Anwendungsbeispiel sollte das Matching anhand von Propensity Scores erfolgen. Obwohl die Stichproben, welche für ALLBUS 2002 und ALLBUS 2004 gezogen wurden, die gleiche Population abbilden sollten, so war

#	UNIT2002	UNIT2004
1	1	3815
2	2	3903
3	3	4416
4	3	4558
5	3	4864
6	4	4250
7	4	4962
8	4	5672
⋮	⋮	⋮
13559	2820	5711

Tabelle 4.1: Tabelle der gefundenen statistischen Zwillinge bei ALLBUS 2002 als Treatment-Gruppe

nicht auszuschließen, dass sich die real betrachteten Gruppen stark voneinander unterschieden. Somit sollte durch Nutzen der Propensity Scores ein möglicher Einfluss der Gruppenzugehörigkeit ausgeschlossen werden. Dazu musste, wie zuvor beschrieben, zunächst eine logistische Regression der 96 Einflussvariablen auf die Treatment-Variable durchgeführt werden. Die Distanzen der Propensity Scores wurden im darauf folgenden „Matching“-Befehl als Maß für die Ähnlichkeit genutzt.

Nach Durchführung des Matchings wurde eine Tabelle[16] erzeugt, welche alle gefundenen Zwillingspaare auflistete. Ein Ausschnitt des Outputs für den Fall, dass die Befragten aus ALLBUS 2002 die Behandlungsgruppe darstellen, ist in Tabelle 4.1 zu sehen.

Da jedoch keine Information über die Qualität des durchgeführten Matchings vorlag, wurde nun die im R-Paket „Matching“ implementierte interne Evaluation betrachtet. Wie in Anhang A zu sehen ist, wurde der Wert schon während des Matching-Algorithmus erfasst. Über den Befehl `MatchBalance` wurden Balance-Statistiken sowohl vor als auch nach dem Matching berechnet, um sie dann miteinander zu vergleichen. Dabei wurden verschiedene statistische Tests angewendet (T-Test, Kolmogorov-Smirnov-Test), um auf Balance in einem Datensatz zu testen. Die ausgegebene Zusammenfassung[17] enthielt ebendiese Statistiken, jeweils vor und nach dem Matching. Die Ergebnisse dieses Vergleichs sind beispielhaft in Anhang B zu finden. Es waren keine signifikanten Unterschiede zwischen den Verteilungen vor und nach dem Matching festzustellen. Somit war die Zwillingssuche erfolgreich und die Paare konnten für die weitere Berechnung verwendet werden.

Nachdem mit Hilfe des „Match“-Befehls gute statistischen Zwillinge gefunden

[16] CD: Kapitel 4 – Statistical Matching/Zwillinge/Twins_g*
[17] CD: Kapitel 4 – Statistical Matching/Match Balance

wurden, konnten die fehlenden Ausprägungen in den Fusionsvariablen nun anhand Tabelle 4.1 geschätzt werden. Dazu wurden für jeden Fall aus ALLBUS 2002 die fehlenden Ausprägungen in den Fusionsvariablen durch die Ausprägungen des statistischen Zwillings ersetzt. Falls für eine Einheit mehrere Zwillinge gefunden worden waren, wurde pro Variable zufällig ein Zwilling ausgewählt und dessen Wert eingesetzt. Hierbei wurde für jede Variable ein neuer zufälliger Zwilling gesucht, um die Unsicherheit zu berücksichtigen, die durch die zufällige Wahl entstanden ist. Eine alternative Vorgehensweise wäre an dieser Stelle gewesen, für jeden Fall aus der Behandlungsgruppe nur einmalig einen statistischen Zwilling zufällig zu selektieren und dessen gesamte Werte für alle Variablen zu übertragen. Dies hätte wiederum den Vorteil, dass mögliche Strukturen innerhalb einer statistischen Einheit beibehalten würden. Das verwendete R-Programm[18], welches diese Imputation für jede fehlende Ausprägung der Fusionsvariablen durchgeführt hat, ist in Anhang A erläutert.

Dieser Algorithmus wurde nun jeweils auf die 9 Datensätze mit ALLBUS 2002 als Behandlungsgruppe und die 9 Datensätze mit ALLBUS 2004 als Behandlungsgruppe angewendet (vgl. Abb. 4.2). So wurden 18 gematchte Datensätze[19] generiert, die jeweils zwar alle 5766 Fälle beinhalteten, jedoch nur entweder Variablen zum Thema „Gesundheit“ oder zum Thema „Religion“ beinhalteten. Denn mit ALLBUS 2002 als Behandlungsgruppe bestand der neu erzeugte Datensatz nur aus den Fusionsvariablen zum Thema „Gesundheit“ und den gemeinsamen Variablen, anhand welchen gematcht wurde. Diese mussten nun zusammengeführt werden, um Datensätze mit allen im Kontext der Fragestellung relevanten Variablen zu erhalten. Für jeden der drei imputierten Datensätze der gemeinsamen Variablen lagen nun sechs gematchte Teildatensätze vor - jeweils drei pro Behandlungsgruppe. Da diese jeweils anhand des gleichen imputierten g-Datensatz gematcht wurden, konnte jeweils ein Datensatz mit ALLBUS 2002 als Behandlungsgruppe mit einem Datensatz mit ALLBUS 2004 als Behandlungsgruppe zusammengeführt werden. So entstanden neun gematchte Datensätze[20], welche 5766 Fälle zu allen betrachteten Variablen enthielten und zudem keinerlei fehlende Werte mehr aufwiesen. Diese sind schematisch in Abbildung 4.3 dargestellt.

Somit waren diese neu generierten Datensätze geeignet, um den Zusammenhang zwischen „Religion“ und „Gesundheit“ zu analysieren.

[18] CD: Kapitel 4 – Statistical Matching/R Algorithmen
[19] CD: Kapitel 4 – Statistical Matching/Gematchte Datensätze/Match*
[20] CD: Kapitel 4 – Statistical Matching/Gematchte Datensätze/Final kombiniert

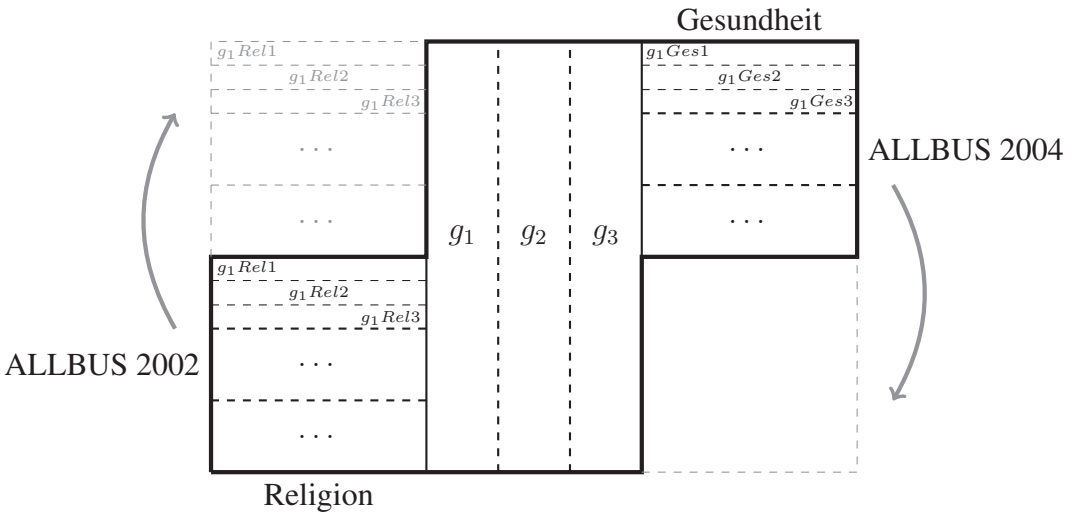


Abbildung 4.2: Grafische Darstellung der Imputation der Fusionsvariablen

g_1Rel1		g_1Ges3
g_1Rel2	g_1	g_1Ges1
g_1Rel3		g_1Ges2
g_2Rel3		g_2Ges1
g_2Rel2	g_2	g_2Ges2
g_2Rel1		g_2Ges3
g_3Rel2		g_3Ges3
g_3Rel1	g_3	g_3Ges2
g_3Rel3		g_3Ges1

Abbildung 4.3: Grafische Darstellung der Zusammenführung zu neun vollständigen Datensets

Kapitel 5

Analyse

> If physicians suggest, either directly or even implicitly, that faith and religious activity are associated with health, [...] then they indirectly suggest the opposite — which is that the disease and illness are associated with insufficient faith and insufficient devotion.
>
> Richard Sloan et al. (2000)

In dem in dieser Arbeit behandelten Anwendungsbeispiel sollte es das Ziel der Analyse sein, Zusammenhänge zwischen „Religiosität“ und „Gesundheit“ zu prüfen. Bei diesen beiden handelt es sich allerdings um latente, also nicht messbare, Variablen, die in den erzeugten Datensätzen nicht direkt abgefragt wurden. Vielmehr wurde eine Vielzahl von Variablen erhoben, von denen Schlüsse auf ebendiese latenten Variablen gezogen werden konnten. Daher mussten diese beiden Faktoren zunächst per Faktoranalyse extrahiert werden.

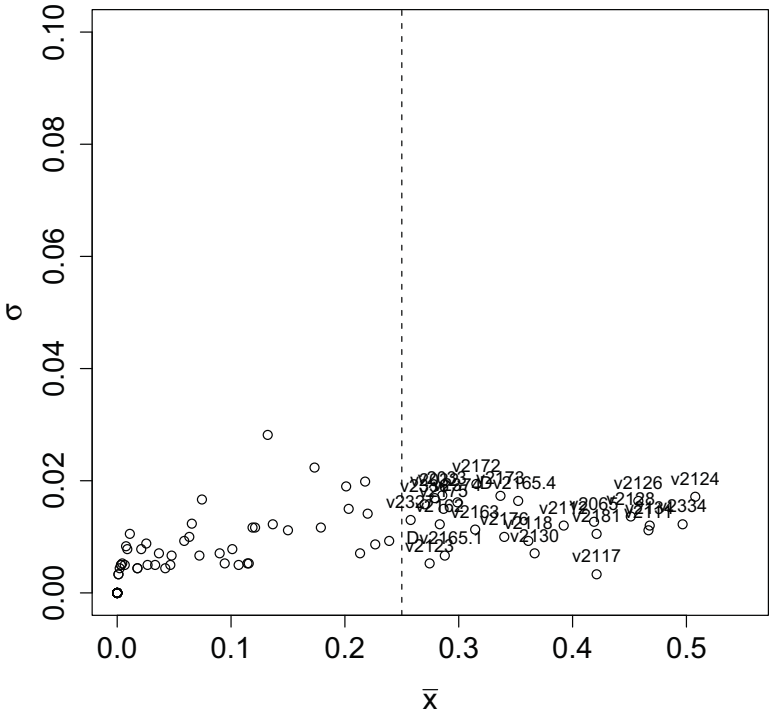
An jedem der neun Datensätze wurde daher eine Hauptkomponentenanalyse durchgeführt. Dabei wurden jeweils nur die Variablen zu den Themen „Religiosität“ und „Gesundheit“ miteinbezogen, da hier nur ihre Faktoren von Interesse waren. Die gemeinsamen Variablen dienten nur zur Vorbereitung der Datensätze und sollten in Anbetracht des Untersuchungsziels nur noch eine untergeordnete Rolle in der Analyse spielen. Nach dem Kaiser-Kriterium (Guttman, 1954) wurden jeweils diejenigen Faktoren extrahiert, deren Eigenwerte größer als 1 waren. Die Faktorladungen wurden anschließend nach dem Quartimax-Kriterium (Neuhaus & Wrigley, 1954) rotiert, um leichter interpretierbare Ergebnisse zu erhalten.

Betrachtete man die neun rotierten Faktorenmatrizen[21], so ließ sich Folgendes beobachten: Es wurden jeweils etwa 40 Faktoren extrahiert, was bei Einbeziehung von

[21] CD: Kapitel 5 – Analyse/FA_Match*

fast 150 Variablen nicht verwunderlich ist. In allen Datensätzen war eine klare Unterscheidung zwischen den Einflussfaktoren der Religionsvariablen und denen der Gesundheitsvariablen zu beobachten. Es wurden kaum Faktoren bestimmt, welche sich auf beide Variablengruppen auswirkten. In den Fällen, in welchen ein Faktor doch gruppenübergreifend wirkte, war der Einfluss verschwindend gering. In allen Datensätzen bestimmte der erste Faktor maßgeblich die Religionsvariablen, während der zweite Faktor den größten Einfluss auf die Gesundheitsvariablen hatte. Darüber hinaus war in allen Datensätzen zu beobachten, dass sich die Variablen zum Thema „Religion“ leichter von wenigen Faktoren beschreiben ließen. Neben dem ersten Faktor, hinter welchem sich in etwa „Religiosität / Wichtigkeit von Glauben“ zu verbergen schien, gab es einige andere Faktoren, welche einen relativ großen Einfluss auf weitere Religionsvariablen hatten. So ließ sich beispielsweise ein weiterer Faktor als „Glaube an Spirituelles“ interpretieren, da er insbesondere Variablen wie „Erfahrung mit Zen-Meditation“ beeinflusste. Bei den Gesundheitsvariablen gab es feinere Klassifikationen, die von den gleichen Faktoren bestimmt wurden. Es schien hier einzelne Faktoren für kleinere Unterthemen wie „Körperliche Gesundheit“, „Stress“ oder „Gesunde Ernährung“ zu geben. Diese Ergebnisse waren in allen der neun Datensätze wiederzufinden. Es war zu beobachten, dass sich die ersten beiden Faktoren und ihre jeweiligen Ladungen in allen Datensätzen kaum voneinander unterschieden. Um die Variation der Faktorladungen zu prüfen, wurde zunächst pro Faktor für jede Variable der Mittelwert und die Standardabweichung der quadrierten Faktorladungen über alle neun Datensätze errechnet. Diese beiden Parameter wurden dann für alle Variablen in einen Plot gezeichnet, um die Streuung der Faktorladungen beurteilen zu können. In den Abbildungen 5.1 und 5.2 sind ebendiese Plots für die ersten beiden Faktoren zu sehen. Zusätzlich wurde in die Plots eine vertikale Linie für $\bar{x} = 0{,}25$ gezeichnet. Große Faktorladungen bedeuten, dass der Faktor einen großen Einfluss auf die betrachtete Variable hat. Bei $\bar{x} = 0{,}25$ entspricht die gemittelte Fakorladung $0{,}5$. Diese soll als Grenzwert für eine große Faktorladung betrachtet werden. Somit wurden diejenigen Variablen, welche rechts von $\bar{x} = 0{,}25$ lagen, besonders stark vom betrachteten Faktor beeinflusst.

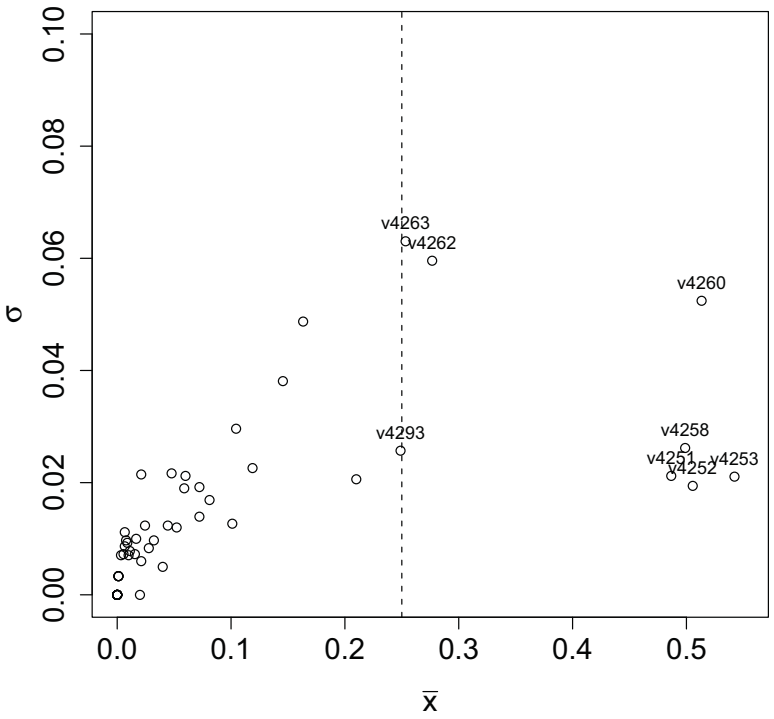
Es war zu sehen, dass die Streuung der Ladungen insbesondere beim ersten Faktor, welcher zuvor als „Religiosität“ interpretiert wurde, sehr klein war. Beim zweiten Faktor, welcher den größten Einfluss auf die Gesundheitsvariablen hatte, war eine zwar größere doch immer noch geringe Varianz zu sehen. Dies wurde deutlich, als die Ladungsstreuung einer der weniger einflussreichen Faktoren betrachtet wurde. Zur Veranschaulichung wurde in Abbildung 5.3 ein gleicher Plot für einen Faktor gezeichnet, welcher nur auf wenige Variablen und mit geringen Ladungen wirkte.



Variablen, dessen $\bar{x} > 0,25$:

v2032	Vertrauen: Katholische Kirche
v2033	Vertrauen: Evangelische Kirche
v2065	Leben mit: Religiosität
v2111	Gott befasst s. persönl. m. Menschen
v2112	Es ist ein Gott, der für uns sein will
v2117	Leben hat nur Bedeut., weil Gott ist
v2118	Lebenssinn, weil es nach Tod etw. gibt
v2123	Kirchliche Beerdigung?
v2124	Religiositätsskala, Befragte<r>
v2126	Egal, ob es Gott gibt
v2128	Folge keiner religiösen Lehre
v2130	Nachdenken über Glaubensfragen
v2134	Erfahrung: Nähe Gott. durch Glauben
v2162	Rel. bringt Menschen einand. näher
v2163	Mehr religiöse Menschen in Ämtern
v2172	Glauben an: Leben nach dem Tod
v2173	Glauben an: Himmel
v2175	Glauben an: Sünde
v2176	Glauben an: Vergebung
v2181	Glaube an Vergebung der Sünden?
v2327	Religiosität Mutter als Befr. Kind war
v2334	Befr.: Wie oft beten Sie? <7er Skala>
v2336	Häufigkeit v. kirchl. Aktivitäten
v2374	2.Haushaltsperson: Kirchgangshäufigkeit
Dv2165.1	Gottesglauben: Persönlicher Gott
Dv2165.4	Gottesglauben: Glaube nicht an Gott

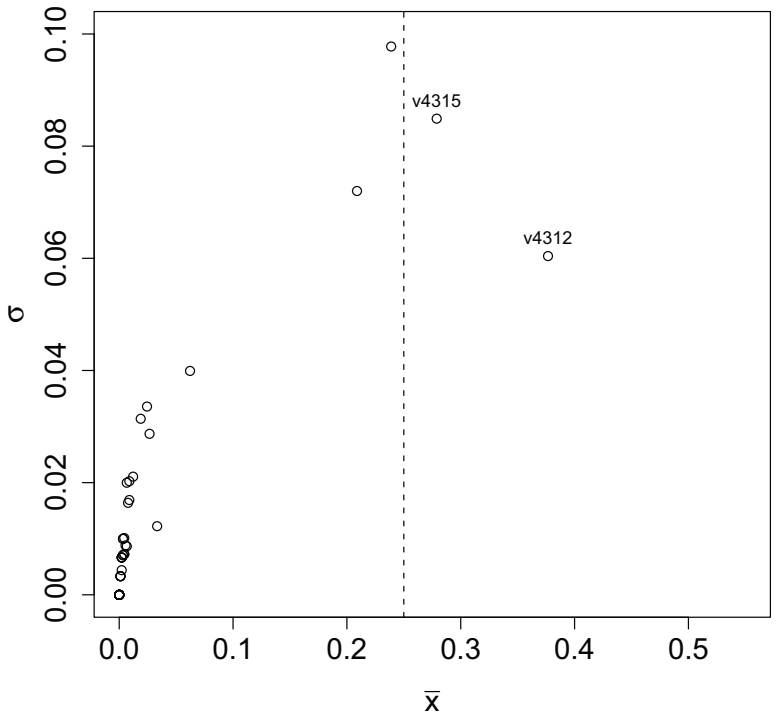
Abbildung 5.1: Geplottete Verteilungsparameter der Ladungen Faktor 1



Variablen, dessen $\bar{x} > 0,25$:

v4251	Gesundheitszustand Befr.
v4252	Gesundheitl. Probleme: Treppensteigen
v4253	Gesundheitl. Probleme: Alltagstätigkeit
v4258	Letzte 4 Wochen: Körperliche Schmerzen
v4260	Letzte 4 W.: Eingeschränkt wg. Körper
v4262	Letzte 4 W.: Eingeschränkt wg. Seele
v4263	Letzte 4 Wochen: Kontakte eingeschränkt
v4293	Befr. schwerbehindert?

Abbildung 5.2: Geplottete Verteilungsparameter der Ladungen Faktor 2



Variablen, dessen $\bar{x} > 0,25$:

v4312 Konsumhäufigkeit: Weissbrot, Toastbrot
v4315 Konsumhäufigkeit: Fleisch, Wurst

Abbildung 5.3: Geplottete Verteilungsparameter der Ladungen „kleiner“ Faktor

Inhaltlich könnte dieser Faktor als “Ungesunde Ernährung“ interpretiert werden. In Abbildung 5.3 ist zu sehen, dass die Streuung der Ladungen deutlich höher ist als bei den ersten beiden Faktoren. Zusätzlich ist zu erkennen, dass viele der Faktorladungen nahe 0 liegen, was zeigt, dass der Faktor nur auf wenige Variablen wirkt. Die ersten beiden Faktoren beeinflussten somit in allen Datensätzen ähnliche Variablen ähnlich stark. In den weniger einflussreichen Faktoren dagegen gab es teilweise größere Abweichungen. Doch auch diese waren zu vernachlässigen, da sie sich nur auf wenige Variablen auswirkten.

Zusätzlich wurde ein Mittelwert-Standardabweichungs-Plot für die Eigenwerte der extrahierten Faktoren gezeichnet. In Abbildung 5.4 ist zu sehen, dass es zwischen den neun Datensätzen auch wenig Streuung bezüglich der Eigenwerte gab. Die Eigenwerte der Faktoren nahmen also in allen neun Datensätzen ähnliche Werte an, wodurch ähnlich viele Faktoren einen Eigenwert von größer 1 aufwiesen. Dies erklärt, warum in allen gematchten Datensätzen in etwa die gleiche Anzahl an Faktoren extrahiert wurde.

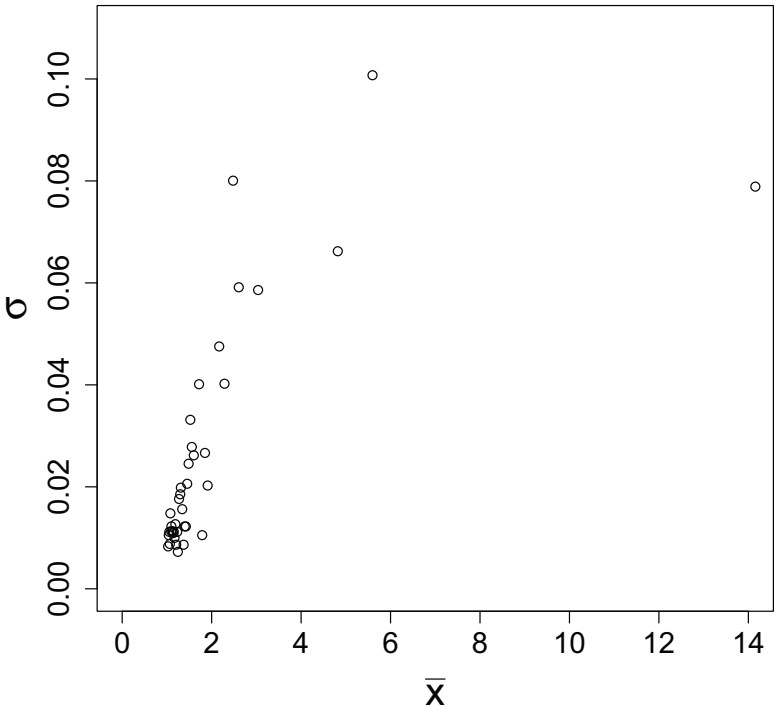


Abbildung 5.4: Geplottete Verteilungsparameter der Eigenwerte

All dies weist darauf hin, dass die durchgeführte Datenbearbeiung erfolgreich war und die Ergebnisse am Ende einer weiterführenden Analyse des Zusammenhangs von Religiosität und Gesundheit gut zu kombinieren wären.

Kapitel 6

Schlussbetrachtung

> Statistiken sind mit Vorsicht zu genießen und mit Verstand einzusetzen.
>
> Carl Hahn (2005)

Statistical Matching ist ein Werkzeug der Datenfusion, welches die Zusammenführung zweier Datensätze zum Ziel hat. Diese Technik erleichtert die Datenbeschaffung, da mithilfe des Matchings Informationen aus bereits erhobenen Daten gefolgert werden können, anstatt diese aufwändig selbst zu erheben.

Im Anwendungsbeispiel zeigte sich, dass durch Matching durchaus zufriedenstellende Ergebnisse zu erreichen sind. Es ist gelungen, zwei Datensätze adäquat zu einem einzigen zu fusionieren, in welchem die Analyse von Interesse umsetzbar ist. Die extrahierten Faktoren zu den Themen „Religiosität“ und „Gesundheit“ könnten im nächsten Schritt als Variablen interpretiert und auf Korrelation überprüft werden.

In dieser Arbeit ist allerdings auch deutlich geworden, dass der Matching-Prozess mit großem Aufwand verbunden ist, wenn man vor das Problem fehlender Werte gestellt ist. Denn um das Matching durchführen zu können, muss zunächst ein vollständiger Datensatz vorliegen. Dies ist in der Realität nur in den seltensten Fällen gegeben, weshalb man dem Schritt der Imputation oft nicht entgehen kann. Im Falle der multiplen Imputation verkompliziert dies den gesamten Prozess, da jeder darauffolgende Schritt an mehreren Datensätzen durchgeführt werden muss.

Als Alternative zum Statistical Matching per Zwillingssuche wäre es auch denkbar, die fehlende Information in den Fusionsvariablen als fehlende Werte zu interpretieren und sie dann mithilfe der multiplen Imputation zu ersetzen. Dies hätte allerdings eine weitere Potenzierung der Anzahl der zu analysierenden Datensätze zur Folge. Eine weiterführende Fragestellung wäre in diesem Zusammenhang, ob diese andere Herangehensweise an die Datenfusion zu anderen Ergebnissen führt als die Methode des Statistical Matchings.

In Zeiten des immer wichtiger werdenden Datenschutzes sollte man jedoch auch die Risiken der Matching-Technik bedenken. Denn mithilfe von Statistical Matching ist es möglich, in zwei unterschiedlichen Stichproben statistische Zwillinge auszumachen. Man stelle sich nun eine anonymisierte Umfrage vor, welche mit einem nichtanonymisierten Datensatz der gleichen Stichprobe (z.B. Einwohnermeldedaten) fusioniert wird. Die Anonymität könnte aus dem betrachteten Datensatz eliminiert werden, indem mithilfe des Melderegisters zu jedem anonymen Befragten die reale Person ausfindig gemacht wird. Dies stellt offensichtlich ein Risiko für den Datenschutz dar.

Anhang A

R Algorithmen

A.1 Auffinden statistischer Zwillinge mit Hilfe des Pakets „Matching“

```
## Am Beispiel für ALLBUS 2004 als Behandlungsgruppe ##

# Einbinden der benötigten Pakete
library(lattice)
library(rbounds)
library(Matching)
library(foreign)

# Schleife für alle drei Imputationen
j<-1
 while(j<=3){

# Einlesen des imputierten Datensatzes der gemeinsamen Variablen
y<-read.csv(sprintf("gImputation%x_TreatGes.csv",j),
header=TRUE, sep=";", dec=",")
dat<-data.frame(y)
  assign(sprintf("datimp%sG",j),dat)

# Durchführung einer logistischen Regression auf die
# "Treatment"-Variable
glmy<-glm(treat~g1001 + g1073 + g1128 + g1137 + g1146 + g1167
+ g1168 + g1208 + g1214 + g1268 + g1305 + g1318 + g1334 + g1353
+ DUM1.1 + DUM1.3 + DUM1.4 + DUM1.5 + DUM1.6 + DUM1.7 + DUM2.1
+ DUM2.2 + DUM2.3 + DUM2.4 + DUM2.5 + DUM2.6 + DUM2.7 + DUM2.9
+ DUM2.10 + DUM2.11 + DUM2.12 + DUM2.13+ DUM2.14 + DUM2.15
+ DUM2.16 + DUM2.17 + DUM3.2 + DUM3.3 + DUM3.4 + DUM3.5 + DUM4.1
+ DUM4.3 + DUM4.4 + DUM4.5 + DUM4.6 + DUM5.2 + DUM6.1 + DUM6.3
```

```
+ DUM6.4 + DUM6.5 + DUM6.6 + DUM6.7 + DUM6.8 + DUM6.9 + DUM6.10
+ DUM6.11 + DUM6.12 + DUM7.1 + DUM7.3 + DUM7.4 + DUM9.2 + DUM9.3
+ DUM10.2 + DUM10.3 + DUM10.4 + DUM10.5 + DUM10.6 + DUM11.1
+ DUM11.3 + DUM11.4 + DUM11.5 + DUM11.6 + DUM12.1 + DUM12.2
+ DUM12.4 + DUM12.5 + DUM12.6 + DUM13.1 + DUM13.3 + DUM13.4
+ DUM13.5 + DUM13.6 + DUM14.1 + DUM14.2 + DUM14.3 + DUM14.4
+ DUM14.5 + DUM14.8 + DUM14.9 + DUM15.2 + DUM15.3 + DUM15.4
+ DUM15.5 + DUM15.6 + DUM15.7 + DUM15.8,
family = gaussian, data = dat)

# Auffinden statistischer Zwillinge
X  <- glmy$fitted
Tr  <- dat$treat

Ma <- Match(Tr=Tr, X=X, M=1)

# Prüfen der Qualität der gefundenen Zwillingspaare
Mb <- MatchBalance(treat~g1001 + g1073 + g1128 + g1137 + g1146
+ g1167 + g1168 + g1208 + g1214 + g1268 + g1305 + g1318 + g1334
+ g1353 + DUM1.1 + DUM1.3 + DUM1.4 + DUM1.5 + DUM1.6 + DUM1.7
+ DUM2.1 + DUM2.2 + DUM2.3 + DUM2.4 + DUM2.5 + DUM2.6 + DUM2.7
+ DUM2.9 + DUM2.10 + DUM2.11 + DUM2.12 + DUM2.13+ DUM2.14 + DUM2.15
+ DUM2.16 + DUM2.17 + DUM3.2 + DUM3.3 + DUM3.4 + DUM3.5 + DUM4.1
+ DUM4.3 + DUM4.4 + DUM4.5 + DUM4.6 + DUM5.2 + DUM6.1 + DUM6.3
+ DUM6.4 + DUM6.5 + DUM6.6 + DUM6.7 + DUM6.8 + DUM6.9 + DUM6.10
+ DUM6.11 + DUM6.12 + DUM7.1 + DUM7.3 + DUM7.4 + DUM9.2 + DUM9.3
+ DUM10.2 + DUM10.3 + DUM10.4 + DUM10.5 + DUM10.6 + DUM11.1
+ DUM11.3 + DUM11.4 + DUM11.5 + DUM11.6 + DUM12.1 + DUM12.2
+ DUM12.4 + DUM12.5 + DUM12.6 + DUM13.1 + DUM13.3 + DUM13.4
+ DUM13.5 + DUM13.6 + DUM14.1 + DUM14.2 + DUM14.3 + DUM14.4
+ DUM14.5 + DUM14.8 + DUM14.9 + DUM15.2 + DUM15.3 + DUM15.4
+ DUM15.5 + DUM15.6 + DUM15.7 + DUM15.8,
data = dat, match.out= Ma, nboots=500)
assign(sprintf("Mb_g%sG",j),Mb)

# Erstellen einer übersichtlichen Liste der gefundenen
# Zwillingspaare
u<-Ma$MatchLoopC[,1]
v<-Ma$MatchLoopC[,2]

y<-data.frame(UNIT2004=u,UNIT2002=v)
```

```
# Ergänzen der Liste um eine zusätzliche Zeile, um Abbruch-
# kriterium zu gewährleisten.
y<-rbind.data.frame(y,c(5767,0))
assign(sprintf("Twins_g%sG",j),y)

j<-j+1}
```

A.2 Imputation der Werte

```
# Wiederholung des Algorithmus für alle drei imputierten Datensätze,
# welche aus dem gleichen imputierten Datensatz der gemeinsamen
# Variablen resultieren
j<-1
while(j<=3){

# Einlesen des Datensatzes, welcher die imputierten
# Gesundheits-Variablen enthält
y<-read.csv(sprintf("gImputation2_TreatGes%x.csv",j),
header=TRUE, sep=";", dec=",")
dat<-data.frame(y)

# Einsetzen der Ausprägungen eines der gefundenen Zwillinge
# (Schleife für alle 2820 Fälle und 55 Variablen)
a<-97
while(a<=190){
 c<-1
 z<-1
 while(z<=2946){
  I<-0
  y<-Twins_g2G[c,1]
  k<-0
   while(y==Twins_g2G[c,1]){
   k<-k+1
   c<-c+1
   }
if (k>1){
  g<-runif(1,1,k)}
else {
  g<-1}
  I<-dat[Twins_g2G[c-g,2],a]
  dat[y,a]<-I
  z<-z+1
  }
a<-a+1}
```

```
# Exportieren des gematchten Datensatzes in eine SPSS-Datei
write.table(dat, sprintf("dat2G%x.xls",j),
sep =";", dec=",", row.names = FALSE)
write.foreign(dat, sprintf("Match2G%x.dat",j),
sprintf("Match2G%x.sps",j), package="SPSS")

j<-j+1}
```

Anhang B

Zusammenfassung der Match Balance

Für (g1001) 2. HH-Person, Alter:

	Before Matching	After Matching
mean treatment.....	49.07739	49.07739
mean control.....	48.53014	48.20969
std mean diff.....	3.376150	5.353135
mean raw eQQ diff.....	0.7056738	0.5133124
med raw eQQ diff.....	1	0
max raw eQQ diff.....	4	4
mean eCDF diff.....	0.007221417	0.005234749
med eCDF diff.....	0.003727015	0.003337681
max eCDF diff.....	0.02762491	0.01695696
var ratio (Tr/Co).....	1.002828	0.9918605
T-test p-value.....	0.1997387	0.04128314
KS Bootstrap p-value..	0.144	0.028
KS Naive p-value.....	0.2215101	0.04715322
KS Statistic.....	0.02762491	0.01695696

Anhang C

Legende zur Daten-CD

- **Kapitel 2 — Ausgangsdatensätze**
  - ALLBUS2002.sav
  - ALLBUS2004.sav
- **Kapitel 3 — Multiple Imputation**
  - **Imputation G**
    - da.out
    - em.out
    - gImputation1.sav
    - gImputation2.sav
    - gImputation3.sav
  - **Imputation Gesundheit**
    - da1g.out
    - da2g.out
    - da3g.out
    - em1g.out
    - em2g.out
    - em3g.out
    - gImputation1_Ges1.sav
    - gImputation1_Ges2.sav
    - gImputation1_Ges3.sav
    - gImputation2_Ges1.sav
    - gImputation2_Ges2.sav
    - gImputation2_Ges3.sav
    - gImputation3_Ges1.sav
    - gImputation3_Ges2.sav
    - gImputation3_Ges3.sav
  - **Imputation Religion**
    - da1r.out
    - da2r.out
    - da3r.out
    - em1r.out
    - em2r.out

        * em3r.out
        * gImputation1_Rel1.sav
        * gImputation1_Rel2.sav
        * gImputation1_Rel3.sav
        * gImputation2_Rel1.sav
        * gImputation2_Rel2.sav
        * gImputation2_Rel3.sav
        * gImputation3_Rel1.sav
        * gImputation3_Rel2.sav
        * gImputation3_Rel3.sav
    - 20022004_28gVariablen.sav
    - 20022004_28gVariablen_Dummies.sav
- **Kapitel 4 — Statistical Matching**
    - **Gematchte Datensätze**
        * **Final kombiniert**
            · M11.sav
            · M12.sav
            · M13.sav
            · M21.sav
            · M22.sav
            · M23.sav
            · M31.sav
            · M32.sav
            · M33.sav
        * M1G1.sav
        * M1G2.sav
        * M1G3.sav
        * M1R1.sav
        * M1R2.sav
        * M1R3.sav
        * M2G1.sav
        * M2G2.sav
        * M2G3.sav
        * M2R1.sav
        * M2R2.sav
        * M2R3.sav
        * M3G1.sav
        * M3G2.sav
        * M3G2.sav
        * M3R1.sav
        * M3R2.sav
        * M3R3.sav
    - **Match Balance**
        * Mb_g1G.txt
        * Mb_g1R.txt

        - Mb_g2G.txt
        - Mb_g2R.txt
        - Mb_g3G.txt
        - Mb_g3R.txt
        - Mb_summaries.txt
    - **R Algorithmen**
        - Imputation_TrG.txt
        - Imputation_TrR.txt
        - Matching_TrG.txt
        - Matching_TrG.txt
    - **Zwillinge**
        - Twins_g1G.txt
        - Twins_g1R.txt
        - Twins_g2G.txt
        - Twins_g2R.txt
        - Twins_g3G.txt
        - Twins_g3R.txt
- **Kapitel 5 — Analyse**
    - FA_Match11.xls
    - FA_Match12.xls
    - FA_Match13.xls
    - FA_Match21.xls
    - FA_Match22.xls
    - FA_Match23.xls
    - FA_Match31.xls
    - FA_Match32.xls
    - FA_Match33.xls
    - Plot_Eigenwerte.pdf
    - Plot_Faktor1.pdf
    - Plot_Faktor2.pdf
    - Plot_Faktor_klein.pdf

Literaturverzeichnis

[1] BACHER, J. Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS. *ZA-Informationen*, 51 (2002), 38–66.

[2] BAKER, K., HARRIS, P., AND O'BRIEN, J. Data Fusion: An Appraisal and Eperimental Evaluation. *Journal of the Market Research Society*, 31 (1989), 153–212.

[3] GESIS – LEIBNIZ-INSTITUT FÜR SOZIALWISSENSCHAFTEN. Allbus – Allgemeine Informationen. Website, 30. Oktober. 2010. http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/allgemeine-informationen/.

[4] GUTTMAN, L. Some Necessary Conditions for Common Factor Analysis. *Psychometrika*, 19 (1954), 149–161.

[5] HAHN, C. *Meine Jahre mit Volkswagen*. Signum, 2005.

[6] KIM, J.-O., AND CURRY, J. The Treatment of Missing Data in Multivariate Analysis. *Sociological Methods & Research 6*, 2 (1977), 215–240.

[7] LITTLE, R. J., AND RUBIN, D. B. *Statistical Analysis With Missing Data*. John Wiley & Sons, 1987.

[8] NEUHAUS, J. ., AND WRIGLEY, C. The Quartimax Method: An Analytic Approach to Orthogonal Simple Structure. *British Journal of Statistical Psychology*, 7 (1954), 81–91.

[9] ROHRSCHNEIDER, L. Behandlung fehlender Daten. Master's thesis, Humboldt Universität zu Berlin, 2007. http://edoc.hu-berlin.de/docviews/abstract.php?id=28131.

[10] RÄSSLER, S. *Statistical Matching*. Springer, 2002.

[11] RUBIN, D. B. *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons, 1987.

[12] SCHAFER, J. L., AND GRAHAM, J. W. Missing Data: Our View of the State of the Art. *Psychological Methods 7*, 2 (2002), 147–177.

[13] SIMS, C. Comments. *Annals of Economic and Social Measurement*, 1 (1972a), 343–345.

[14] SIMS, C. Rejoinder. *Annals of Economic and Social Measurement*, 1 (1972b), 355–357.

[15] SLOAN, R., BAGIELLA, E., VANDECREEK, L., HOVER, M., CASALONE, C., HIRSCH, T., HASAN, Y., KREGER, R., AND POULOS, P. Should Physicians prescribe Religious Activities? *New England Journal of Medicine*, 342 (2000), 1913–1916.

[16] SMITH, H. L. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. *Sociological Methodology 27* (1997), 325–353.

[17] VAN DER PUTTEN, P., KOK, J. N., AND GUPTA, A. Data Fusion Through Statistical Matching. MIT Sloan School of Management, 2002. Paper 185.

[18] ZAHN, D. Psychosoziale Belastung bei Herztransplantationskandidaten. Master's thesis, Johannes Gutenberg-Universität Mainz, 2009. http://ubm.opus.hbz-nrw.de/volltexte/2009/2104/pdf/diss.pdf.

Erklärung zur Urheberschaft

Hiermit erkläre ich, Sarah Asmah, dass ich die vorliegende Arbeit allein und nur unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.
Die Prüfungsordnung ist mir bekannt. Ich habe in meinem Studienfach bisher keine Bachelorarbeit eingereicht bzw. diese nicht endgültig nicht bestanden.

_____

Sarah Asmah

Berlin, den 8. November 2010