

# 5 Multiple Imputationen

## 5.1 Einleitung

Ein bei allen freiwilligen Erhebungen auftretendes Problem ist der partielle Antwortausfall (Item-Non-Response), d. h. die Tatsache, dass nicht alle Befragten sämtliche Fragen beantworten.<sup>1</sup> Dazu kommt es häufig, wenn komplizierte oder als sensibel erachtete Fragen (z. B. zu Einkommen und Vermögen) gestellt werden.

Würde das Problem der fehlenden Angaben außer Acht gelassen werden, wären die aus Analysen resultierenden Schätzergebnisse verzerrt. Um dieses Problem zu behandeln, wird im HFCS die Methode der multiplen Imputation anhand von zusammenhängenden Gleichungen (Chained Equations) angewendet.

Dabei werden fehlende Werte (Missing Values) im Datensatz jeweils durch mehrere Werte ersetzt, die auf Grundlage eines iterativen Bayesschen Modells geschätzt werden. Das Hauptziel dieses Verfahrens ist es, dass die imputierten Werte den Zusammenhang zwischen allen Variablen im Sinne der Erhaltung der Korrelationsstrukturen des Datensatzes bewahren. Daher werden die fehlenden Werte jeder Variable unter Berücksichtigung einer maximalen Anzahl verfügbarer Variablen geschätzt. Um der statistischen Unsicherheit bezüglich der fehlenden Werte Rechnung zu tragen, wird nicht nur ein Wert für jeden Missing Value, sondern mehrere (im HFCS sind es fünf) imputiert.

Andere, dem HFCS ähnliche Erhebungen, wie der *Survey of Consumer Finances* (SCF – siehe Kennickell, 1998) oder die spanische *Encuesta Financiera de las Familias* (EFF – siehe Barceló, 2006), beruhen auf demselben Ansatz der Imputation fehlender Werte.

Da multiple Imputationen ein ausgesprochen zeitintensiver Prozess sind, stellen die meisten Institutionen, wie auch der HFCS, den Nutzern bereits imputierte Datensätze zur Verfügung. Dies stellt sicher, dass jeder Datennutzer mit denselben imputierten Datensätzen arbeiten kann. Im Fall des HFCS können Nutzer alle imputierten Werte einer Variable anhand der entsprechenden Flag-Variable (siehe Abschnitt 4.5) erkennen und haben somit auch die Möglichkeit, eigenständig Non-Response-Analysen oder Imputationen durchzuführen bzw. andere Arten der Berücksichtigung der Non-Response in den Analysen zu verwenden.

Dieses Kapitel ist wie folgt aufgebaut: In Abschnitt 5.2 werden Daten zur Item-Non-Response im HFCS präsentiert. Danach folgt in Abschnitt 5.3 eine Darstellung des angewandten Imputationsverfahrens. In Abschnitt 5.4 wird erklärt, wie das Imputationsmodell spezifiziert und die Imputationen durchgeführt wurden. Abschließend werden in Abschnitt 5.5 einige Imputationsergebnisse präsentiert.

## 5.2 Item-Non-Response

Tabelle 4 zeigt ausgewählte Informationen zur Item-Non-Response. Im Durchschnitt weist jeder Haushalt 17,3 Missing Values auf, was einen Antwortausfall bei lediglich etwa 2 % der insgesamt abgefragten Variablen darstellt. Bei den Betragsvariablen beträgt der betroffene Prozentsatz allerdings 6,9 %. Dies dokumentiert,

<sup>1</sup> Ein weiteres bei Erhebungen häufig auftretendes Problem ist der vollständige Antwortausfall einer Erhebungseinheit (Unit-Non-Response), d. h. dass ein Haushalt überhaupt keine Fragen beantwortet, da er z. B. die Teilnahme an der betreffenden Erhebung ablehnt. Dieses Problem wird durch die Berechnung von Non-Response-Gewichten für den HFCS behandelt (siehe Kapitel 7).

Tabelle 4

**Item-Non-Response je Haushalt (ungewichtet)**

	Mittelwert	Median	Minimum	Maximum
Anzahl der abgefragten Variablen				
Alle Variablen	826,8	824,0	637	1.242
Euro-Variablen	52,1	53,0	17	98
Anzahl der Variablen mit Missing Values				
Alle Variablen	17,3	8,0	0	474
Euro-Variablen	3,6	2,0	0	54
Anteil der Variablen mit Missing Values in %				
Alle Variablen	2,0	1,0	0	39,5
Euro-Variablen	6,9	4,2	0	78,8

Quelle: HFCS Austria 2010, OeNB.

Anmerkung: Intervallangaben werden als Missing Values der entsprechenden Betragsvariable, und nicht als eigene Variable erfasst. Wird eine Frage mehreren Haushaltsmitgliedern gestellt, wird für die Antwort eines jeden Haushaltsmitglieds eine eigene Variable erfasst.

dass die diesbezüglichen Fragen als sensibel empfunden werden dürften bzw. ihre Beantwortung besonders schwierig ist.

Es gibt verschiedene Ansätze zur Analyse von Datensätzen, in denen für bestimmte Variablen nicht alle Werte verfügbar sind.<sup>2</sup> In den meisten Statistikpaketen wird standardmäßig ein fallweises Ausschlussverfahren eingesetzt (auch Complete Case Analysis genannt). Dabei werden alle Haushalte, bei denen eine der in der Analyse verwendeten erhobenen Variablen fehlende Werte aufweist, gelöscht und ausschließlich vollständige Beobachtungen in die Analyse miteinbezogen. Allerdings ergeben sich aus dem daraus resultierenden Informationsverlust zwei Probleme: Zum einen führt er zu einer Verzerrung der Schätzer, wenn systematische Unterschiede zwischen vollständigen und unvollständigen Beobachtungen vorliegen, zum anderen – selbst wenn der Schätzer unverzerrt wäre – könnte durch den Verlust von Beobachtungen der Schätzer weniger präzise geschätzt werden. Um zu zeigen, wie groß das Ausmaß des Informationsverlusts im Fall des HFCS wäre, sind in Tabelle 5 Item-Non-Response-Quoten für ausgewählte Variablen dargestellt.

Anhand der Tabelle 5 lässt sich z. B. erkennen, dass – nach dem Wert ihres Hauptwohnsitzes gefragt – 75,5 % der Haushalte einen Betrag nennen (Spalte 3). Bei den übrigen 24,5 % der Haushalte kam es zu einem partiellen Antwortausfall, d. h., dass sie entweder ein (vorgegebenes oder individuelles) Intervall nannten (15,3 %, Spalte 4), mit „Weiß nicht“ antworteten bzw. „Keine Angabe“ machten (6,4 %, Spalte 5) oder auf Missing editiert wurden<sup>3</sup> (2,7 %, Spalte 6). Die Non-Response-Quoten<sup>4</sup> fallen je nach abgefragter Position sehr unterschiedlich aus. Eine hohe Non-Response-Quote weisen etwa die Fragen nach dem ausstehenden Betrag der durch den Hauptwohnsitz besicherten Hypothek (100 % – 63,5 % = 36,5 %) bzw. nach dem Bruttoeinkommen des Haushalts aus Finanzanlagen (100 % – 34,3 % = 65,7 %) auf. Im Hinblick auf letztere Variable geben 40,7 % der Haus-

<sup>2</sup> Siehe dazu ausführlich Little und Rubin (2002).

<sup>3</sup> Nähere Details finden sich in Kapitel 4.

<sup>4</sup> Die Non-Response-Quote errechnet sich als 100% abzüglich des Werts in der Spalte „Betrag“ in Tabelle 5.

**Item-Non-Response bei ausgewählten Variablen (ungewichtet)**

	Haushalt verfügt über das Item		Angaben jener Haushalte, die über das Item verfügen			
	Ja	Unbekannt	Betrag	Intervall	„Weiß nicht“/ „Keine Angabe“	Sonstiges Missing <sup>1</sup>
	(1)	(2)	(3)	(4)	(5)	(6)
	in %					
Wert des Hauptwohnsitzes <sup>2</sup>	49,6	0,0	75,5	15,3	6,4	2,7
Durch Hauptwohnsitz besicherte Hypothek 1: ausstehender Kapitalbetrag	15,1	1,4	63,5	21,2	12,3	3,1
Monatliche Miete	44,1	0,0	97,1	2,3	0,5	0,1
Sonstiges Immobilieneigentum 1: Marktwert	12,6	0,2	74,1	15,3	9,6	1,0
Durch sonstige Immobilien besicherte Hypothek 1: ausstehender Kapitalbetrag	1,7	0,4	70,7	7,3	14,6	7,3
Guthaben auf Girokonten	98,9	0,0	72,0	13,3	14,4	0,3
Guthaben auf Sparkonten	86,0	1,6	64,6	18,6	16,0	0,8
Wert börsennotierter Aktien	5,4	0,4	71,1	12,5	16,4	0,0
Geldschulden gegenüber dem Haushalt	9,3	0,5	90,5	5,0	4,5	0,0
Beschäftigungsstatus (Hauptbeschäftigung) (Person 1)	100,0	0,0	99,9	0,0	0,1	0,0
Bruttoeinkommen aus abhängiger Beschäftigung (Person 1)	48,8	0,1	76,7	9,9	3,4	9,9
Bruttoeinkommen aus der Arbeitslosenunterstützung (Person 1)	6,1	0,1	83,3	9,7	6,3	0,7
Bruttoeinkommen aus Finanzanlagen	70,9	6,6	34,3	40,7	24,0	0,9
Schenkung/Erbschaft 1: Wert	21,4	1,3	71,1	16,3	10,0	2,6
Ausgaben für Lebensmittel zu Hause	100,0	0,0	96,3	3,4	0,3	0,0

Quelle: HFCS Austria 2010, OeNB.

<sup>1</sup> Missing Values aufgrund von Editierungsmaßnahmen und dem Ausstieg aus Schleifen.

<sup>2</sup> Hierfür wurde die Variable HB0900 verwendet.

halte zumindest ein Intervall an, in dem das Einkommen liegt. Dies zeigt deutlich, wie wichtig es ist, bei Nichtbeantwortung von Betragsfragen zusätzlich nach numerischen Intervallen zu fragen. Sie liefern wertvolle und oft sehr genaue Informationen (siehe Online-Anhang und Abschnitt 2.6.1 für den Fragebogen und den Aufbau der Euro-Schleifen). Variablen mit niedriger Non-Response-Quote sind z. B. Nicht-Betragsvariablen wie Beschäftigungsstatus ( $100\% - 99,9\% = 0,1\%$ ) oder die Ausgaben für zu Hause verzehrte Lebensmittel ( $100\% - 96,3\% = 3,7\%$ ).

Aus Tabelle 5, Spalte 2, geht ein weiterer Aspekt der Item-Non-Response im HFCS hervor. Denn es gibt Variablen – sogenannte Branch-Variablen (siehe Grafik 3 in Kapitel 4) –, die aufgrund der Nichtbeantwortung einer übergeordneten Frage (Head-Variable) fehlende Werte aufweisen können (und somit auf „Missing“ gesetzt werden). Zum Beispiel wird Haushalten noch vor der Betragsfrage zum Bruttoeinkommen aus Finanzvermögen die Entscheidungsfrage gestellt, ob sie über derartiges Einkommen verfügen, und nur jene Haushalte, die dies bestätigen (70,9%), gelangen überhaupt zur Frage nach der Höhe des Einkommens. Bei den übrigen Haushalten – einschließlich jener 6,6%, die die Entscheidungsfrage nicht beantworten – wird die Betragsfrage übersprungen. Im Hinblick auf den Antwortausfall bei der Betragsfrage muss allerdings die Antwortverweigerung jener Haushalte (6,6%), die die binäre Frage nicht beantwortet haben, als fehlender Wert zweiter (oder höherer) Ordnung berücksichtigt werden, da nicht bekannt ist, ob diese Haushalte ein positives Bruttoeinkommen aus Finanzvermögen haben oder nicht.

Tabelle 6

### Logit-Regression des Antwortausfalls bei der Betragsfrage zu Girokontoguthaben (ungewichtet)

Kovariaten	Koeffizient
Weiblich (Person 1)	0,0775 (0,0950)
Alter (Person 1)	-0,0012 (0,00344)
Hochschulabschluss (Person 1)	-0,259* (0,156)
Unselbstständig bzw. selbstständig erwerbstätig (Person 1)	-0,195* (0,113)
Wohnhaft in Wien	-0,194 (0,134)
Wohnfläche Hauptwohnsitz	0,00274*** (0,000863)
Haushaltsgröße	0,119*** (0,0421)
Konstante	-1,331*** (0,256)
Beobachtungen <sup>1</sup>	2.330

Quelle: HFCS Austria 2010, OeNB.

Anmerkung: Angabe von Standardfehlern in Klammern; \*\*\*  $p < 0,01$ , \*\*  $p < 0,05$ , \*  $p < 0,1$ .

<sup>1</sup> Die restlichen 50 Beobachtungen des Datensatzes weisen fehlende Werte bei einer der Kovariaten auf bzw. Filter Missing bei der abhängigen Variable und werden daher in der Regression nicht berücksichtigt.

Wenn bei den HFCS-Daten ein fallweises Ausschlussverfahren eingesetzt würde, wären der Informationsverlust und die daraus resultierende Verminderung der Präzision von unverzerrten Schätzern auch aufgrund der zahlreichen Variablen mit fehlenden Werten höherer Ordnung also potenziell beträchtlich. Da außerdem vollständige Beobachtungen normalerweise systematisch von unvollständigen abweichen, führt Complete Case Analysis zu einer Verzerrung der Schätzwerte.

Zur Illustration ist in Tabelle 6 eine Regression des Antwortausfalls bei der Frage zu Guthaben auf Girokonten (1 bei fehlendem Wert, ansonsten 0) auf mehrere Erklärungsvariablen dargestellt. Es zeigt sich, dass sich Item-Respondenten signifikant von Item-Non-Respondenten unterscheiden, da Erstere in kleineren Wohnungen und in kleineren Haushalten leben, über ein höheres Ausbildungsniveau verfügen und häufiger beschäftigt sind. Bei Anwendung des fallweisen Ausschlusses im Hinblick auf Girokontoguthaben würden die Schätzer demnach zugunsten einer Population mit diesen Haushaltscharakteristika verzerrt werden.

### 5.3 HFCS-Imputationsverfahren

Zur Imputation von HFCS-Daten wählen wir ein von Royston (2004) in der Statistiksoftware Stata<sup>®</sup> implementiertes Verfahren, in dem alle zu imputierenden Variablen in Regressionsgleichungen geschätzt werden (Chained Equations)<sup>5</sup>. Das Verfahren<sup>6</sup> kann in die folgenden Schritte unterteilt werden:

- Schritt 1: Auswahl der zu imputierenden  $P$  Variablen  $Y_1, Y_2, \dots, Y_p$
- Schritt 2: Ersetzen der fehlenden Werte von  $Y_1, Y_2, \dots, Y_p$  mit zufälligen Ziehungen aus tatsächlich beobachteten Werten
- Schritt 3: Für  $Y_1, Y_2, \dots, Y_p$ 
  - wird jeweils eine Bayessche Regression von der zu imputierenden Variable auf ein umfangreiches Set unabhängiger Variablen durchgeführt, die aus den HFCS-Variablen ohne fehlende Werte und aus den in Schritt 1 ausgewählten Variablen (außer jener Variable, für die die Regression durchgeführt wird) ausgewählt werden; die Regression ist auf Beobachtungen zu beschränken, die in der abhängigen Variable nicht fehlen,

<sup>5</sup> Dieses Verfahren ist in der englischen Fachsprache auch unter anderen Namen bekannt wie Stochastic Relaxation, Regression Switching, Sequential Regression, Incompatible MCMC oder Fully Conditional Specification.

<sup>6</sup> Zu weiterführenden technischen Details siehe Albacete (2012), in dem das Imputationsverfahren der Immobilienvermögenserhebung genauer beschrieben wird, welches dem des HFCS ident ist.

- wird nach dem Zufallsprinzip ein Vektor von Regressionsparametern aus deren A-posteriori-Verteilung gezogen,
  - werden die entsprechenden vorhergesagten Werte berechnet und als imputierte Werte verwendet,
  - werden die fehlenden Werte der imputierten Variable durch ihre imputierten Werte ersetzt.
- Schritt 4: Schritt 3 ist  $t$ -mal zu wiederholen. Dabei sind die jeweils zuvor verwendeten imputierten Werte durch aktualisierte, aus der jeweils letzten Regression gewonnene zu ersetzen. Auf dieser Grundlage wird das erste Imputationssample geschaffen.
  - Schritt 5: Schritt 3 und 4 sind  $M$ -mal unabhängig voneinander zu wiederholen, um  $M$  Imputationssamples zu generieren.

Die Grundidee, auf der dieses Verfahren basiert, ist die Imputation fehlender Werte für jede der  $P$  Missing-Variablen auf Grundlage von Prognosen mittels eines Bayesschen Regressionsmodells, das für jede Variable eigens konstruiert wird (Schritt 3). Um die gemeinsame Verteilung der Variablen mit beobachteten und fehlenden (wahren) Werten möglichst zu erhalten, enthält jedes Regressionsmodell ein umfassendes Set unabhängiger *beobachteter* Variablen.

Darüber hinaus ist das Verfahren *multivariat*, da die Schätzung der fehlenden Werte wiederholt ( $t$ -mal) durchlaufen wird, wobei die Variablen, die in jeder Regression konditioniert werden, durch die beobachteten oder aktuell imputierten Werte ersetzt werden (Schritt 4). Es ist wichtig, dass jedes Regressionsmodell zusätzlich auch ein umfangreiches Set unabhängiger *Missing*-Variablen enthält, um die gemeinsame Verteilung der Variablen mit fehlenden Werten zu erhalten. Wenn  $t$  gegen unendlich geht, wird erwartet, dass die Imputationen im Laufe der Zyklen gegen eine Approximation einer Ziehung aus der gemeinsamen a-posteriori-prädiktiven Verteilung (Joint Posterior Predictive Distribution) der fehlenden Werte von  $Y_1, Y_2, \dots, Y_p$  konvergieren.

Im letzten Schritt des Verfahrens (Schritt 5) wird jeder fehlende Wert multipel imputiert, indem Schritt 3 und 4 unabhängig voneinander  $M$  mal wiederholt werden. Dies trägt der statistischen Unsicherheit imputierter Werte bei der Schätzung von Varianzen mit imputierten unvollständigen Variablen Rechnung. Die  $M$  Imputationen der fehlenden Werte von  $Y_1, Y_2, \dots, Y_p$  konvergieren unseren Erwartungen zufolge gegen eine Approximation von  $M$  Ziehungen aus der gemeinsamen a-posteriori-prädiktiven Verteilung der fehlenden Werte.

Wenn es auch theoretisch denkbar wäre, dass die Folge der Ziehungen auf Grundlage der oben abgebildeten Regressionen nicht gegen eine stationäre prädiktive Verteilung konvergiert, so haben Simulationsstudien den Nachweis geliefert, dass dieser Ansatz unverzerrte Schätzwerte liefert (siehe Van Buuren et al., 2006). Außerdem spiegelt – im Fall der HFCS-Daten, die eine große Anzahl von Variablen umfassen, welche wiederum zahlreiche Schranken, Filter-Missings, Intervallrückmeldungen, Interaktionen oder Einschränkungen in Bezug auf andere Variablen aufweisen – die Verwendung separater Regressionen für jede Variable die Daten besser wider und erscheint daher sinnvoller als die Spezifikation einer gemeinsamen Verteilung für alle Variablen, wie dies beim Joint Modeling-Ansatz der Fall ist.<sup>7</sup>

<sup>7</sup> Ein Überblick zu den verschiedenen Imputationsmethoden findet sich ebenfalls in Little und Rubin (2002).

Schließlich wird darauf hingewiesen, dass das HFCS-Imputationsverfahren auf der Annahme basiert, dass die Non-Response-Wahrscheinlichkeiten von Variablen mit Missing Values nur von beobachteten Informationen abhängen und nie von unbeobachteten, wie z. B. von den Variablen mit fehlenden Werten selbst. Diese Annahme wird in der Literatur Ignorierbarkeitsannahme (Ignorability Assumption) genannt.

Bevor die eben dargestellten fünf Schritte durchlaufen werden können, sind die Daten vorzubereiten und alle Parameter unseres Imputationsmodells zu spezifizieren, z. B. die Wahl der zu imputierenden Variablen, die Imputationsreihenfolge, das Regressionsmodell für jede Variable, die Anzahl der Zyklen  $t$ , die Anzahl der Imputationssamples  $M$  etc. Im nächsten Abschnitt beschreiben wir, wie dabei vorgegangen wurde.

## 5.4 Durchführung der Imputationen

### 5.4.1 Auswahl der zu imputierenden Variablen

Im ersten Schritt des HFCS-Imputationsverfahrens sind die zu imputierenden Variablen  $Y_1, Y_2, \dots, Y_p$  auszuwählen. Unsere Strategie ist es, so viele Variablen mit Missing Values wie möglich zu imputieren (in unserem Fall rund 70 % der Variablen). Die übrigen Variablen mit Missing Values werden nicht mittels des HFCS-Imputationsverfahrens imputiert, da sie entweder nicht ausreichend Varianz aufweisen oder zu wenige Beobachtungen vorliegen, um eine Schätzung mittels Regression zuzulassen.<sup>8</sup>

Die Imputation einer möglichst umfassenden Zahl von Variablen soll die Anzahl der Beobachtungen, bei denen der Datennutzer gezwungen ist, ein fallweises Ausschlussverfahren in Bezug auf HFCS-Daten anzuwenden, weil die Variablen, an denen er interessiert ist, nicht imputiert wurden, auf ein Minimum reduzieren. Ein weiterer wichtiger Grund für diese Strategie ist, dass wir die Korrelationsstruktur der Daten mit unseren Imputationen nicht verzerren wollen. Würden wir auf die Imputation zahlreicher Variablen verzichten, könnten wir diese auch nicht in den Regressionsmodellen als unabhängige Variablen mit Missing Values verwenden und würden die Beziehungen zwischen den nicht imputierten und imputierten Variablen mit Missing Values verzerren.

### 5.4.2 Imputationsreihenfolge

Wie im vorangegangenen Abschnitt zum HFCS-Imputationsverfahren erwähnt, besteht eine der Schwächen des Verfahrens darin, dass nicht theoretisch nachgewiesen werden kann, dass die Folge von auf Grundlage Bayesscher Regressionen gezogenen Prädiktionen gegen eine stationäre prädiktive Verteilung konvergiert. In der Praxis kann allerdings die Auswahl einer bestimmten Reihenfolge von  $Y_1, Y_2, \dots, Y_p$  häufig zu Konvergenz beitragen. Daher ordnen wir die zu imputierenden Variablen gemäß dem Ausmaß ihrer Unvollständigkeit (Missingness), d. h., wir beginnen die Imputation bei jenen Variablen, die die wenigsten fehlenden Werte aufweisen, und beenden sie bei den Variablen mit den meisten fehlenden

<sup>8</sup> Es wurde ein sehr geringer Anteil der Variablen, die nicht mittels HFCS-Imputationsverfahren imputiert werden können, auf Grundlage von Ad-hoc-Methoden wie dem Hotdeck-Verfahren nach Abschluss des HFCS-Verfahrens imputiert. Der Grund dafür ist, dass deren Imputation als sehr wichtig erachtet wird, da sie z. B. zur Konstruktion wichtiger aggregierter Variablen (wie gesamtes Haushaltseinkommen) verwendet werden.

Werten. Variablen, die im selben Ausmaß Missingness aufweisen, werden in einer zufälligen Abfolge imputiert, wobei jedoch diese Abfolge dann immer gleich bleibt. Die Imputation der Head-Variablen erfolgt immer vor jener der entsprechenden Branch-Variablen. So wird die Antwort auf die Frage, ob der Haushalt einen hypothekarisch besicherten Kredit offen hat, immer vor dem Betrag der Hypothek imputiert, auch wenn beide Variablen denselben Grad an Missingness aufweisen würden.

### 5.4.3 Arten von Regressionsmodellen

In einem dritten Schritt wurde für jede zu imputierende Variable ein Regressionsmodell definiert. Abhängig vom jeweiligen Variablentyp wählen wir zwischen vier verschiedenen Arten von Regressionsmodellen aus. Für stetige Variablen verwenden wir ein Intervallregressionsmodell<sup>9</sup>, da alle unsere stetigen Variablen nach oben und/oder unten hin beschränkt sind (nähere Details dazu finden sich im Abschnitt 5.4.6). Für binäre Variablen verwenden wir ein Logit-Modell, für Ordinal- und Nominalvariablen greifen wir auf geordnete Logit- und multinomiale Logit-Modelle zurück.

### 5.4.4 Verwendung von Gewichten bei Regressionen

Die Notwendigkeit zur Verwendung von Gewichten zur Schätzung deskriptiver Parameter (Mittelwerte, Proportionen, Gesamtwerte etc.) ist im Allgemeinen unumstritten. Die Verwendung von Gewichten bei der Schätzung von Regressionsmodellen auf der Grundlage von Erhebungsdaten ist hingegen umstritten. Diese konkrete Frage stellt sich auch in Bezug auf die Schätzung der Regressionen in Schritt 3 des HFCS-Imputationsverfahrens. Wir haben uns aus den folgenden Gründen für die Verwendung von Gewichten entschieden: Wenn die Regressionsmodelle fehlspezifiziert sind, ist eine gewichtete Regression einer ungewichteten vorzuziehen, da Erstere unverzerrte Schätzer der Regressionskoeffizienten liefert. Nur bei korrekter Spezifikation von Regressionsmodellen kann von gewichteter Regression abgesehen werden, da diese zu erhöhten Standardfehlern führen würde. Da wir den datengenerierenden Prozess nicht kennen, verwenden wir sicherheitshalber Gewichte, um (definitiv) an unverzerrte Imputationen zu gelangen, auch wenn dies (potenziell) zu leicht erhöhten Standardfehlern führen könnte. Dies ist vor allem aufgrund der Veröffentlichung des Datensatzes von Bedeutung, da wir davon ausgehen, dass die meisten Datennutzer mit den imputierten Datensätzen arbeiten werden.<sup>10</sup>

<sup>9</sup> Das Intervallregressionsmodell stellt eine generalisierte Version des Tobit-Modells dar. So wird dem Umstand Rechnung getragen, dass die Daten nach unten und oben hin zensiert sind. Siehe Cameron und Trivedi (2005) für weiterführende Details.

<sup>10</sup> Eine weitere Möglichkeit, zu einer Entscheidung für bzw. gegen die Verwendung von Gewichten zu gelangen, besteht darin, eine Imputation mit Gewichten und ohne Gewichte durchzuführen und beide Modelle zu vergleichen. Ergeben die beiden Ansätze signifikante Unterschiede hinsichtlich der Parameterschätzungen, so spricht das für die Verwendung von Gewichten, damit zumindest unverzerrte Schätzwerte sichergestellt sind. Zeigen sich hingegen keine relevanten Abweichungen bei den Parameterschätzungen, aber große Unterschiede hinsichtlich der Standardfehler, deutet dies auf eine geeignete Spezifikation hin. Die Verwendung ungewichteter Modelle wäre somit unproblematisch. Eine zweimalige Durchführung des HFCS-Imputationsverfahrens – einmal gewichtet, einmal ungewichtet – wäre allerdings ein zeitaufwändiger Prozess, der bislang noch nicht durchlaufen wurde. Dies könnte der Gegenstand künftiger Forschung sein.

### 5.4.5 Variablentransformationen

Vor der Imputation führen wir Transformationen einiger Variablen mit Missing Values durch, da dies nicht nur deren imputierte Werte, sondern auch alle imputierten Werte im Allgemeinen deutlich verbessert. Nach Abschluss der Imputationen wird eine Rücktransformation aller Variablen in ihre ursprüngliche Form vorgenommen.

Eine wichtige Transformation ergibt sich aus der Verwendung des natürlichen Logarithmus bei stetigen Variablen. Diese Arten von Variablen haben üblicherweise eine äußerst schiefe Verteilung; die Verwendung des Logarithmus trägt dazu bei, dass die Verteilung näher bei der Normalverteilungsannahme liegt, die für die Prognose notwendig ist. Eine weitere äußerst hilfreiche Transformation für Jahresvariablen besteht darin, Zeitspannen anstelle von Jahren zu imputieren. So imputieren wir z. B. statt des Anschaffungsjahres eines Hauses den Zeitraum, der seit dem Hauskauf verstrichen ist. In solchen Fällen wurde die oben erwähnte logarithmische Transformation auf Grundlage der Zeiträume und nicht der Jahre durchgeführt.

Bei kategorialen Variablen können zwei Typen von Transformationen verwendet werden. Erstens kann bei einigen Variablen durch eine Neuordnung von Kategorien eine Transformation von Nominalvariablen in Ordinalvariablen vorgenommen werden. Dies verbessert die Stabilität des Imputationsmodells, da ordinale Regressionsmodelle die Schätzung einer geringeren Anzahl von Parametern erfordern als multinomiale Regressionsmodelle. Zweitens werden Mehrfachnennungsfragen mittels Generierung einer binären Variable für jede Antwortkategorie in mehrere binäre Variablen umgewandelt (1 falls die Kategorie zutrifft, ansonsten 0). Dies ermöglicht die Imputation mehrerer Antwortkategorien bei ein und derselben Frage pro Imputationssample.

Eine Transformation, die sowohl bei stetigen als auch kategorialen Variablen mit Missing Values durchgeführt wird, ist die Teilung der ursprünglichen Variable in Head- und Branch-Variablen, wenn die ursprüngliche Variable ein gewisses Maß an Heterogenität aufweist. Zum Beispiel haben manche Kreditlaufzeitvariablen den Wert  $-4$ , mit der Bedeutung, dass „keine fixe Laufzeit vereinbart“ wurde. Es wäre bei der Imputation einer solchen Kreditlaufzeitvariable nicht sinnvoll, die Regression über diese Beobachtungen zusammen mit Beobachtungen mit einem bestimmten Kreditlaufzeitwert laufen zu lassen. In derartigen Fällen teilen wir die jeweiligen Variablen auf zwei Variablen auf, und zwar eine binäre Head-Variable, die anzeigt, ob ein Kredit eine fixe Laufzeit hat oder nicht (Imputation mittels Logit-Regressionsmodell), und eine stetige Branch-Variable, die im Fall einer fixen Laufzeit diese anzeigt (Imputation mittels Intervallregression).

Eine weitere Transformation, die sowohl bei stetigen als auch kategorialen Variablen mit Missing Values durchgeführt wird, ist die der Personen-IDs<sup>11</sup>. Da zur Vermeidung von verzerrten Imputationen Personenvariablen für jede Personen-ID einzeln modelliert und imputiert werden (siehe Abschnitt 5.4.8), sollte sichergestellt werden, dass Personen mit gleichen IDs relativ homogen sind, wenn sie gemeinsam modelliert werden. Deswegen werden vor den Imputationen die Personen in neue, speziell für die Imputationen geschaffene Personen-IDs

<sup>11</sup> Standardmäßig ist im Datensatz die Person mit der ID=1 der Kompetenzträger; alle weiteren Personen sind dem Alter nach gereiht.



gruppiert. Die Kriterien dafür sind folgende: Erste Personen (mit Personen-ID gleich 1) werden alle Kompetenzträger, die männlich sind, alle Partner von Kompetenzträgern, die Person 2 waren und männlich sind, und alle übrigen Kompetenzträger. Zweite Personen (mit Personen-ID gleich 2) werden alle weiblichen Partner von Kompetenzträgern, die schon vorher Person 2 waren, und werden alle Frauen, die Person 1 waren, bevor ihr männlicher Partner erste Person wurde. Alle restlichen Personen werden nach absteigendem Alter geordnet und nummeriert.

Schließlich wird bei Haushalten mit Landwirten eine spezielle Transformation verwendet, nämlich die der Variablen der Werte der Unternehmen des Haushalts (HD0801 bis HD0803) und die der Variable des Werts des Hauptwohnsitzes (HB0900). Anstatt diese Variablen einzeln zu imputieren, wird in einem ersten Schritt die Summe dieser Variablen imputiert und zusätzlich der Anteil davon in Prozent, der zur Landwirtschaft gehört. Anschließend werden in einem zweiten Schritt die einzelnen Variablen (HD0801 bis HD0803 und HB0900) aus dieser Summe und diesen Anteilen berechnet. Der Grund für diese Transformation ist, dass sie die imputierten Werte deutlich verbessert, da manche Haushalte mit Landwirtschaften den Wert ihres Hauptwohnsitzes nicht getrennt vom Wert ihrer Landwirtschaft angegeben haben, sondern als Summe (siehe Abschnitt 4.6.2.8 für weitere Details).

#### 5.4.6 Schranken

Wie bereits erwähnt, verwenden wir Intervallregressionsmodelle für die Imputation stetiger Variablen in Schritt 3, da diese nach oben und/oder unten hin beschränkt sind. Diese Schranken werden eingesetzt, um die Imputation von Werten zu vermeiden, die entweder nicht definiert sind oder die im Widerspruch zu anderen erhobenen Variablen stehen. Wir unterscheiden zwischen allgemeinen und individuellen Schranken.

Die allgemeinen Schranken sind für alle Haushalte und Personen gleich und werden eingesetzt, um eine Imputation nicht definierter oder höchst unrealistischer Werte zu vermeiden. Beispiele für diese Art von Schranken sind von stetigen oder Zählvariablen (Einkommen, Alter) zu erfüllende Nichtnegativitätsbedingungen. Die untere Schranke für diese Variablen ist für alle Haushalte null. Außerdem wenden wir auf jede stetige Variable die folgende Regel an: Für jeden Haushalt wird die Hälfte des niedrigsten beobachteten Werts der Variable als untere Schranke und das Doppelte des höchsten beobachteten Werts als obere Schranke definiert. Durch die Anwendung dieser Regel soll die Imputation von extremen Outliern vermieden werden, ohne dass es zu einer Verzerrung der Ergebnisse kommt. Weitere Beispiele für allgemeine Schranken finden sich bei Prozentvariablen (z. B. Anteil des Haushalts am Wohnungseigentum), in deren Fall die untere Schranke auf null und die obere auf 100 gesetzt wird, bzw. bei einigen Jahresvariablen (z. B. Jahr des Erwerbs oder der Erbschaft des Hauptwohnsitzes im Eigentum des Haushalts), die nach oben hin mit 2011 beschränkt sind (Jahr, in dem die letzten Interviews der Erhebung durchgeführt wurden).

Im Gegensatz zu den allgemeinen Schranken nehmen individuelle Schranken je nach Haushalt oder Person unterschiedliche Werte an; für gewöhnlich dienen sie der Gewährleistung von Konsistenz gegenüber anderen Variablen desselben Haushalts. Die meisten der beim HFCS angewendeten Schranken fallen in diese

letztere Kategorie. Für die Zwecke der Imputation der Ausgaben für zu Hause verzehrte Lebensmittel definieren wir z. B. die gesamten (vom Haushalt) geschätzten Konsumausgaben als obere Schranke. Umgekehrt wird bei der Imputation der gesamten Konsumausgaben die Summe der Ausgaben für zu Hause und außer Haus konsumierte Lebensmittel, Mahlzeiten und Getränke als untere Schranke festgesetzt. Individuelle Schranken werden auch eingesetzt, wenn Haushalte bei einer Betragsfrage ein (vorgegebenes oder individuelles) Intervall anstelle eines Betrags angeben. Derartige Intervalle werden nach jeder nicht beantworteten Betragsfrage abgefragt und erweisen sich für die Imputation als sehr nützlich, da sie wertvolle und präzise Informationen über den fehlenden Wert in der Betragsfrage liefern (siehe auch Abschnitt 5.2 im Zusammenhang mit Tabelle 5).

Individuelle Schranken werden im Rahmen des HFCS z. B. auch bei Imputation von Mietzahlungen (wobei die Warmmiete als obere Schranke für die Kaltmiete definiert wird und Letztere wiederum als untere Schranke für die Warmmiete), Krediten (z. B. ursprüngliche Höhe des Kredits als obere Schranke für den noch ausstehenden Kapitalbetrag und vice versa) oder bei Imputation mehrerer Zählvariablen (z. B. Geburtsjahr des ältesten Haushaltsmitglieds als untere Schranke für das Jahr des Erwerbs des Hauptwohnsitzes) angewendet. Im Fall von Beobachtungen, auf die mehrere untere Schranken bzw. mehrere obere Schranken zutreffen (z. B. allgemeine und individuelle Schranken), wählen wir die jeweils restriktivste untere bzw. obere Schranke.

#### 5.4.7 Prädiktorauswahl

Wie bereits erwähnt, ist eines der Hauptziele der Imputation die gemeinsame Verteilung zwischen unvollständig und vollständig beobachteten Variablen sowie auch zwischen den Variablen mit Missing Values untereinander zu erhalten. Daher reicht es bei der Auswahl der Prädiktoren für das Imputationsmodell nicht aus, gute Prädiktoren für jede zu imputierende Variable zu wählen. Eine derartige Vorgehensweise könnte die Korrelationsstruktur zwischen der zu imputierenden Variable und den ausgeschlossenen Variablen verzerren. Außerdem würde die Ignorierbarkeitsannahme, auf der unser Imputationsmodell beruht (siehe Abschnitt 5.3), weniger plausibel erscheinen, wenn wir Variablen außer Acht lassen, die den Antwortausfall der zu imputierenden Variablen bestimmen.

Deshalb wählen wir eine möglichst große Anzahl an Prädiktoren (Broad Conditioning Approach). Bei einem großen Datensatz, wie im Fall des HFCS mit mehreren hundert Variablen, können aber nicht alle mit eingeschlossen werden, denn einerseits würden sich daraus Multikollinearitätsprobleme und andererseits rechnerische Schwierigkeiten ergeben. Ähnlich wie Van Buuren et al. (1999) bzw. Barceló (2006) verwenden wir daher die folgende Strategie zur Auswahl von Prädiktorvariablen:

1. Einschluss jener Variablen, die Determinanten des Antwortausfalls der zu imputierenden Variable sind. Diese sind zur Erfüllung der unserem Imputationsmodell zugrunde liegenden Ignorierbarkeitsannahme erforderlich (siehe Abschnitt 5.3). Typische Determinanten des Antwortausfalls, die wir verwendet haben, sind z. B.: Variablen zur Beschreibung des Haushalts (geschätztes Haushaltseinkommen, Haushaltsgröße, Anzahl der Kinder), Variablen zur Beschreibung der Haushaltsmitglieder (Alter, Ausbildung, Geschlecht und Beschäftigungsstatus der ersten Person sowie des Partners bzw. der Partnerin

der ersten Person), Stratifizierungsvariablen (Bundesland, Ortsgröße), von den Interviewern angeführte Informationen (Lebensstandard, Lage des Wohnsitzes, Art und Zustand des Gebäudes, Atmosphäre des Interviews etc.). Letztere Informationen (Paradaten) waren für die Imputationen äußerst wichtig, da sie den Antwortausfall bei vielen Variablen gut erklären konnten.

2. Darüber hinaus sind solche Variablen einzuschließen, die gut geeignet sind, die relevante zu imputierende Variable zu prognostizieren und zu erklären. Dies ist das klassische Kriterium für Prädiktoren und trägt dazu bei, die statistische Unsicherheit der Imputationen zu senken. Diese Prädiktoren werden aufgrund ihrer Korrelation mit der zu imputierenden Variable identifiziert. Bei den ausstehenden Kapitalbeträgen im Rahmen verschiedener Kreditarten verwenden wir z. B. als Prädiktoren den ursprünglichen Kreditbetrag und die Jahre, die seit Aufnahme des Kredits verstrichen sind, da dies bei den meisten Regressionen ein beträchtliches Ausmaß an Varianz erklärt. Bei einer Imputation des Marktwerts verschiedener Formen von Immobilienvermögen schließen wir üblicherweise dessen Anschaffungswert, die Zeitspanne (in Jahren), in der sich das betreffende Vermögen bereits im Eigentum des Haushalts befindet, und den Gesamtwert der Immobilien im Eigentum des Haushalts ein. Bei der Imputation von Kreditvariablen werden (wie oben beschrieben) typischerweise ursprüngliche Kredithöhe, Kreditrückzahlungsbetrag oder ausstehender Kreditbetrag verwendet. Diese Variablen sind oft miteinander in einer gewissen logischen Art verbunden (z. B. ist der ausstehende Kreditbetrag die ursprüngliche Kredithöhe abzüglich der Summe der Rückzahlungen). Jedoch ist es bei den Imputationen nicht möglich, alle diese logischen Querverknüpfungen zu bewahren, insbesondere, wenn mehrere Variablen imputiert werden.
3. Darüber hinaus entfernen wir im Subsample der fehlenden Beobachtungen der zu imputierenden Variable jene der oben genannten Prädiktor-Variablen, die zu viele Missing Values aufweisen, und ersetzen sie durch vollständigere Prädiktoren dieser Prädiktoren. Als Faustregel kann davon ausgegangen werden, dass Prädiktoren, zu denen Beobachtungen im Ausmaß von weniger als 50% innerhalb des erwähnten Subsamples vorliegen, entfernt und durch vollständigere Prädiktoren ersetzt werden. Dieses Kriterium trägt zu erhöhter Robustheit der Imputationen bei. Üblicherweise handelt es sich bei solchen Prädiktoren von Prädiktoren um essenzielle Haushaltsmerkmale wie Haushaltsgröße, Anzahl der Kinder, Region, Alter, Beschäftigungsstatus und Familienstand der ersten Person.
4. Darüber hinaus sind alle Variablen jener Modelle einzuschließen, die nach der Imputation auf die Daten angewendet werden sollen. Anders gesagt erwägt man zuerst, welche verschiedenen ökonomischen Theorien auf Grundlage der Daten getestet werden könnten, und inkludiert jene Variablen als Prädiktoren, von denen zu erwarten ist, dass sie die zu imputierende Variable gemäß diesen Theorien beeinflussen oder erklären werden. Würde man diese Variablen ausschließen, könnte dies tendenziell zu einer Verzerrung der Ergebnisse potenzieller Datennutzer bei Überprüfung der Hypothese eines bestimmten Modells führen. Die HFCS-Daten bieten etwa detaillierte Informationen über verschiedene Vermögenskomponenten der Haushalte, z. B. Sach- oder Finanzvermögen. Diese Information wird für die Analyse von Vermögens-effekten auf den Konsum verwendet. Daher verwenden wir diese Variablen

sowohl bei der Imputation der Konsumausgaben als auch bei jener der Vermögensvariablen.

Es ist einleuchtend, dass viele Variablen in der Erhebung mehrere der oben genannten Kriterien zur Auswahl von Prädiktoren gleichzeitig erfüllen, wie z. B. Einkommen, Alter oder Bildung der ersten Person.

In allen Regressionsmodellen inkludieren wir auch einen Interaktionsterm und eine Haupteffekt-Dummy-Variable für jede der oben genannten Prädiktorvariablen, die nicht bei allen Haushalten abgefragt wurde, bei denen die zu imputierende Variable abgefragt wurde. Nehmen wir z. B. an, wir möchten eine Imputation der Konsumausgaben des Haushalts unter Verwendung des Hypothekenbetrags als einem unserer Prädiktoren durchführen. Während die Konsumausgaben für alle Haushalte in der Stichprobe erhoben wurden, trifft das auf Hypotheken nicht zu. Würden wir für jene Haushalte, die keinen hypothekarisch besicherten Kredit offen haben, den Hypothekenbetrag einfach nur auf null setzen (entspricht einem Interaktionsterm), würde dies zu verzerrten Schätzwerten führen, da die Information, ob der Haushalt über eine Hypothek verfügt oder nicht, vernachlässigt würde. Diese Information ist somit zusätzlich als Haupteffekt-Dummy-Variable in das Regressionsmodell aufzunehmen. Da in diesem Fall die Frage nach etwaigen Hypotheken selber auch nicht an alle Haushalte gestellt wurde, sondern ausschließlich an Wohnungs-/Hauseigentümer, ist hier sinnvollerweise auch eine Wohnungs-/Hauseigentümer-Dummy-Variable in die Regression einzubeziehen.

Die Anzahl der Prädiktoren ist letztlich durch die Größe des Subsamples, über das die Regression geschätzt wird, beschränkt. Dort, wo die Anzahl der nach der oben genannten Strategie gewählten Prädiktoren die Größe des Subsamples übersteigt, verwenden wir Akaikes Informationskriterium zur Bestimmung der Prädiktoren mit der besten Anpassungsgüte, wobei – soweit möglich – jede der oben genannten vier Prädiktorkategorien in jeder Regressionsgleichung vertreten sein sollte. Üblicherweise entspricht bei kleinen Subsamples die Anzahl der für jedes Regressionsmodell verwendeten Prädiktoren etwa 20% der Anzahl der Beobachtungen für die zu imputierende Variable. Bei größeren Subsamples liegt die Anzahl der Prädiktoren für gewöhnlich zwischen 5% und 10%. Weitere Details zur Spezifikation von Subsamples finden sich im nächsten Abschnitt.

#### **5.4.8 Spezifikation von Subsamples**

Jede Regression in Schritt 3 wird über ein Subsample geschätzt, welches aus allen Haushalten bzw. Personen besteht, denen die jeweilige Frage zu der zu imputierenden Variable gestellt wurde. Wenn ein Haushalt z. B. zwei hypothekarisch besicherte Kredite offen hat und wir den offenen Betrag der zweiten Hypothek imputieren möchten, dann imputieren wir diesen fehlenden Wert mittels Regression über das Subsample der Haushalte, die über mindestens zwei Hypotheken verfügen. Eine Berücksichtigung von Haushalten mit nur einer Hypothek bei Imputation der Beträge von Zweihypotheken würde bedeuten, dass wir systematische Unterschiede zwischen Erst- und Zweihypotheken ignorieren. Beispielsweise würde dabei die Tatsache außer Acht gelassen, dass die erste Hypothek

höher ist als die zweite, da Haushalte Hypotheken nach deren Bedeutung ordnen, was zu einem Bias unserer Schätzwerte führen würde.<sup>12</sup>

Ein weiteres Beispiel ist die Imputation von Personenvariablen. Diese werden auch nur über das Subsample der Personen mit derselben Personen-ID regressiert. Um die Homogenität der Personen mit gleichen IDs zu sichern, werden vor den Imputationen die Personen in neue, speziell für die Imputationen geschaffene Personen-IDs gruppiert (siehe Abschnitt 5.4.5), welche dann die erwähnten Subsamples bilden. Wird auf Einzelfragenbasis imputiert, wie wir dies tun, fällt der Bias sehr gering aus, auch wenn dies zu Lasten der Präzision geht, da die Subsample-Größen dadurch manchmal klein sind.

#### 5.4.9 Anzahl der Zyklen

Im vierten Schritt bestimmt die Anzahl der Zyklen (oder Iterationen)  $t$ , wie oft Schritt 3 wiederholt wird. Wenn  $t$  gegen unendlich geht, sollten die imputierten Werte gegen eine Ziehung aus der gemeinsamen a-posteriori-prädiktiven Verteilung der Variablen mit fehlenden Werten konvergieren. Allerdings stellt sich laut Van Buuren et al. (1999) in der Praxis in diesen Modellen Konvergenz gewöhnlich sehr rasch während der ersten paar Iterationen ein. Angesichts des großen, mit dem HFCS-Imputationsmodell verbundenen rechnerischen Aufwands und nach dem Vorbild anderer ähnlicher Erhebungen wie dem SCF (Kennickell, 1998) und EFF (Barceló, 2006) setzen wir die Anzahl der Zyklen für das HFCS-Imputationsmodell mit  $t=6$  fest.

Im Normalfall überprüfen wir die Konvergenz grafisch, indem wir den Mittelwert der imputierten Werte mit der Iterationszahl  $t$  in Beziehung setzen, wie z. B. in Grafik 4, in der dies für die Girokontovariablen (HD1110) dargestellt ist. Konvergenz gilt als erzielt, sobald das Muster der imputierten Mittelwerte nur mehr zufallsbedingt erscheint. In Grafik 4 scheint dies bald der Fall zu sein: Von der ersten Iteration an lässt die geglättete Kurve der imputierten Mittelwerte des Girokontoguthabens keinen eindeutigen Trend mehr erkennen. Darüber hinaus

Grafik 4

#### Durchschnittliches Girokontoguthaben im Zuge der verschiedenen Iterationen (ungewichtet)



Quelle: HFCS Österreich 2010, OeNB.

<sup>12</sup> Natürlich könnten wir in einem solchen Fall viele Interaktionsterme in unser Modell aufnehmen, um den Bias zu reduzieren; dennoch könnte es noch unbeobachtete Unterschiede zwischen beiden Gruppen geben.

zeigt Grafik 4, dass die Schwankungsbreite der imputierten Mittelwerte sehr klein ist, was ein weiterer Indikator für Konvergenz ist. Natürlich können derartige Überprüfungen (wie jeder andere Check beim Chained Equations-Ansatz) niemals das Vorliegen von Konvergenz bestätigen, sind aber geeignet, Schwächen des Imputationsmodells bzw. andere ungewöhnliche Ergebnisse, die auf Nicht-Konvergenz hindeuten könnten, aufzuzeigen.

#### 5.4.10 Anzahl der Imputationssamples

Im letzten Schritt (Schritt 5) wählen wir die Anzahl  $m = 1, 2, \dots, M$  der Realisationen, die aus der gemeinsamen a-posteriori-prädiktiven Verteilung der fehlenden Daten zu ziehen sind, oder – einfacher ausgedrückt – die Anzahl der durch die multiple Imputation zu generierenden Samples. Wird  $M$  zu niedrig angesetzt, resultiert dies in zu geringen Standardfehlern der Schätzergebnisse und in zu kleinen  $p$ -Werten. Schafer und Olsen (1998) haben aber gezeigt, dass die Effizienzgewinne eines Schätzers nach den ersten paar  $M$ -Imputationssamples rapide nachlassen. Ihnen zufolge sind solide Schlussfolgerungen bereits ab einer Größenordnung von  $M=3$  bis  $M=5$  möglich. In Einklang mit der internationalen Vorgabe der EZB und mit anderen Erhebungen (wie SCF oder EFF) legen wir die Anzahl der Imputationen daher auf  $M=5$  fest.

### 5.5 Ausgewählte Ergebnisse

Nach der Imputation ist der HFCS-Datensatz fünfmal so groß, da er aus  $M=5$  multipel imputierten Samples (auch Implicates genannt) besteht. Tabelle 7 bietet erste Einblicke in die Imputationsergebnisse. So sind die gewichteten Mittelwerte ausgewählter Beträgsvariablen in den multipel imputierten Samples und im ursprünglichen, nicht imputierten Sample dargestellt.

Ein interessantes Ergebnis ist, dass die Mittelwerte der meisten Variablen im Durchschnitt nach der Imputation höher ausfallen als vor der Imputation. Liegen die Imputationswerte nahe den wahren Werten, deutet dies darauf hin, dass Haushalte, die eine Antwort hinsichtlich der relevanten Variablen verweigern, tendenziell höhere (nicht beobachtete) Beträge bei diesen Variablen besitzen. Zum Beispiel liegt der Mittelwert der ersten Schenkung/Erbschaft (ohne Hauptwohnsitz) vor Imputation bei 88.019 EUR. Nach den jeweiligen Imputationen erhöht sich dieser auf 110.526 EUR in  $m=2$ , 94.873 EUR in  $m=3$ , 190.532 EUR in  $m=4$ , bzw. 125.350 EUR in  $m=5$ . In  $m=1$  sinkt der Mittelwert geringfügig auf 87.819 EUR. Das bedeutet, dass die Imputationen den Mittelwert der ersten Schenkung/Erbschaft im Durchschnitt um 38 % von 88.019 EUR auf 121.820 EUR erhöhen, wobei rund die Hälfte der hier imputierten Werte auf Intervallsangaben seitens der Haushalte basiert. Dies deutet darauf hin, dass Haushalte mit höherwertigen Erbschaften eher dazu neigen, die Antwort auf die Frage zu verweigern oder mit Intervallangaben zu antworten als jene mit kleineren Erbschaften. Die substanziellsten Erhöhungen im Vergleich zum nicht imputierten Sample treten bei Imputationen von Finanzvermögen (z. B. Marktwert von Aktien) auf. Auch hier spielen die Intervallsangaben der Haushalte eine wichtige Rolle, da sie wertvolle und oft sehr genaue Informationen für die Imputationen liefern (siehe auch Tabelle 5).

Bei anderen Variablen wiederum ändert sich der Mittelwert nicht wesentlich oder sinkt sogar. Der Mittelwert der Ausgaben für zu Hause verzehrte Lebensmittel ändert sich durch die Imputation aufgrund der geringen Item-Non-

**Mittelwerte für ausgewählte Variablen vor und nach multipler Imputation (gewichtet)**

	Mittelwert vor der Imputation	Mittelwerte der multipl imputierten Samples				
	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
<i>in EUR</i>						
Wert des Hauptwohnsitzes <sup>1</sup>	246.203	261.468	271.337	266.286	275.096	268.015
Durch Hauptwohnsitz besicherte Hypothek 1: ausstehender Kapitalbetrag	55.745	94.427	84.670	47.135	58.619	47.657
Monatliche Miete	363	334	332	335	330	334
Sonstiges Immobilieneigentum 1: Marktwert	231.583	195.953	193.173	265.195	198.880	218.387
Durch sonstige Immobilien besicherte Hypothek 1: ausstehender Kapitalbetrag	68.300	58.141	90.563	70.627	74.631	67.420
Guthaben auf Girokonten	2.406	3.343	3.255	3.130	2.908	3.220
Guthaben auf Sparkonten	21.989	28.230	29.700	33.696	29.781	28.905
Wert börsennotierter Aktien	30.440	23.554	36.887	22.753	28.553	22.573
Bruttoeinkommen aus abhängiger Beschäftigung (Person 1)	25.871	25.075	25.254	26.517	26.230	29.403
Bruttoeinkommen aus der Arbeitslosen- unterstützung (Person 1)	6.263	6.361	6.880	6.300	6.225	6.295
Bruttoeinkommen aus Finanzanlagen	836	800	763	730	771	787
Schenkung/Erbschaft 1: Wert	88.019	82.842	130.673	95.338	90.959	94.404
Ausgaben für Lebensmittel zu Hause	379	381	380	381	380	380

Quelle: HFCS Austria 2010, OeNB.

<sup>1</sup> Hierfür wurde die Variable HB0900 verwendet.

Anmerkungen: Alle Mittelwerte werden über die Beobachtungen „Haushalt verfügt über das Item = ja“ geschätzt. Die Anzahl dieser Beobachtungen kann je nach Imputationssample  $m$  variieren, wenn imputiert wird, ob Haushalte über das betreffende Item verfügen oder nicht.

Response-Quote (siehe Tabelle 5) dieser Variable nicht wesentlich. Der Mittelwert des Bruttoeinkommens aus Finanzanlagen ist nach der Imputation sogar geringer als zuvor. Das zeigt, dass Haushalte, die Fragen in Bezug auf diese Variable unbeantwortet lassen, tendenziell niedrigere Einkünfte aus Finanzvermögen haben.

Nicht zuletzt geht aus Tabelle 7 auch hervor, dass die statistische Unsicherheit von Imputationen je nach Variable stark schwanken kann. Bei einigen Variablen (z. B. mit Hauptwohnsitz besicherte Hypothek 1) zeigen die Mittelwerte eine relativ hohe Varianz zwischen den fünf multipl imputierten Samples, was die Unsicherheit der imputierten Werte widerspiegelt und auf die niedrigere Anzahl von Beobachtungen zu diesen Variablen zurückzuführen ist. Bei anderen Variablen (z. B. Bruttoeinkommen aus der Arbeitslosenunterstützung oder monatliche Miete) weisen die Mittelwerte eine relativ niedrige Varianz zwischen den fünf multipl imputierten Samples auf, was auf eine höhere Präzision der imputierten Werte hindeutet. Hätten wir die Variablen 1-fach statt multipl imputiert – also mit nur einem Imputationssample –, dann wäre die Varianz der Schätzer zu niedrig, da die Unsicherheit hinter den imputierten Werten ignoriert würde und diese wie wahre Werte behandelt würden.

## 5.6 Abschließende Bemerkungen

Wir haben gezeigt, dass Imputation für die Analyse des HFCS-Datensatzes notwendig ist, da sie im Vergleich zum fallweisen Ausschlussverfahren – bei Vorliegen systematischer Unterschiede zwischen vollständigen und unvollständigen

gen Beobachtungen – den Non-Response-Bias von Schätzergebnissen reduziert. Imputation verringert ebenfalls den Informationsverlust bei Analysen, da keine Beobachtungen gelöscht werden müssen. Mithilfe eines multiplen Imputationsverfahrens, in dem alle zu imputierenden Variablen in Regressionsgleichungen geschätzt werden (Chained Equations), haben wir fünf multipel imputierte Samples generiert. Informationen zur korrekten Analyse multipel imputierter Daten in Stata<sup>®</sup> finden sich im HFCS-User Guide (siehe Kapitel 9).