Imputationen

Nicolás Albacete, Pirmin Fessler, Peter Lindner

(Abteilung für volkswirtschaftliche Analysen, OeNB)

Technischer Workshop zum HFCS in Östereich

25. und 26.2.2013

Outline

1. Partieller Antwortausfall im HFCS
2. Methoden zur Analyse von Daten mit missing values
   - Complete Case Analyse
   - 1-Fache Imputation
   - Multiple Imputation
3. Imputationen im HFCS
   - Imputationsmodell
   - Ausgewählte Ergebnisse
4. Auswertung des multipel imputierten Datensatzes
   - Prinzip
   - Rubin Rules
   - Beispiel

Partieller Antwortausfall im HFCS ↪ pro Haushalt

Item-Non-Response (ungewichtet) je Haushalt

	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	0,0	39,5
Euro-Variablen	6,9	4,2	0,0	78,8

Quelle: HFCS Austria 2010, OeNB.

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

Partieller Antwortausfall im HFCS ↪ pro Variable

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"	Editierungen[1]
	in %					
Wert des Hauptwohnsitzes[2]	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
Unternehmen 1: Wert des Unternehmens[1]	5,8	0,1	45,3	21,9	26,3	6,6
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
Freiwillige private Altersvorsorge – Vermögensstand (Person 1)	13,1	0,8	49,5	19,0	29,9	1,6
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.

[1] *Die letzte Spalte enthält Missing Values aufgrund von Editierungsmaßnahmen und dem Ausbruch aus einer Schleife.*

[2] *Hierfür wurde die Variable HB0900 verwendet.*

Partieller Antwortausfall im HFCS ↪ Ländervergleich

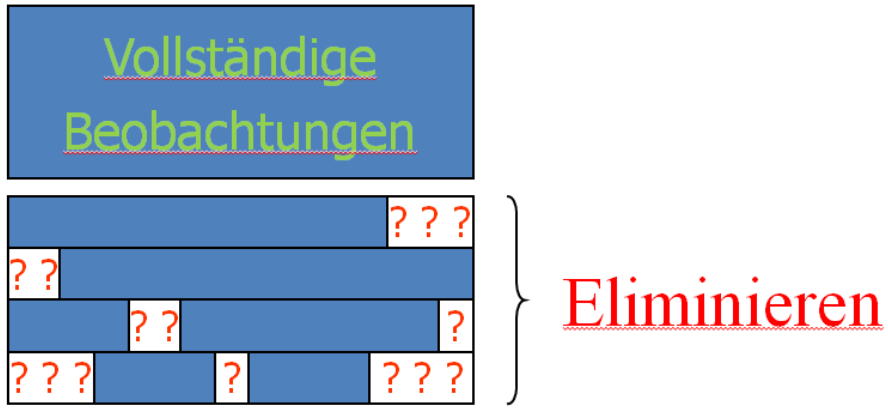
Table 6.3 Item non-response rates: Current value of household main residence

Country	% having item		Of those having item*			Conditional mean (EUR)	
	Reported having item	Imputed as having item	Collected	Imputed from ranges	Imputed from missing	All	Collected#
Belgium	74.0	0.1	92.4	4.6	2.4	273,100	272,800
Germany	56.0	0.4	89.9	5.3	3.6	205,800	206,200
Greece	66.8	0.0	91.0	5.6	3.3	123,400	124,100
Spain	86.9	0.0	90.9	4.9	4.3	211,100	212,300
France	66.7	0.0	0.0	80.9	19.1	222,200	230,200
Italy	71.2	0.0	99.4	0.0	0.0		-
Cyprus	80.0	0.0	81.8	0.0	17.5	317,500	334,700
Luxembourg	70.0	0.0	88.3	8.9	2.9	611,900	611,500
Malta	76.3	0.0	67.5	30.3	0.2		-
Netherlands	74.1	0.0	94.5	0.0	5.5	270,600	269,900
Austria	48.4	1.2	74.8	15.3	9.1	258,100	258,600
Portugal	69.4	0.0	90.0	0.0	7.0	113,800	115,700
Slovenia	82.2	0.9	82.1	7.4	10.5	126,500	128,600
Slovakia	77.3	0.0	81.1	14.6	4.1	68,700	69,200
Finland	77.0	0.0	All values estimated				-

* In addition to collected and imputed values, observations can be edited or estimated, which is why the columns do not always add up to 100%.

Includes observations edited, estimated or collected as range values and then imputed. Provided only for countries with >15 imputed cases.

Complete Case Analyse ↪ Prinzip



- Standardmäßige Analyse bei Statistik-Software
- Einfach aber beschränkt

Complete Case Analyse $\hookrightarrow$ Nachteile

Informationsverlust der unvollständigen Beobachtungen hat zwei Aspekte:

- Erhöhte Varianz der Schätzer
- Bias falls vollständige Beobachtungen systematisch unterschiedlich sind von unvollständigen Beobachtungen
  - Beschränkung auf vollständige Beobachtungen impliziert die Annahme, dass diese Beobachtungen repräsentativ für alle sind
  - Diese Annahme ist oft bedenklich!

Complete Case Analyse $\hookrightarrow$ Antwortausfall nicht zufällig

Logit-Regression des Antwortausfalls bei Betragsfrage zu Girokontoguthaben

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

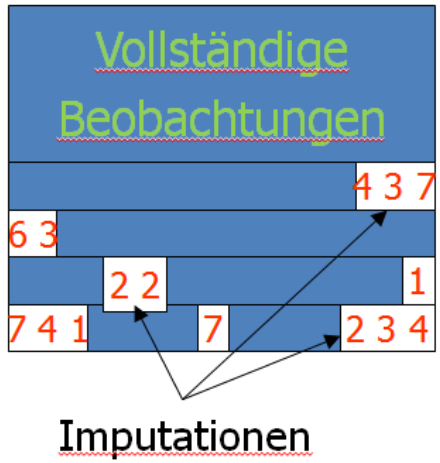
Kovariate	Koeffizient
Weiblich (Person 1)	0,0775
	(0,0950)
Alter (Person 1)	–0,0012
	(0,00344)
Hochschulabschluss (Person 1)	–0,259*
	(0,156)
Unselbstständig bzw. selbstständig erwerbstätig (Person 1)	–0,195*
	(0,113)
Wohnhaft in Wien	–0,194
	(0,134)
Wohnfläche Hauptwohnsitz	0,00274***
	(0,000863)
Haushaltsgröße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

Quelle: HFCS Austria 2010, OeNB.

Anmerkung: Standardfehler in Klammern.

**** p<0,01, ** p<0,05, * p<0,1*

1-Fache Imputation ↪ Prinzip



Vorteile

- Rechteckiger Datensatz
- Behält beobachtete Werte
- Einmalige Behandlung fehlender Werte
- Verwertet fehlende Beobachtungen

Nachteile

- Naive Imputationsmethoden können schlecht sein
- Erfindet Daten – unterschätzt Unsicherheit

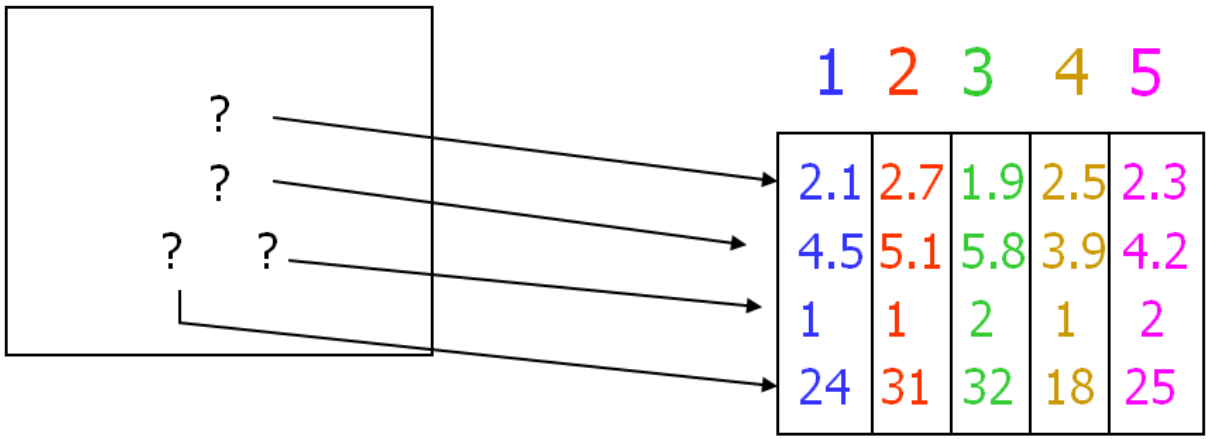
1-Fache Imputation ↪ Zu den Nachteilen

- Laut Little und Rubin (2002) sollten Imputationen im Allgemeinen:
  - Konditional auf Variablen mit vollständigen Beobachtungen sein
  - Multivariat sein, um Assoziationen zwischen Variablen mit unvollständigen Beobachtungen zu bewahren
  - Zufällig gezogen werden, statt Mittelwerte zu sein
- Problem bei 1-fachen Imputationen: Unsicherheit der Imputationen werden in den Standardfehlern nicht berücksichtigt
  - Multiple Imputation

Multiple Imputation ↪ Prinzip

Erstelle M Sätze von Imputationen, jeder Satz ist gezogen aus der Prognoseverteilung der fehlenden Werte

- Z.B. M=5



Multiple Imputation ↪ Vorteile

- Behält Vorteile der 1-fachen Imputation
  - Konsistente Analysen
  - Enthält Wissen des Datenerhebers
  - Vollständige Datensätze
- Korrigiert Nachteile der 1-fachen Imputation
  - Spiegelt Unsicherheit der imputierten Werte wider
  - Korrigiert Ineffizienz der Imputation von gezogenen Werten
  - Schätzer haben hohe Effizienz bei bereits niedrigem M, z.B. 5

HFCS-Imputationsmodell $\hookrightarrow$ Algorithmus

- Arten von Variablen
  - ▷ Stetig (Intervallregressionsmodell)
  - ▷ Binär (Logit-Modell)
  - ▷ Ordinal oder Nominal (geordnete oder multinomiale Logit-Modelle)

- Vollständig beobachtete Variablen: U
- $Y_1, Y_2, ..., Y_p$ zu imputierende Variablen, wobei Y_1 die kleinste Anzahl an fehlenden Werten hat und Y_p die größte

Iteration 1

- Ersetze fehlende Werte von $Y_1, Y_2, ..., Y_p$ mit zufälligen Ziehungen aus tatsächlich beobachteten Werten (Startwerte)
- Regressiere Y_1 auf ein umfangreiches Set unabhängiger Variablen aus U und $Y_2, ..., Y_p$; Imputiere fehlende Y_1-Werte anhand Bayesschen Regressionsmodells (gezogen)

- Regressiere Y_2 auf ein umfangreiches Set unabhängiger Variablen aus U und $Y_1, Y_3, Y_4, ..., Y_p$; Imputiere fehlende Y_2-Werte anhand Bayesschen Regressionsmodells (gezogen)
- Usw.

Iteration 2,3,...

- Wiederhole die Schritte der ersten Iteration t mal

- Dabei sind die jeweils zuvor verwendeten imputierten Werte durch aktualisierte, aus der jeweils letzten Regression gewonnene zu ersetzen
- Auf dieser Grundlage wird der erste Satz von Imputationen erstellt

HFCS-Imputationsmodell $\hookrightarrow$ Weitere Spezifikationen

- Einschränkungen des Samples
- Jedes Regressionsmodell wird nur über relevantes Subsample geschätzt
- Variablentransformationen
- Schranken
  - Ziehung aus der Prognoseverteilung einer generalisierten Version des Tobit-Regressionsmodells (Zensierung der Daten nach oben UND unten hin)
- Prädiktorauswahl
- Modeliert jede konditionale Verteilung einzeln. Keine Garantie, dass eine gemeinsame Verteilung mit diesen konditionalen Verteilungen exisitiert (aktives Forschungsfeld)
- Wie viele Iterationen?
  - Empirische Studien zeigen, dass nach 5 oder 6 Iterationen, sich nicht mehr viel ändert

HFCS-Imputationsmodell ↪ Ausgewählte Ergebnisse

Gewichtete Mittelwerte für ausgewählte Variablen vor und nach multipler Imputation

	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]	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.

[1] 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.

Auswertung des multipel imputierten HFCS-Datensatzes $\hookrightarrow$ Prinzip

- M vollständige Datensätze (z.B. $M = 5$)
- Analysiere jeden vollständigen Datensatz einzeln
- Kombiniere Ergebnisse auf einfacher Art und Weise für Inferenz
- Besonders praktisch für öffentlich zugängliche Datensätze
  - Datenerheber erstellt Imputationen für die User, welche die Daten mit complete-data Methoden auswerten können

Auswertung des multipel imputierten HFCS-Datensatzes $\hookrightarrow$ Rubin Rules

- θ = der zu schätzende Parameter
- $\hat{\theta}_l$ = der Schätzer des l-ten vollständigen Datensatzes ($l = 1, \ldots, M$)
- U_l = der Schätzer der Varianz von $\hat{\theta}_l$ von der l-ten Analyse
- Dann ist der MI-Schätzer von θ:

$$\bar{\theta} = \frac{1}{M} \sum_{l=1}^{M} \hat{\theta}_l$$

- Der MI-Schätzer der Varianz ist:

$$V = \bar{U} + \frac{M+1}{M} B$$

  - $\bar{U} = \frac{1}{M} \sum_{l=1}^{M} U_l$ ist die within-Imputation Varianz
  - $B = \frac{1}{M-1} \sum_{l=1}^{M} (\hat{\theta}_l - \bar{\theta})^2$ ist die between-Imputation Varianz

Auswertung des multipel imputierten HFCS-Datensatzes ↪ Beispiel

Datensatz (l)	$\widehat{\theta_{1,l}}$	$U_{1,l}$	$\widehat{\theta_{2,l}}$	$U_{2,l}$
1	12,6	3,6^2	4,32	1,95^2
2	12,6	3,6^2	4,15	2,64^2
3	12,6	3,6^2	4,86	2,09^2
4	12,6	3,6^2	3,98	2,14^2
5	12,6	3,6^2	4,50	2,47^2
Mittelwert	**12,6**	**3,6^2**	**4,36**	**2,27^2**
Varianz	**0**		**0,339**	

θ	$\bar{\theta}$	$\bar{U}$	B	$\sqrt{V} = \sqrt{\bar{U} + \frac{6}{5}B}$
θ_1	12,6	3,6^2	0	3,6
θ_2	4,36	2,27^2	0,339	2,36