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.[1] 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 5 zeigt ausgewählte Informationen zur Item-Non-Response. Im Durchschnitt weist jeder Haushalt 29,9 Missing Values auf, was einen Antwortausfall bei lediglich etwa 2,1 % der insgesamt abgefragten Variablen darstellt. Bei den Betragsvariablen beträgt der betroffene Prozentsatz allerdings 4,7 %. Dies dokumentiert,

[1] *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 5

Item-Non-Response je Haushalt (ungewichtet)

	Mittelwert	Median	Minimum	Maximum
Anzahl der abgefragten Variablen				
Alle Variablen	1.392,0	1.391,0	1.109,0	1.889,0
Betragsvariablen	63,0	64,0	36,0	106,0
Anzahl der Variablen mit Missing Values				
Alle Variablen	29,9	18,0	0,0	487,0
Betragsvariablen	3,0	2,0	0,0	36,0
Anteil der Variablen mit Missing Values in %				
Alle Variablen	2,1	1,3	0,0	32,0
Betragsvariablen	4,7	3,0	0,0	49,2

Quelle: HFCS Austria 2014, 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.[2] 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 6 Item-Non-Response-Quoten für ausgewählte Variablen dargestellt.

Anhand von Tabelle 6 lässt sich z. B. erkennen, dass – nach dem Wert ihres Hauptwohnsitzes gefragt – 77,1 % der Haushalte einen Betrag nennen (Spalte 3). Bei den übrigen 22,9 % der Haushalte kam es zu einem partiellen Antwortausfall, d. h., dass sie entweder ein (vorgegebenes oder individuelles) Intervall nannten (19,4 %, Spalte 4), mit „Weiß nicht“ antworteten bzw. „Keine Angabe“ machten (3,2 %, Spalte 5) oder auf Missing editiert wurden[3] (0,4 %, Spalte 6). Die Non-Response-Quoten[4] fallen je nach abgefragter Position sehr unterschiedlich aus. Eine hohe Non-Response-Quote weisen etwa die Fragen nach dem Wert börsennotierter Aktien (100 % – 51,9 % = 48,1 %) bzw. nach dem Bruttoeinkommen des Haushalts aus Finanzanlagen (100 % – 46,6 % = 53,4 %) auf. Im Hinblick auf letztere Variable geben 33 % der Haushalte zumindest ein Intervall an, in dem das

[2] *Siehe dazu ausführlich Little und Rubin (2002).*

[3] *Nähere Details dazu finden sich in Kapitel 4.*

[4] *Die Non-Response-Quote errechnet sich als 100 % abzüglich des Wertes in der Spalte „Betrag“ in Tabelle 6.*

Tabelle 6

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[1]
	(1)	(2)	(3)	(4)	(5)	(6)
	in %					
Wert des Hauptwohnsitzes[2]	42,9	0,0	77,1	19,4	3,2	0,4
Durch Hauptwohnsitz besicherte Hypothek 1: ausstehender Kapitalbetrag	12,7	0,6	69,4	14,9	15,4	0,3
Monatliche Miete	50,9	0,0	61,7	37,8	0,4	0,1
Sonstiges Immobilieneigentum 1: Marktwert	11,1	0,3	78,4	14,7	6,0	0,9
Durch sonstige Immobilien besicherte Hypothek 1: ausstehender Kapitalbetrag	1,3	0,3	77,5	10,0	12,5	0,0
Guthaben auf Girokonten	99,2	0,0	80,4	9,5	9,5	0,6
Guthaben auf Sparkonten	84,0	1,2	72,5	14,2	9,6	3,7
Wert börsennotierter Aktien	4,5	1,2	51,9	28,9	19,3	0,0
Geldschulden gegenüber dem Haushalt	7,6	0,4	94,3	2,6	3,1	0,0
Beschäftigungsstatus (Hauptbeschäftigung) (Person 1)	100,0	0,0	100,0	0,0	0,0	0,0
Bruttoeinkommen aus abhängiger Beschäftigung (Person 1)	48,7	0,0	85,0	9,9	3,7	1,4
Bruttoeinkommen aus der Arbeitslosenunterstützung (Person 1)	5,5	0,1	89,8	5,4	3,0	1,8
Bruttoeinkommen aus Finanzanlagen	63,0	15,0	46,6	33,0	19,1	1,3
Schenkung/Erbschaft 1: Wert	26,7	1,3	77,8	9,6	9,0	3,5
Ausgaben für Lebensmittel zu Hause	100,0	0,0	98,4	1,4	0,2	0,0

Quelle: HFCS Austria 2014, OeNB.

[1] *Missing Values aufgrund von Editierungsmaßnahmen und dem Ausstieg aus Schleifen.*
[2] *Dafür wurde die Variable HB0900 verwendet.*

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 2.6.2 für den Fragebogen und den Aufbau der Euro-Schleifen). Eine Variable mit niedriger Non-Response-Quote ist z. B. jene zu den Ausgaben für zu Hause verzehrte Lebensmittel (100 % – 98,4 % = 1,6 %).

Aus Tabelle 6, Spalte 2, geht ein weiterer Aspekt der Item-Non-Response im HFCS hervor. 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 (63 %), gelangen überhaupt zur Frage nach der Höhe des Einkommens. Bei den übrigen Haushalten – einschließlich jener 15 %, die die Entscheidungsfrage nicht beantwortet haben – wird die Betragsfrage übersprungen. Im Hinblick auf den Antwortausfall bei der Betragsfrage muss allerdings die Antwortverweigerung jener Haushalte (15 %), 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 7

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

Kovariaten	Koeffizient
Weiblich (Person 1)	–0,0777
	(0,0966)
Alter (Person 1)	–0,00382
	(0,00338)
Hochschulabschluss (Person 1)	–0,00624
	(0,127)
Unselbstständig bzw. selbstständig erwerbstätig (Person 1)	–0,315***
	(0,115)
Wohnhaft in Wien	–0,642***
	(0,132)
Wohnfläche Hauptwohnsitz	0,00225**
	(0,000987)
Haushaltsgröße	0,282***
	(0,0440)
Konstante	–1,750***
	(0,248)
Beobachtungen[1]	2.940

Quelle: HFCS Austria 2014, OeNB.

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

[1] *Die restlichen 57 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 werden 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 7 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, eher in Wien wohnhaft 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 implementiertes Verfahren, in dem alle zu imputierenden Variablen in Regressionsgleichungen geschätzt werden (Chained Equations).[5] Das Verfahren[6] kann in die folgenden Schritte unterteilt werden:

- Schritt 1: Auswahl der zu imputierenden P Variablen $Y_1, Y_2, ..., Y_P$
- Schritt 2: Ersetzen der fehlenden Werte von $Y_1, Y_2, ..., Y_P$ mit zufälligen Ziehungen aus tatsächlich beobachteten Werten
- Schritt 3: Für $Y_1, Y_2, ..., Y_P$
  - 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,

[5] *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.*

[6] *Zu weiterführenden technischen Details siehe Albacete (2014), in dem das Imputationsverfahren der Immobilienvermögenserhebung genauer beschrieben wird, das mit 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: Schritte 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 $Y_1, Y_2, ..., Y_P$ konvergieren.

Im letzten Schritt des Verfahrens (Schritt 5) wird jeder fehlende Wert multipel imputiert, indem Schritte 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 $Y_1, Y_2, ..., Y_P$ konvergieren im Erwartungswert 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 etwa beim Joint-Modeling-Ansatz der Fall ist.[7]

[7] *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 $Y_1, Y_2, ..., Y_P$ 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 da zu wenige Beobachtungen vorliegen, um eine Schätzung mittels Regression zuzulassen.[8]

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 $Y_1, Y_2, ..., Y_P$ 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

[8] *Es wurde ein sehr geringer Anteil der Variablen, die nicht mithilfe des HFCS-Imputationsverfahrens 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 z. B. das gesamte 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 aus vier verschiedenen Arten von Regressionsmodellen aus. Für stetige Variablen verwenden wir ein Intervallregressionsmodell[9], 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 und multinomiale Logit-Modelle zurück.[10]

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 in der zweiten Welle entschieden, Gewichte lediglich als Prädiktoren (siehe Abschnitt 5.4.7), aber nicht für gewichtete Regressionen zu verwenden, da dies auch der aktuelle Trend in der Imputationsliteratur ist (siehe z. B. Frumento et al., 2012). Als Grund wird in der Literatur genannt, dass multiple Imputationen lediglich dazu da sind, fehlende Werte (und deren Unsicherheit) gut zu prognostizieren. Eine Gewichtung der Einheiten sollte erst zu einem späteren Zeitpunkt stattfinden, wenn anhand einer Analyse des finalen Datensatzes Aussagen über die Population getroffen werden sollen.

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.

[9] *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.*

[10] *Einzige Ausnahmen sind die Nominalvariablen zur 3-stelligen ISCO-Klassifikation von Berufsbezeichnungen und zur 3-stelligen NACE-Klassifikation von Betriebsaktivitäten, die durch ihre sehr hohe Anzahl an Kategorien (74 bzw. 121) schwer mit einem multinomialen Logit-Modell zu schätzen waren. In diesen zwei Fällen wurde jeweils das sogenannte Predictive-Mean-Matching-Verfahren angewendet, bei dem in einem ersten Schritt anhand einer linearen Regression Werte für die Missing Values prognostiziert werden und schließlich in einem zweiten Schritt für jeden Missing Value jener beobachtete Wert imputiert wird, der den prognostizierten am nächsten kommt.*

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.

Eine weitere Transformation, die bei manchen Variablen mit Werten zwischen 0 und 1 verwendet wird, ist die log-odds-Transformation *(log(y/(1–y)))* – z. B. bei der ausstehenden Konsumkredithöhe (HC0801 bis HC0803). Anstatt diese Variablen einzeln zu imputieren, wird in einem ersten Schritt zunächst die ursprüngliche Kredithöhe (HC0601 bis HC0603) imputiert. Ebenso wird ein Indikator, ob die ausstehende Kredithöhe kleiner ist als die ursprüngliche, und – falls ja – der Anteil der ausstehenden an der ursprünglichen Kredithöhe in Prozent imputiert. Dieser Anteil wird als log-odds-Transformation imputiert, welche die Qualität der imputierten Werte deutlich verbessert. Anschließend werden in einem zweiten Schritt die einzelnen Variablen (HC0801 bis HC0803) aus den entsprechenden ursprünglichen Kredithöhen und Anteilen berechnet.

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.[11] Da

[11] *Standardmäßig ist im Datensatz die Person mit der ID=1 der Kompetenzträger; alle weiteren Personen sind dem Alter nach gereiht.*

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 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 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 Wertes des Hauptwohnsitzes (HB0900). Anstatt diese Variablen einzeln zu imputieren, wird in einem ersten Schritt die Summe dieser Variablen imputiert und zusätzlich der Anteil davon, der zur Landwirtschaft gehört, in Prozent. 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 4.6.2.7 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. Bei manchen stetigen Variablen wird angenommen, dass ein Wert unter oder über einer bestimmten allgemeinen Schranke in der Praxis nicht vorkommen kann. So ist z. B. die untere Schranke für das Jahr der Kreditaufnahme (HB1301 bis HB1303) 1945. Es wird angenommen, dass es in Österreich keinen Kredit gibt, bei dem die Aufnahme/Neuverhandlung/Refinanzierung länger als 70 Jahre zurückliegt. Durch die Anwendung solcher „empirischer" Schranken soll die Imputation von extremen Outliern in diesen Variablen 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), für die 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 2015 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 6).

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) 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 ließen, 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) sind 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 Marktwertes 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 die ursprüngliche Kredithöhe, der Kreditrückzahlungsbetrag oder der ausstehende 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 dieser Variablen gleichzeitig imputiert werden.
3. Darüber hinaus entfernen wir im Subsample der fehlenden Beobachtungen der zu imputierenden Variable jene der oben genannten Prädiktorvariablen, 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 dieser 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 Informationen werden für die Analyse von Vermögenseffekten 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 die finalen Survey-Gewichte (siehe Diskussion in Abschnitt 5.4.4) und einen Interaktionsterm sowie für jede der oben genannten Prädiktorvariablen eine Haupteffekt-Dummy-Variable, 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 die Anzahl der für jedes Regressionsmodell verwendeten Prädiktoren etwa 20 % der Anzahl der Beobachtungen für die zu imputierende Variable. 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, das 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 Regres-

sion über das Subsample der Haushalte, die über mindestens zwei Hypotheken verfügen. Eine Berücksichtigung von Haushalten mit nur einer Hypothek bei Imputation der Beträge von Zweithypotheken würde bedeuten, dass wir systematische Unterschiede zwischen Erst- und Zweithypotheken 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.[12]

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 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 bei diesen Modellen Konvergenz gewöhnlich sehr rasch während der ersten paar Zyklen ein. Angesichts des großen, mit dem HFCS-Imputationsmodell verbundenen rechnerischen Aufwands setzen wir die Anzahl der Zyklen für das HFCS-Imputationsmodell mit $t = 10$ fest. Andere ähnliche Erhebungen, wie der SCF (Kennickell, 1998) und der EFF (Barceló, 2006), verwenden sogar nur $t = 6$.

Im Normalfall überprüfen wir die Konvergenz grafisch, indem wir den Mittelwert der imputierten Werte mit der Zyklenzahl t in Beziehung setzen. Konvergenz gilt als erzielt, sobald das Muster der imputierten Mittelwerte nur mehr zufallsbedingt erscheint und kein eindeutiger Trend mehr erkennbar ist.

Zusätzlich überprüfen wir in der zweiten Welle des HFCS die Konvergenz für ausgewählte Variablen auch anhand des Gelman-Rubin-Kriteriums, das in der Literatur sehr oft verwendet wird (siehe für mehr Details z. B. Cowles und Carlin, 1996). Demnach ist Konvergenz bei einer Variable erreicht, wenn die Varianz eines Schätzers dieser Variable (z. B. Mittelwert, Median oder andere Perzentilwerte) zwischen den verschiedenen multiplen Imputationssamples relativ klein im Vergleich zur Varianz desselben Schätzers zwischen den verschiedenen Zyklen ist.[13] Dies wird in der zweiten Welle des HFCS bei allen überprüften Variablen erfüllt.[14]

[12] *Auch wenn wir in einem solchen Fall viele Interaktionsterme in unser Modell aufnehmen könnten, um den Bias zu reduzieren, könnte es dennoch unbeobachtete Unterschiede zwischen beiden Gruppen geben.*

[13] *Das Gelman-Rubin-Kriterium ist gleich der Wurzel von [(t−1)/t) + (BV/WV)], wo BV und WV die Between- und Within-Varianz sind. Wenn die Gelman-Rubin-Werte unter 1,2 bis 1,1 sind, dann spricht man üblicherweise von Konvergenz.*

[14] *Überprüft wurden folgende Variablen: HB0900, HD1110, HD1210, HD1510, HB1701, HB2801, HB4400, HI0100, HI0200 und HI0310.*

Natürlich können derartige Überprüfungen (wie jeder andere Check beim Chained-Equations-Ansatz) niemals das Vorliegen von Konvergenz bestätigen (siehe Abschnitt 5.3). Sie 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 $m = 1, 2, ..., M$ 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 8 bietet erste Einblicke in die Imputationsergebnisse. So sind die gewichteten Mittelwerte ausgewählter Betragsvariablen 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 87.202 EUR. Nach den jeweiligen Imputationen erhöht sich dieser auf 92.620 EUR in $m = 1$, 91.502 EUR in $m = 2$, 92.076 EUR in $m = 3$, 100.621 EUR in $m = 4$ bzw. 97.088 EUR in $m = 5$. Das bedeutet, dass die Imputationen den Mittelwert der ersten Schenkung/Erbschaft im Durchschnitt um 9 % von 87.202 EUR auf 94.781 EUR erhöhen, wobei rund ein Drittel der hier imputierten Werte auf Intervallangaben 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 Sparkontoguthaben und Hypothekarkredit beim Hauptwohnsitz auf. Auch hier spielen die Intervallangaben der Haushalte eine wichtige Rolle, da sie wertvolle und oft sehr genaue Informationen für die Imputationen liefern (siehe auch Tabelle 6).

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-Response-Quote (siehe Tabelle 6) dieser Variable nicht wesentlich. Der Mittelwert des

Tabelle 8

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

	Mittelwert vor der Imputation	Mittelwerte der multipel imputierten Samples				
	m = 0	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 5*
	in EUR					
Wert des Hauptwohnsitzes[1]	285.996	290.833	290.995	292.706	290.210	292.890
Durch Hauptwohnsitz besicherte Hypothek 1: ausstehender Kapitalbetrag	73.205	80.468	86.705	81.603	81.151	85.650
Monatliche Miete	407	393	399	401	396	391
Sonstiges Immobilieneigentum 1: Marktwert	249.384	237.947	248.696	258.458	233.246	246.517
Durch sonstige Immobilien besicherte Hypothek 1: ausstehender Kapitalbetrag	78.480	81.357	70.089	74.470	67.089	74.713
Guthaben auf Girokonten	2.623	2.689	2.695	2.624	2.612	2.528
Guthaben auf Sparkonten	23.201	26.925	27.293	26.375	26.526	27.389
Wert börsennotierter Aktien	27.584	26.222	32.490	25.038	26.693	31.007
Bruttoeinkommen aus abhängiger Beschäftigung (Person 1)	27.319	27.677	27.587	27.695	27.509	27.560
Bruttoeinkommen aus der Arbeitslosen-unterstützung (Person 1)	6.437	6.504	6.502	6.363	6.482	6.664
Bruttoeinkommen aus Finanzanlagen	706	523	564	553	596	587
Schenkung/Erbschaft 1: Wert	87.202	92.620	91.502	92.076	100.621	97.088
Ausgaben für Lebensmittel zu Hause	373	374	373	373	373	373

Quelle: HFCS Austria 2014, OeNB.

[1] Dafür wurde die Variable HB0900 verwendet.

Anmerkung: 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.*

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 8 auch hervor, dass die statistische Unsicherheit von Imputationen je nach Variable stark schwanken kann. Bei einigen Variablen (z. B. Sonstiges Immobilieneigentum 1) zeigen die Mittelwerte eine relativ hohe Varianz zwischen den fünf multipel 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 multipel imputierten Samples auf, was auf eine höhere Präzision der imputierten Werte hindeutet. Hätten wir die Variablen einfach statt multipel imputiert – also mit nur einem Imputationssample –, dann wäre die Varianz der Schätzer zu niedrig, da die Unsicherheit hinter den imputierten Werten ignoriert werden würde und diese wie wahre Werte behandelt werden 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 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 finden sich im HFCS-User Guide (siehe Kapitel 9).